Section of Biostatistics • Department of Public Health
Faculty of Health and Medical Sciences
University of Copenhagen • Denmark
PhD thesis • 2025

Zehao Su
Combined evidence of treatment effects

Printed by Campus Print • University of Copenhagen

Zehao Su
Combined evidence of treatment effects
2025

UNIVERSITY OF COPENHAGEN
FACULTY OF HEALTH AND MEDICAL SCIENCES

PhD thesis

Zehao Su

Combined evidence of treatment effects

Errata to "Combined evidence of treatment effects"

Chapter 2

Page 7, line 28	The overlap condition $P(D = 1 \mid X) > 0$ should read $P(D = 0 \mid X) > 0$.
Page 12, line 23	The notation for empirical average $\mathbb{P}_n f = \sum_{i=1}^n f(O_i)$ should read $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(O_i)$.

Chapter 4

Page 28, equation (4.16)	The summation in the second term on the right hand side of the equation $\sum_{d' \in \{0,1\}} \eta(d \mid X)[\cdots]$ should read $\sum_{d' \in \{0,1\}} \eta(d' \mid X)[\cdots]$.
Page 29, line 9	The inverse-variance weight should instead be $[E\{\text{var}^{-1}(\tilde{\epsilon} \mid X, D) \mid X, G = 0\}]^{-1}\text{var}^{-1}(\tilde{\epsilon} \mid X, D)$.

Chapter 5

Page 38, line 21	The condition for function h_1 should be $E\{h_1(t, X, D) \mid T \geq t, A = 0, X\} = 0$.
Page 39, line 8	The notation of the orthocomplement of the tangent space should be $\dot{\mathcal{P}}^{\perp}_{\ll}$. The functions h_1 and h_2 also depend on d.

Manuscript I

Page 15, Table 1	The parameter γ should be defined as $E\{Y(-1) - Y(1) \mid S = 0\}$.

Manuscript III

Page 10, Theorem 1	Condition (ii) for consistency of $\hat{\theta}_1(0)$ should be $\bar{\mathrm{A}}_{\bullet 1} = \mathrm{A}_{\bullet 1}$, $\bar{e}_1 = e_1$, and $\bar{\mathrm{A}}_1^c = \mathrm{A}_1^c$.

Zehao Su

Combined evidence of treatment effects

This thesis has been submitted to the Graduate School of Health and Medical Sciences, University of Copenhagen on February 28, 2025.

PhD thesis

Academic advisors

Frank Eriksson	*University of Copenhagen*
Henrik Ravn	*Novo Nordisk A/S*
Helene Rytgaard	*University of Copenhagen*

Assessment committee

Brice Ozenne	*University of Copenhagen*
Christian Dehlendorff	*Danish Medicines Council*
Shu Yang	*North Carolina State University*

Acknowledgement

This thesis was completed at the Section of Biostatistics, University of Copenhagen. I would like to thank everyone at the section for making it a wonderful place to work. I extend my deepest gratitude to my academic advisors, Frank Eriksson, Henrik Ravn, and Helene Rytgaard, for their guidance and encouragement. I could not have asked for a better advisor than Frank, with whom I shared many gratifying and inspiring discussions throughout the three years. I am thankful to Stijn Vansteelandt for hosting my research stay at Ghent University and other colleagues there for their hospitality. I would like to thank Troels for his emotional support during this journey. I am forever grateful to my family, especially my parents and grandparents, who have always supported me in my decision to pursue academic degrees.

Abstract

Although randomized controlled trials are the gold standard for assessing treatment effects, they can lack external validity and are increasingly costly and time-consuming. Combining existing data provides an opportunity to generate evidence of treatment effects in external populations without the need to conduct new trials. Incorporating external data into the trials can also improve the precision of treatment effect estimates. This thesis contributes three works of statistical methods for causally interpretable parameters through data combination. The first work establishes the identifiability of the target population average treatment effect in the presence of unmeasured effect modifiers using proxy variables. The second work is concerned with efficient estimation of the marginal treatment effect with data from multiple sources under transportability of the stratified treatment effects. The third work proposes estimators of cumulative incidences in trials with competing risks, incorporating external controls. The developed methods are illustrated with data from placebo-controlled trials investigating the effect of semaglutide and liraglutide (glucagon-like peptide-1 receptor agonists) on body weight and cardiovascular events.

Resumé

Randomiserede kontrollerede forsøg betragtes som guldstandarden for vurdering af behandlingseffekter. Dog kan de mangle ekstern validitet og er i stigende grad både omkostningstunge og tidskrævende. En løsning er at anvende kombinerede data til at generere evidens for behandlingseffekter i eksterne populationer uden at skulle udføre nye forsøg. I eksisterende studier kan denne metode også bruges til at øge præcisionen af behandlingseffektsestimater ved at integrere eksterne data. Denne afhandling præsenterer tre artikler om statistiske metoder for kausalt fortolkelige parametre ved kombination af data. Den første artikel beskriver identificerbarheden af målpopulationens gennemsnitlige behandlingseffekt i tilfældet med umålte effektmodifikatorer ved brug af såkaldte stedfortrædervariable. Den anden artikel fokuserer på efficient estimation af den marginale behandlingseffekt med data fra flere kilder ved transportabilitet af stratificerede behandlingseffekter. Den tredje artikel foreslår estimatorer for kumulative incidenser i studier med konkurrerende afgangsårsager, som integrerer eksterne kontrolpersoner. De udviklede metoder illustreres med data fra placebokontrollerede studier, der undersøger effekten af semaglutid og liraglutid (glukagonagtige peptid 1-receptoragonister) på kropsvægt og kardiovaskulære hændelser.

Contents

Acknowledgement i
Abstract iii
Resumé v

Synopsis

1 Introduction 3
1.1 *Motivation* 3
1.2 *Objectives* 4
1.3 *Organization* 6
2 Assumptions and estimators 7
2.1 *Typical assumptions* 7
2.2 *Plug-in estimators* 9
2.3 *Geometric approach to estimation* 10
2.4 *Robust estimation* 12
3 Proximal causal inference 15
3.1 *Unmeasured confounding* 15
3.2 *Proximal identification and estimation* 15
3.3 *Selection diagrams* 19
3.4 *Relations to unmeasured effect modifiers* 21
4 Multisource data fusion 25
4.1 *Setup and identifiability* 25
4.2 *Efficiency bounds in nested models* 27
4.3 *Variational dependence and non-collapsibility* 31
5 Competing risks analysis 33
5.1 *Setup and identifiability* 33
5.2 *Two transportability assumptions* 35
5.3 *Efficiency in three models* 36
6 Summary of manuscripts 41
6.1 *Manuscript I* 41
6.2 *Manuscript II* 41
6.3 *Manuscript III* 42
7 Discussion and outlook 45

Bibliography 47

Manuscripts

I Proximal indirect comparison
by Su, Z., Rytgaard, H. C. W., Ravn, H., and Eriksson, F.. . 55

II Efficient estimation of target population average treatment effect from multi-source data
by Su, Z., Rytgaard, H. C. W., Ravn, H., and Eriksson, F.. . 101

III Improving precision of cumulative incidence estimates in randomized controlled trials with external controls
by Su, Z., Rytgaard, H. C. W., Ravn, H., and Eriksson, F.. . 165

Synopsis

1. Introduction

Collecting evidence of treatment effects is a fundamental task in biostatistics. In practice, available data may be combined to generate new evidence or to improve existing evidence. This thesis is concerned with statistical inference of treatment effects through data combination and contribute with three methodological works.

1.1. Motivation

A new treatment under development is usually evaluated first in clinical trials for efficacy and safety. Ideal randomized controlled trials (RCTs) are the gold standard for evaluating treatment effects, since the evidence can be interpreted causally and enjoys internal validity. This constitutes the primary source of evidence, which comes from well-controlled experimental settings and often targets the treatment effect of interest through study designs. After obtaining approval from regulatory bodies, the treatment is then applied by medical practitioners and reaches a broader population.

However, conducting an RCT for every new target population under investigation may not be time- nor cost-effective, when data from other RCTs comparing the same treatments in similar populations are readily available. On the other hand, existing RCT results may not be directly applicable to new target populations by their lack of external validity due to treatment effect heterogeneity and selection on effect modifiers. In some situations, administration of certain treatments may be difficult or unethical according to the standard of care, necessitating the use of non-experimental data or historical data.

Integrating evidence from other sources into the target population is desirable when data obtained from the target population alone do not best describe the underlying treatment effect (Cole et al., 2023). Depending on the scientific question, the target population may be a subset of the source population or a separate population on its own. Using information from the source population to make statistical inferences of treatment effects is usually referred to as generalizability in the former scenario and transportability in the latter. While both scenarios are common in applications, we focus on transportability in the current work. Most of the statistical theory for transportability can be easily adapted to accommodate generalizability. Transportability studies usually follow a non-nested design, where samples are collected separately from the source and target populations (Dahabreh et al., 2021). See Degtiar and Rose (2023); Colnet et al. (2024c) for comprehensive reviews of data combination methods.

1.2. Objectives

In the following, we discuss three particular use cases of data combination methods studied in the manuscripts. We also briefly describe the RCTs in the data examples that illustrate the methods developed. Note that the problems presented below are the main objectives of the thesis but are far from being exhaustive; see Dahabreh (2024) for additional use cases where researchers may wish to combine information from different sources.

1.2.1. *Indirect comparison*

Indirect comparison is a specific form of meta-analysis, where the head-to-head comparison of two treatments of interest is not available from the same RCT. It is often the case that these treatments are previously studied in two separate RCTs, both of which contain an arm for a third treatment. In this case, the shared treatment arm allows for evidence synthesis using the contrast between treatments rather than information within the treatment arms. This is referred to as an anchored comparison (Phillippo et al., 2018).

SCALE (Davies et al., 2015) and STEP-2 (Davies et al., 2021) are two RCTs whose main objective was to study the effect of glucagon-like peptide-1 (GLP-1) agonists on body weight loss among overweight or obese adults with type 2 diabetes. The active treatments were once-daily liraglutide, 3.0 mg or 1.8 mg in SCALE and once-weekly semaglutide, 2.4 mg or 1.0 mg in STEP-2. Both RCTs were placebo-controlled, but the frequency of administration and the injection volume of placebos were matched to their respective active treatments. Since SCALE was initiated 5 years earlier than STEP-2, it is unclear whether liraglutide would have achieved the same weight loss effect as that reported in SCALE, had SCALE been conducted within the time frame of STEP-2.

Suppose we are interested in comparing the effect of the higher dose liraglutide against that of the higher dose semaglutide in the study population of STEP-2. The direct comparison of these treatment is not available from any superiority trial except STEP-8 (Rubino et al., 2022), which instead studied overweight or obese patients without diabetes in the United States (US). Although SCALE and STEP-2 employed highly comparable inclusion-exclusion criteria, they recruited patients from 9 and 12 countries, respectively. This leads to a potential discrepancy between the study populations of the RCTs, and the effect of liraglutide versus placebo may not be naively transported for indirect comparison. Nevertheless, if the study population of STEP-2 is a subset of the study population of SCALE, the effect of the missing liraglutide against semaglutide in STEP-2 may still be obtained from the treatment-outcome information in both RCTs.

1.2.2. *Multisource data*

It is common to observe evidence of treatment effects in multiple datasets collected from different sources. For instance, in a nationwide RCT, patients often visit different clinics across the country where they receive the same treatment options. Hospitals located in different countries may independently conduct RCTs for the same treatments to produce a reliable treatment effect estimate within their own cohorts. When the source studies contain fine-grained information on treatment effect heterogeneity, we may combine these RCTs to produce an interpretable treatment effect estimate in the actual target

population of interest.

STEP-1 (Wilding et al., 2021) is a placebo-controlled, multi-center global RCT for the effect of once-weekly semaglutide, 2.4 mg on body weight. The study population of STEP-1 comprised overweight or obese patients without diabetes, and the study sample was collected from 129 centers in a total of 16 countries. Suppose we are interested in evaluating the weight loss effect of semaglutide in the US sub-population of the entire study population of STEP-1 recruited from a large region, had the treatment and outcome information not been collected. The US sample was the largest among all countries, making up about 39% of the entire sample of STEP-1. However, due to the diverse racial and ethnic demography of the US, transporting experiment results using samples from any single country or region may produce a biased treatment effect. Viewing participants in the United Kingdom, the rest of European countries, and the East Asian countries as three separate RCT samples, we might be able to obtain an estimate of the US sub-population treatment effect based on all three source studies. If this was possible, future study plannings may focus on sampling schemes that more realistically reflect the composition of potential target patients of the treatment.

1.2.3. *External controls*

Many existing RCTs investigating treatments in similar patient groups have a control arm in which either a placebo or standard of care is administered. From an economic viewpoint, it makes sense to allocate more participants to receiving new treatments with less evidence, if some controls can be borrowed from existing RCTs (Viele et al., 2014). Even for sufficiently balanced designs, augmenting the control arm in an RCT may further strengthen the evidence of a treatment effect. In studies of treatments for rare diseases, where the number of eligible patients can be relatively small, it may be more feasible to consider single-arm designs that rely completely on external controls for the comparison of treatment (Davi et al., 2020).

SUSTAIN-6 (Marso et al., 2016a) and LEADER (Marso et al., 2016b) are two cardiovascular outcome trials for assessing the safety of once-weekly semaglutide, 1.0 mg and 0.5 mg, and once-daily liraglutide, 1.8 mg, in patients with type 2 diabetes. In SUSTAIN-6, patients were randomized to one of the two doses of semaglutide or their volume-matching placebos to create arms of equal size. In LEADER, patients were randomized to liraglutide or placebo in a 1 to 1 ratio. The primary outcome used in both RCTs was the composite event of the first occurrence of three major adverse cardiovascular events. LEADER ran with 9340 randomized patients and a median follow-up of 3.8 years, and SUSTAIN-6 had a total of 3297 patients with a median follow-up of 2.1 years. Since the control arm in LEADER is almost 3 times as large as the combined control arms of SUSTAIN-6 and patients were generally followed over longer periods of time, it is particularly interesting to incorporate LEADER controls into SUSTAIN-6 as external controls. A re-analysis of the data may be able to consolidate the superiority of semaglutide for protection against the primary event in SUSTAIN-6 by increasing the precision of effect estimates.

1.3. Organization

The thesis includes a synopsis of seven chapters and three manuscripts. The remaining chapters of the synopsis are organized as follows. Chapter 2 presents a selective overview of existing estimators of target population treatment effects in transportability. Chapter 3 reviews the proximal causal inference framework for unmeasured confounding, which inspires the work on proximal indirect comparison in Manuscript I. Chapter 4 covers the connection between testable implications of transportability assumptions and efficient estimation of target population treatment effects, which we extend to a multi-source data setting in Manuscript II. Chapter 5 discusses transportability assumptions in competing risks analysis, and Manuscript III adopts one such assumption to accommodate external controls in estimation of the target population treatment effect. Chapter 6 provides a summary of the manuscripts, and finally Chapter 7 contains a general discussion on the thesis and future research directions.

2. Assumptions and estimators

2.1. Typical assumptions

To ground ideas, we restrict attention to a specific setup of data combination. We work with a non-nested study design where samples originate from a source population ($D = 0$) and a target population ($D = 1$), indicated by a binary variable D. For subjects from the target population, we do not have information on the binary treatment $A \in \{0, 1\}$ nor the outcome Y, but both variables are observed for subjects from the source population. Additionally, a set of baseline covariates X is available from all subjects. We assume the joint sample is an i.i.d. sample of size n from the distribution P over $O = \{(1 - D)Y, (1 - D)A, X, D\}$. We use n_0 and n_1 to denote the number of samples from the source and target populations.

We consider the target population average treatment effect (TATE) defined as

$$\theta = E\{Y(1) - Y(0) \mid D = 1\},$$

where $Y(a)$ denotes the potential outcome under the treatment assignment $A = a$. Ideally, we would like to have access to the full data distribution over $\{Y(1), Y(0), X, D\}$, where θ is immediately identifiable. However, we do not even observe the factual outcomes in the target population. To proceed from here, it appears necessary to make assumptions on the compatibility of the two populations, namely transportability assumptions.

There are multiple choices of transportability assumptions, all of which are conditional on the baseline covariates. If the baseline covariates consist of only categorical variables, this is identical to performing a stratification and claiming that the subjects within the same stratum are transportable across the populations in a certain sense. Therefore, it is important that an overlap condition holds, so that we do not extrapolate outside of the support of the source population. Denote the supports of the baseline covariates in the target and source populations by $\mathcal{X}_1$ and $\mathcal{X}_0$. The overlap condition is

$$\mathcal{X}_1 \subset \mathcal{X}_0, \tag{2.1}$$

or equivalently, $P(D = 1 \mid X) > 0$.

Now we present transportability assumptions on the conditional distributions of the potential outcomes, on the conditional treatment-specific means, on the conditional distribution of individual treatment effect, and on the conditional average treatment

Table 2.1. *Transportability assumptions appearing in publications.*

Assumption	Publications
(2.2)	Stuart et al. (2011); Cole and Stuart (2010); Hotz et al. (2005); Buchanan et al. (2018); Vo et al. (2019)
(2.3)	Dahabreh et al. (2020a, 2019)
(2.4)	Tipton (2013); Kern et al. (2016); Nguyen et al. (2018)
(2.5)	Tipton et al. (2014); Dahabreh et al. (2023); Lee et al. (2023).

effect (CATE):

$$Y(a) \perp\!\!\!\perp D \mid X, \tag{2.2}$$

$$E\{Y(a) \mid X, D\} = E\{Y(a) \mid X\}, \tag{2.3}$$

$$\{Y(1) - Y(0)\} \perp\!\!\!\perp D \mid X, \tag{2.4}$$

$$E\{Y(1) - Y(0) \mid X, D\} = E\{Y(1) - Y(0) \mid X\}. \tag{2.5}$$

All the above assumptions are ubiquitous in the literature on transportability and generalizability of treatment effects; see Table 2.1 for a list of publications that make use of these assumptions. Assumptions 2.2 and 2.4 are apparently stronger than Assumptions 2.3 and 2.5, respectively. The former two assumptions are on the conditional distribution of the counterfactuals, while the latter two are on the conditional means in these distributions. Since the mean scale is a common choice to summarize the distribution of continuous and binary outcomes, Assumptions 2.3 and 2.5 often suffice. Here we have chosen to evaluate the treatment effect through the difference of potential outcomes, and assumptions 2.4 and 2.5 exploit this particular specification of the target parameter. Assumption 2.2 alone does not imply Assumption 2.4. However, Assumption 2.2 is often stated in the joint distribution of the potential outcomes as $\{Y(1), Y(0)\} \perp\!\!\!\perp D \mid X$, and in this formulation it becomes strictly stronger than Assumption 2.4.

These transportability assumptions are sufficient for the external validity of the treatment effect in the source population. The TATE can be identified through the conditional treatment effects in the source population by standardization. Then, if internal validity can be established for the source population, we can identify θ in the observed data distribution. Let $e_0(a \mid x) = P(A = a \mid X = x, D = 0)$ be the propensity score for receiving treatment a in the source population. A minimum set of assumptions frequently made in causal inference is consistency, positivity of treatment assignment, and conditional exchangeability of treatment assignment:

$$(1 - D)Y(A) = (1 - D)Y, \tag{2.6}$$

$$0 < e_0(a \mid x) < 1, \quad x \in \mathcal{X}_0, \tag{2.7}$$

$$Y(a) \perp\!\!\!\perp A \mid \{X, D = 0\}. \tag{2.8}$$

In particular, these assumptions are fulfilled if subjects from the source population are solicited from a randomized controlled trial. Clinical trials have well-defined, concrete treatments so consistency holds. Under proper randomization, which is allowed to depend on the baseline covariates, positivity and conditional mean exchangeability also hold for treatment assignment. If subjects from the source population are derived from observational data, the validity of these causal assumptions requires further examination.

Let $\mu_0(a,x) = E(Y \mid A = a, X = x, D = 0)$ denote the conditional mean outcome under treatment a in the source population. Let $\pi(x) = P(D = 1 \mid X = x)$ denote the sampling score (or selection score) of an individual being selected into the target population and $\alpha = P(D = 1)$ the target population's mixing proportion in the combined population. Under assumptions (2.6)–(2.8) and one of assumptions (2.2)–(2.5), the target parameter is identifiable as a functional of P:

$$\theta = E\{\mu_0(1,X) - \mu_0(0,X) \mid D = 1\}. \tag{2.9}$$

This is a type of g-formula. The conditional mean outcome difference between treatment 1 and treatment 0 quantifies the treatment effect heterogeneity among subjects in the source population. This conditional effect is then standardized according to the distribution of the baseline covariates in the target population to produce the target-population average treatment effect. Alternatively, the parameter may be written in a weighting formulation, such that

$$\theta = E\left\{\frac{1-D}{\alpha}\frac{\pi(X)}{1-\pi(X)}\frac{2A-1}{e_0(A \mid X)}Y\right\}. \tag{2.10}$$

Here, an outcome associated with a subject in the source population undergoes two weightings. It is first weighted by the inverse propensity score within the population to achieve balance of the treatment groups. Then it is recalibrated with the odds of being in the target population to account for the discrepancy between covariate distributions in the two populations, up to a constant scaling by the ratio of sampling proportions.

2.2. Plug-in estimators

Akin to the central role of the propensity score in observational studies for inferring causal effects (Rosenbaum and Rubin, 1983), the sampling score plays an important part in transportability studies. As a motivating argument, we notice that assumption (2.4), transportability of the individual treatment effect, implies strong ignorability of sample selection given the sampling score (Tipton, 2013)

$$\{Y(1) - Y(0)\} \perp\!\!\!\perp D \mid \pi(X).$$

Therefore, if strong ignorability of treatment assignment also holds, then

$$\begin{aligned}\theta = E[&E\{Y \mid A = 1, e_0(1 \mid X), \pi(X), D = 0\} \\ &- E\{Y \mid A = 0, e_0(1 \mid X), \pi(X), D = 0\} \mid D = 1]\end{aligned} \tag{2.11}$$

This provides an intuition for simultaneous matching on the sampling score and the propensity score; see Stuart et al. (2011); Tipton (2013); O'Muircheartaigh and Hedges (2014) for a detailed account for treatment effect generalization. If stratified matching is used, the resulting estimator will behave like a weighting estimator, when the number of strata and the sample size approach infinity (Rubin, 2001). The weighting estimator is the sample analog of (2.10):

$$\frac{1}{n_1}\sum_{i:D_i=0}\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\frac{2A_i-1}{\hat{e}_0(A_i \mid X_i)}Y_i.$$

Every outcome from the source population is weighted by the product of the inverse propensity score and the odds of being sampled into the target population. Weighting estimators are found in both generalizability (Cole and Stuart, 2010; Lesko et al., 2017; Dahabreh et al., 2019) and transportability (Westreich et al., 2017). For protection against pathologically large odds, the Hájek estimator can be used in place of the Horvitz-Thompson weighting estimator above, by replacing the normalizing factor with the stable weights (Dahabreh et al., 2020a)

$$\frac{\sum_{i:D_i=0} \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} \frac{A_i}{\hat{e}_0(1 \mid X_i)} Y_i}{\sum_{i:D_i=0} \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} \frac{A_i}{\hat{e}_0(1 \mid X_i)}} - \frac{\sum_{i:D_i=0} \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} \frac{1-A_i}{\hat{e}_0(0 \mid X_i)} Y_i}{\sum_{i:D_i=0} \frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)} \frac{1-A_i}{\hat{e}_0(0 \mid X_i)}}.$$

An alternative approach relies on the outcome regression estimator $\hat{\mu}_0(a, x)$ from the source population. The g-estimator based on (2.9) is thus (Zhang, 2009; Dahabreh et al., 2020a)

$$\hat{\theta}_{\mathrm{g}} = \frac{1}{n_1} \sum_{i:D_i=1} \{\hat{\mu}_0(1, X_i) - \hat{\mu}_0(0, X_i)\}.$$

Samples from the source population are used to provide the conditional treatment effects, which are subsequently standardized on the empirical distribution of the target population to produce the average treatment effect.

Consistency of the g-estimator and the weighting estimator can be established under the correct specification of $\hat{\mu}_0$ and $\{\hat{\pi}, \hat{e}_0\}$, respectively (Colnet et al., 2022). These estimators can be reformulated as Z-estimators that solve certain estimation equations. If the nuisance parameter models are parametric, the estimators will then typically be root-n consistent and converge to a normal distribution, whose standard error may be estimated by the sandwich estimator (Buchanan et al., 2018) or nonparametric bootstrap (Dahabreh et al., 2020a). Suppose the study design grants knowledge on the sampling probability and the propensity score. Then we would expect that parametric models are adequate for the weighting estimator. However, the g-estimator is blind to study designs and therefore does not enjoy this privilege. To avoid model misspecifications, one option is to forego parametric models and turn to nonparametric models like machine learning. Yet the resulting estimators will inherent the sub-parametric convergence rates from these models, leading to plug-in bias that ruins the asymptotic normality.

Fortunately, in many cases, the study of target parameters as functionals of the underlying probability distribution reveals estimators with robustness against model misspecifications and slow convergence rates due to estimation of the nuisance parameters. A classical example in causal inference is the augmented inverse probability weighting estimator for the average treatment effect (Hirano et al., 2003; Bang and Robins, 2005).

2.3. Geometric approach to estimation

This section largely follows van der Vaart (1998, Chapter 25). A set of probability measures $\mathcal{P}$ on the sample space $\mathcal{O}$ equipped with a suitable σ-algebra is called a (semiparametric) model. A model solely indexed by a Euclidean parameter is parametric. The goal of the theory we present is to study the statistical inference of a real-valued

functional $\theta : \mathcal{P} \to \mathbb{R}$, which is also called the target parameter or the parameter of interest.

Suppose the observed data $\{O_i : i = 1, \ldots, n\}$ is an independent and identically distributed (i.i.d.) sample from a distribution $P \in \mathcal{P}$ with density p. To determine the characteristics of the "best" estimator in the asymptotic sense is hard, but some intuition from parametric models carries over to the general semiparametric case, where the model may be indexed by an infinite-dimensional parameter. Specifically, we want to use the Cramér-Rao lower bound, which requires two ingredients: the score of the indexing parameter in some parametric model and the derivative of the target parameter with respect to the indexing parameter. To this end, we first define smooth one-dimensional paths that pass through the underlying density p that are differentiable in an appropriate sense. Heuristically, for every path $\{p_\varepsilon\}$ with $p_\varepsilon|_{\varepsilon=0} = p$, the associated score function is

$$s(o) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \log\{p_\varepsilon(o)\}\bigg|_{\varepsilon=0}.$$

We call the set of all the score functions of the differentiable paths the tangent set and denote it by $\dot{\mathcal{P}}$. For simplicity, we also assume that $\dot{\mathcal{P}}$ is a closed linear subspace of the Hilbert space of mean-zero $L_2(P_0)$-functions, which we denote by $L_2^0(P_0)$. Suppose the target parameter is pathwise differentiable relative to the tangent space; that is, there exists a continuous linear functional $\dot{\theta} : \dot{\mathcal{P}} \to \mathbb{R}$ such that for every differentiable path $\{p_\varepsilon\}$ with score s,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta(P_\varepsilon)\bigg|_{\varepsilon=0} = \dot{\theta}(s).$$

The pathwise derivative $\dot{\theta}$ behaves like a Hadamard derivative in the sense that uniform convergence along the paths is guaranteed (Bickel et al., 1993, Appendix A). We call any function $\phi \in L_2^0(P)$ such that

$$\dot{\theta}(s) = E\{\phi(O)s(O)\}, \quad \text{for all } s \in \dot{\mathcal{P}},$$

an influence function (IF) of the parameter θ. A parameter can have more than one influence function if the tangent space $\dot{\mathcal{P}}$ is a proper subspace of $L_2^0(P)$. However, it is always possible to find an almost-everywhere unique IF that lies within the tangent space by projecting any IF onto the space. We call this projection the efficient influence function (EIF), which is $\varphi \in \dot{\mathcal{P}}$ such that

$$\dot{\theta}(s) = E\{\varphi(O)s(O)\}, \quad \text{for all } s \in \dot{\mathcal{P}},$$

The EIF is also the Riesz representer of the functional $\dot{\theta} : \dot{\mathcal{P}} \to \mathbb{R}$ that satisfies

$$\sup_{s \in \dot{\mathcal{P}}} \frac{|\dot{\theta}(s)|}{[E\{s^2(O)\}]^{1/2}} = [E\{\varphi^2(O)\}]^{1/2}.$$

Any influence function ϕ can be conveniently obtained by (Tsiatis, 2006, Chapter 4)

$$\phi = \varphi + g, \quad \text{for some } g \in \dot{\mathcal{P}}^{\perp}.$$

IFs and the EIF are also known as gradients and the canonical gradient (Pfanzagl, 1982).

Estimation problems in one-parameter parametric sub-models are conceivably easier than those in the larger, more complicated semiparametric model $\mathcal{P}$. Therefore, we

should not expect any "unbiased" estimator of θ to have a smaller variance in $\mathcal{P}$ than the supremum of the Cramér-Rao lower bound of θ in all parametric sub-models,

$$\sup_{\{P_\varepsilon\}\in\mathcal{P}} \frac{\left(\mathrm{d}\theta(P_\varepsilon)/\mathrm{d}\varepsilon|_{\varepsilon=0}\right)^2}{E\{s^2(O)\}} = \sup_{s\in\dot{\mathcal{P}}} \frac{[E\{s(O)\varphi(O)\}]^2}{E\{s^2(O)\}} = E\{\varphi^2(O)\}. \tag{2.12}$$

This intuition is formalized by the convolution theorem. However, we need to restrict the class of estimators to the so-called regular estimators. Heuristically, the limiting distribution of a regular estimator does not depend on the local data generating process, so that its convergence is in some sense locally uniform. Any regular estimator T_n converging in distribution such that

$$n^{1/2}(T_n - \theta) \rightsquigarrow L$$

satisfies that the law of the random element L is a convolution of the law of a zero-mean normal random variable Z with variance $E\{\varphi^2(O)\}$ with another probability measure. We say that a regular estimator is efficient if its asymptotic variance achieves the semiparametric efficiency bound $E\{\varphi^2(O)\}$. A regular estimator T_n is asymptotically linear if there exists a function $f \in L_2^0(P)$ such that

$$T_n - \theta = \frac{1}{n}\sum_{i=1}^n f(O_i) + o_{P_0}(n^{-1/2}). \tag{2.13}$$

The function f is called the influence function of the estimator. It can be shown that any regular efficient estimator is asymptotically linear with influence function φ.

2.4. Robust estimation

We show how the EIF (or more generally, the IFs) of a pathwise differentiable target parameter may be useful in statistical inference. Suppose the EIF $\varphi(o) = \varphi(o, \theta, \nu)$ of θ depends explicitly on the target parameter and some (possibly infinite-dimensional) nuisance parameter ν. For a possibly random function $f(o)$ of the data, let $\mathbb{P}_n f = \sum_{i=1}^n f(O_i)$ and let $Pf = \int f(o)\mathrm{d}P(o)$. Let $\hat{\nu}$ be an estimator of ν. We define a Z-estimator $\hat{\theta}$ solving the estimating equation based on the EIF:

$$\frac{1}{n}\sum_{i=1}^n \varphi(O_i, \hat{\theta}, \hat{\nu}) = o_P(n^{-1/2}).$$

The nuisance parameter often comprises several components, and for exposition we assume that ν can be partitioned into two components (ν_1, ν_2). In certain cases, the mean of the EIF satisfies (Chernozhukov et al., 2022b)

$$\left|E[\varphi\{O, \theta, (\nu_1^*, \nu_2^*)\}]\right| = \left|E\{(\nu_1^* - \nu_1)(\nu_2^* - \nu_2)\}\right|,$$

for any reasonable choice $\nu^* = (\nu_1^*, \nu_2^*)$ in the nuisance parameter space. Then the EIF is a valid score of θ, so the consistency of $\hat{\theta}$ may be established, if either ν_1 or ν_2 is correctly specified. This is called robustness against model misspecification.

There is also a close connection between IFs and Neyman orthogonality (Chernozhukov et al., 2018, 2022a), which means that the first-order impact of the misspecification of nuisance parameter is ignorable. We say that φ is Neyman orthogonal to ν if for any reasonable ν^*, it holds that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} E[\varphi\{O, \theta, \nu + \varepsilon(\nu^* - \nu)\}]\Big|_{\varepsilon=0} = 0. \tag{2.14}$$

Further assuming differentiability of φ in the target parameter and expanding the estimating equation around the true parameters $\{\theta, \nu\}$ yield the bias decomposition

$$\begin{aligned}\hat{\theta} - \theta = \mathbb{P}_n\varphi(\cdot, \theta, \nu) + (\mathbb{P}_n - P)\{\varphi(\cdot, \hat{\theta}, \hat{\nu}) - \varphi(\cdot, \theta, \nu)\} \\ + O_P(\|\hat{\nu} - \nu\|^2) + o_P(n^{-1/2}),\end{aligned} \tag{2.15}$$

for an appropriate norm on the nuisance parameter space. Comparing this to (2.13), the current estimator $\hat{\theta}$, if regular, is efficient if the non-leading terms all share the order $o_P(n^{-1/2})$. The cross-product term can be handled by empirical process theory and a Donsker class condition for the class of functions $\{\varphi_P : P \in \mathcal{P}\}$ (van der Vaart and Wellner, 2023). More generally, crossfitting can be applied for nuisance parameters that cannot be reached by any Donsker class (Zheng and van der Laan, 2011; Chernozhukov et al., 2018; Kennedy, 2024). Irrespective of the approach, we typically require consistency of $\{\hat{\theta}, \hat{\nu}\}$. The structure of the error term $O_P(\|\hat{\nu} - \nu\|^2)$ provides robustness for target parameter estimation. Many nonparametric nuisance estimators achieve the subparametric rate $\|\hat{\nu} - \nu\| = o_P(n^{-1/4})$. The second-order error accommodates slow convergence rates from flexible (regression) models This is called rate robustness.

Returning to the specific transportability setup, the semiparametric efficiency bound of the TATE θ is characterized by the EIF

$$\varphi(o) = \frac{1-d}{\alpha}\frac{\pi(x)}{1-\pi(x)}\frac{2a-1}{e_0(a \mid x)}\{y - \mu_0(a, x)\} + \frac{d}{\alpha}\{\mu_0(1, x) - \mu_0(0, x) - \theta\}.$$

The EIF is indeed a Neyman orthogonal score for θ, since for arbitrary functions $\pi^*(x)$, $e_0^*(a \mid x)$, and $\mu_0^*(a, x)$, the Gateaux derivative vanishes:

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} E(\varphi[O, \theta, \{\pi + \varepsilon(\pi^* - \pi), e_0 + \varepsilon(e_0^* - e_0), \mu_0 + \varepsilon(\mu_0^* - \mu_0)\}])\Big|_{\varepsilon=0} = 0.$$

Upon closer inspection, it also has a double robustness property in the sense that

$$E\{\varphi(O, \theta, \{\pi^*, e_0^*, \mu_0\})\} = 0 \quad \text{and} \quad E\{\varphi(O, \theta, \{\pi, e_0, \mu_0^*\})\} = 0,$$

that is, the efficient influence function yields valid estimating equations when either $\{\pi, e_0\}$ or μ_0 is fixed at the truth. The Z-estimator using the obvious estimator $\hat{\alpha} = n_1/n$ and plugging in all nuisance parameter estimates is given by (Dahabreh et al., 2020a)

$$\begin{aligned}\hat{\theta} = \frac{1}{n_1}\sum_{i:D_i=0}\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\frac{2A_i - 1}{\hat{e}_0(A_i \mid X_i)}\{Y_i - \hat{\mu}_0(A_i, X_i)\} \\ + \frac{1}{n_1}\sum_{i:D_i=1}\{\hat{\mu}_0(1, X_i) - \hat{\mu}_0(0, X_i)\}.\end{aligned}$$

Clearly, the estimator $\hat{\theta}$ coincides with the so-called one-step estimator taking the g-estimator as the first-stage estimator:

$$\hat{\theta} = \hat{\theta}_{\mathrm{g}} + \mathbb{P}_n \varphi(\cdot, \hat{\theta}_{\mathrm{g}}, \{\hat{\pi}, \hat{e}_0, \hat{\mu}_0\}).$$

The asymptotic properties of the estimator can be established under suitable regularity conditions. We present the results without enumerating these conditions. Suppose the nuisance parameter estimators have probability limits with respect to the $L_2(P)$-norm such that $\|(\hat{\pi} - \bar{\pi})(X)\| = o_P(1)$, $\|(\hat{e}_0 - \bar{e}_0)(1 \mid X)\| = o_P(1)$, and $\|(\hat{\mu}_0 - \bar{\mu}_0)(a, X)\| = o_P(1)$. It turns out that the estimator $\hat{\theta}$ has both rate robustness and robustness against model misspecification. If either $\bar{\pi} = \pi$ and $\bar{e}_0 = e_0$, or $\bar{\mu}_0 = \mu_0$, then

$$\hat{\theta} - \theta = o_P(1).$$

Furthermore, if $\bar{\pi} = \pi$, $\bar{e}_0 = e_0$, and $\bar{\mu}_0 = \mu_0$ and

$$\big\{\|(\hat{\pi} - \pi)(X)\| + \|(\hat{e}_0 - e_0)(1 \mid X)\|\big\}\|(\hat{\mu}_0 - \mu_0)(a, X)\| = o_P(n^{-1/2}),$$

then $\hat{\theta}$ is efficient:

$$\hat{\theta} - \theta = \mathbb{P}_n \varphi + o_P(n^{-1/2}).$$

While we require the augmentation term

$$\frac{1}{n_1} \sum_{i:D_i=0} \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \frac{2A_i - 1}{\hat{e}_0(A_i \mid X_i)} \{Y_i - \hat{\mu}_0(A_i, X_i)\}$$

to stay bounded, if the estimated odds are large, the estimator $\hat{\theta}$ can still move the g-estimator by a considerably large margin, especially for small samples. For binary outcomes, the estimator $\hat{\theta}$ may not be a value between -1 and 1. One remedy is to first consider Z-estimators for the transformation

$$\gamma(a) = \mathrm{logit}[E\{\mu_0(0, X) \mid D = 1\}],$$

and then use the delta-method to obtain a confidence interval for the target parameter $\theta = \mathrm{expit}\{\gamma(1)\} - \mathrm{expit}\{\gamma(0)\}$. Another proposal is to apply targeted maximum likelihood estimation (Rudolph and van der Laan, 2017) by explicitly constructing a fluctuation around the estimated conditional mean model. The resulting estimate is simply a difference of averages of probabilities and thus respects the natural boundaries of the parameter.

3. Proximal causal inference

3.1. Unmeasured confounding

In this chapter, we first digress from the track of transportability of treatment effects and briefly review a framework for causal effect estimation in the presence of unmeasured confounding. A typical concern with drawing causal conclusions from observational studies is the existence of unmeasured confounders which, in principle, can never be controlled for.

Epidemiologists make use of negative controls to alert unmeasured confounding (Lipsitch et al., 2010; Shi et al., 2020b; Yang et al., 2024). For example, we may have good reasons to believe that the treatment (or exposure) cannot causally affect a variable, so the variable serves as a negative control outcome. In the case of residual association between the treatment and this variable after adjusting for the set of observed confounders, we should be more cautious about interpreting the findings causally, if sensible at all. Likewise, negative control exposures, variables that cannot be the causes of the outcome of interest, may be applied similarly to confounding bias detection.

A more quantitative approach to confounding bias relies on additional untestable assumptions. The difference-in-difference (DiD) estimator (Abadie, 2005; Heckman et al., 1997) of the average treatment effect on the treated relies on the assumption that the change in counterfactual outcome means under the control treatment from baseline to end-of-study does not depend on the treatment groups. In a certain sense, the outcomes at baseline play the role of a negative control outcome, since their conditional means in the respective treatment groups should not differ under unconfoundedness (Sofer et al., 2016). Then the parallel-trends assumption allows for the removal of confounding bias from the naive g-estimator with the difference in conditional outcome means.

Another popular method for identifying causal effects under unmeasured confounding is based on instrumental variables (IVs, Greenland, 2000; Angrist and Krueger, 2001). A valid IV must be correlated with the treatment, must only affect the outcome causally through the treatment, and cannot be confounded with the outcome after controlling for observed variables (Hernán and Robins, 2006). Under the no-defiers assumption, the complier average treatment effect is identifiable (Angrist and Krueger, 2001). Mendelian randomization is a special case of IV analysis when genetic variants such as single-nucleotide polymorphisms are used as instruments (Burgess et al., 2017).

3.2. Proximal identification and estimation

Besides the DiD and the IV approach, proximal causal inference is another closely related approach to estimation of causal effects in the presence of unmeasured confounders. It

makes use of so-called proxies as auxiliary variables that provide information on the confounding bias not captured by other observed confounders. Unlike the aforementioned methods, this framework is applicable in settings where the parallel-trends assumption is implausible and where candidate strong instruments are broken by their association with the unmeasured confounders. Like virtually all causal inference methods, its guarantee relies on a separate set of untestable assumptions related to the proxies. We refer to Tchetgen Tchetgen et al. (2024) for an extended introduction to proximal causal inference.

We consider the setting where the target parameter is the average treatment effect (ATE)

$$\theta = E\{Y(1) - Y(0)\},$$

but, provided access to the baseline variables X, conditional exchangeability of the treatment assignment (2.8) is not valid. We hypothesize that such a violation results from the existence of latent variables U that are common causes of both the treatment and the outcome. If they were measured in the data, exchangeability would be restored by adjusting for these factors:

$$\begin{aligned} &Y(A) = Y; \\ &0 < \text{pr}(A = a \mid X, U) < 1; \\ &Y(a) \perp\!\!\!\perp A \mid \{X, U\}. \end{aligned} \tag{3.1}$$

If these conditions hold, we have latent identifiability of the ATE

$$\theta = E\{E(Y \mid A = 1, X, U) - E(Y \mid A = 0, X, U)\} = E\left\{\frac{2A - 1}{\text{pr}(A \mid X, U)} Y\right\}. \tag{3.2}$$

We cannot proceed further without additional information.

To this end, suppose we have collected a pair of proxies, a treatment-inducing proxy Z and an outcome-inducing proxy W, that satisfy the following conditional independences

$$\begin{aligned} &Y \perp\!\!\!\perp Z \mid \{A, X, U\}, \\ &W \perp\!\!\!\perp \{A, Z\} \mid \{X, U\}. \end{aligned} \tag{3.3}$$

The proxies Z and W can be treated as a negative control treatment and outcome, respectively. The treatment-inducing proxy Z cannot be associated with the outcome within each treatment arm after controlling for covariates $\{X, U\}$. On the other hand, the outcome-inducing proxy W must not vary with the level of the treatment nor Z conditional on $\{X, U\}$. Assumptions (3.1) and (3.3) can be jointly replaced by the stronger exchangeability condition

$$\{Y(a), W\} \perp\!\!\!\perp \{A, Z\} \mid \{X, U\}.$$

See Figure 3.1 for a directed acyclic graph (DAG) encoding a data generating mechanism compatible with assumption (3.3). Immediately, we see that there is no independence requirement between Z and U, so the treatment-inducing proxy need not be an IV. Also important to note is that Z may be a cause of A and W may be a cause of Y. The latter is reminiscent of W being the baseline outcome in DiD.

Here, rather than making a structural assumption or working with some form of confounding bias, we posit the existence of bridge functions that mimic the behavior of

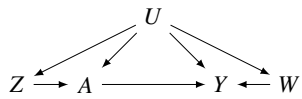

Figure 3.1. A DAG compatible with assumption (3.3). The node X is not drawn but may point to any other node.

the nuisance parameters in the latent identification formulas involving the unobserved factors. Consider the following function classes:

$$\mathcal{H}_u = \big\{h_u(a,x,w) : E\{h_u(A,X,W) \mid A,X,U\} = E(Y \mid A,X,U)\big\},$$
$$\mathcal{Q}_u = \left\{q_u(a,x,z) : E\{q_u(A,X,Z) \mid A,X,U\} = \frac{1}{\mathrm{pr}(A \mid X,U)}\right\}.$$

Any $h_u \in \mathcal{H}_u$ is an outcome bridge function of the outcome-inducing proxy W, and any $q_u \in \mathcal{Q}_u$ is a treatment bridge function of the treatment-inducing proxy Z, if they exist. The bridge functions and the proxies therein are related to the identifiability of the target parameter because their projections onto the L_2-space spanned by $\{A,X,U\}$ coincide with the unknown nuisance parameters. Formally for identifiability, we require that

$$\mathcal{H}_u \neq \emptyset \quad \text{or} \quad \mathcal{Q}_u \neq \emptyset. \tag{3.4}$$

We discuss sufficient conditions for non-emptiness of $\mathcal{H}_u$, and similar conditions can be stated for $\mathcal{Q}_u$. Consider the case where W and U are categorical variables with d_w and d_u levels. Let $M(a,x)$ denote the d_w-by-d_u matrix whose entry (j,k) is

$$\mathrm{pr}(W = w_j \mid A = a, X = x, U = u_k).$$

Then $\mathcal{H}_u \neq \emptyset$ if all $M(a,x)$ have full column rank, which automatically implies that $d_w \geq d_u$ (Miao et al., 2018; Kallus et al., 2022). The intuition is that if we want W to mimic the information contained in U, then W needs to be at least as fine-grained as U. In the general case, $\mathcal{H}_u$ is the solution set to a Fredholm integral equation of the first kind. Such integral equations are generally ill-posed problems that are challenging to solve. The sufficient and necessary condition for the existence of a solution for a special class of such equations is given by Picard's theorem (Kress, 2014, p. 311).

The bridge function classes $\mathcal{H}_u$ and $\mathcal{Q}_u$ consist of functions of observed variables only, but the characterization of the bridge functions depends on the latent variables U. Moreover, the target parameter is not identified with assumption (3.4), in the sense that there may exist latent distributions corresponding to different target parameter values but yielding the same observed distribution when U is marginalized out (Kallus et al., 2022). We consider the observed counterpart of the latent bridge function classes:

$$\mathcal{H} = \big\{h(a,x,w) : E\{h(A,X,W) \mid A,X,Z\} = E(Y \mid A,X,Z)\big\},$$
$$\mathcal{Q} = \left\{q(a,x,z) : E\{q(A,X,Z) \mid A,X,W\} = \frac{1}{\mathrm{pr}(A \mid X,W)}\right\}.$$

Comparing the definitions of $\mathcal{H}$ and $\mathcal{Q}$ to those of $\mathcal{H}_u$ and $\mathcal{Q}_u$, we notice that the unobserved confounders U are simply swapped out for the other proxy. Under assumption (3.3), the latent and observed bridge classes have the relations that

$$\mathcal{H}_u \subset \mathcal{H}, \quad \mathcal{Q}_u \subset \mathcal{Q}.$$

Assumption (3.4) implies that $\mathcal{H} \neq \emptyset$ or $\mathcal{Q} \neq \emptyset$, but we need a stronger assumption to identify the target parameter.

$$\mathcal{H} \neq \emptyset \quad \text{and} \quad \mathcal{Q} \neq \emptyset. \tag{3.5}$$

Then the ATE is identifiable in the observed data distribution as

$$\begin{aligned} \theta &= E\{h(1, X, W) - h(0, X, W)\}, && \text{for any } h \in \mathcal{H}, \\ \theta &= E\{(2A - 1)q(A, X, Z)Y\}, && \text{for any } q \in \mathcal{Q}. \end{aligned} \tag{3.6}$$

Assumption (3.5) can also be justified from the perspective of estimation. Naturally, we desire $n^{1/2}$-consistent estimators for the observed data functional θ whenever possible. It turns out that under mild regularity conditions, both $\mathcal{H}$ and $\mathcal{Q}$ being non-empty is necessary for the existence of such estimators (Severini and Tripathi, 2012; Zhang et al., 2023). Conditions for the existence of observed bridge functions are analogous to the discussion above for h_u.

Alternatively, identifiability of the ATE can be reached with assumption (3.5) and the completeness assumption that for any f,

$$\begin{aligned} &E\{f(U) \mid A, X, Z\} = 0 \Rightarrow f(U) = 0 \quad \text{or} \\ &E\{f(U) \mid A, X, W\} = 0 \Rightarrow f(U) = 0. \end{aligned} \tag{3.7}$$

Similar assumptions are used in non-parametric IV regression (Newey and Powell, 2003) and further explicated in D'Haultfœuille (2011) and Andrews (2017). This is the approach taken by, for example, Cui et al. (2024), Miao et al. (2018), and Shi et al. (2020a). The additional completeness assumption ensures that the observed bridge functions are identical to the latent ones.

The identification formulas suggest two obvious estimators. Suppose the observed bridge functions $\hat{h}$ and $\hat{q}$ are estimated from n i.i.d. copies of $O = (Y, A, X, W, Z)$. The proximal g-estimator is given by

$$\hat{\theta}_h = \frac{1}{n} \sum_{i=1}^{n} \{\hat{h}(1, X_i, W_i) - \hat{h}(0, X_i, W_i)\},$$

and the proximal inverse propensity score weighting estimator is given by

$$\hat{\theta}_q = \frac{1}{n} \sum_{i=1}^{n} (2A_i - 1)\hat{q}(A_i, X_i, Z_i)Y_i.$$

In a linear structural equation model (SEM) of continuous variables, the proximal g-estimator is easily obtained by off-the-shelf implementations of two-stage-least-squares (2SLS) estimators for IV analysis (Tchetgen Tchetgen et al., 2024). For count, binary, and categorical outcomes and outcome-inducing proxies of the same type, analogous 2SLS procedures have been shown to produce the correct conditional log-mean difference and log-odds ratios in specific SEMs (Liu et al., 2024).

Under local regularity conditions including the uniqueness of the bridge functions,

$$\mathcal{H} = \{h\} \quad \text{and} \quad \mathcal{Q} = \{q\},$$

guaranteed by completeness assumptions similar to those in assumption (3.7), the efficient influence function (EIF) of θ is (Cui et al., 2024; Kallus et al., 2022)

$$(2a-1)q(a,x,z)\{y-h(a,x,w)\}+h(1,x,w)-h(0,x,w)-\theta_0. \tag{3.8}$$

The proximal augmented inverse probability weighting estimator of θ based on the EIF is

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}\left[(2A_i-1)\hat{q}(A_i,X_i,Z_i)\{Y_i-\hat{h}(A_i,X_i,W_i)\}+\hat{h}(1,X_i,W_i)-\hat{h}(0,X_i,W_i)\right].$$

The function (3.8) is a doubly robust score in the sense that

$$E\{(2A-1)q^*(A,X,Z)\{Y-h^*(A,X,W)\}+h^*(1,X,W)-h^*(0,X,W)-\theta\}=0$$

if either $h^*=h$ or $q^*=q$. Hence, we should expect the estimator $\hat{\theta}$ to be robust against nuisance model misspecification. To establish the asymptotic normality of $\hat{\theta}$, we can exploit the rate robustness such that

$$\|T(\hat{h}-h)\|_P\|\hat{q}-q\|_P=o_P(n^{-1/2}), \tag{3.9}$$

where $\{T(f)\}(a,x,z)=E\{f(A,X,W)\mid A=a,X=x,Z=z\}$. The first factor can be replaced by the stronger version $\|\hat{h}-h\|_P$ without projection, but sufficiently fast convergence rates under this norm are more difficult to establish. Under some more regularity conditions including a Donsker class condition, $\hat{\theta}$ will be regular and locally efficient.

However, the ill-posedness of the integral equations in the definitions of $\mathcal{H}$ and $\mathcal{Q}$ creates challenges for the estimation of bridge functions. Furthermore, in the asymptotic analysis of estimators such as $\hat{\theta}_h$ and $\hat{\theta}_q$, standard arguments will not be applicable if $\hat{h}$ and $\hat{q}$ fail to converge to some fixed functions in the usual L_2-norm. A common remedy is regularization of the empirical risk minimization with a suitable loss function (Singh, 2023; Kompa et al., 2022; Mastouri et al., 2021; Kallus et al., 2022). In finite-dimensional linear classes without regularization, the minimax problems become equivalent to the estimating equations given in Cui et al. (2024). We refer to Zhang et al. (2023) and Bennett et al. (2023) for elaborate discussions and proposed solutions to sound statistical inference under ill-posedness.

3.3. Selection diagrams

In the rest of the chapter, we return to the data setup in Chapter 2 with the goal to estimate the TATE

$$\theta=E\{Y(1)-Y(0)\mid D=1\}.$$

In Manuscript I, we study anchored indirect comparison in the presence of unmeasured effect modification that threatens the validity of transportability of the treatment effect. For comparison between the proxies and assumptions in proximal causal inference for unmeasured confounding and the those used in Manuscript I for transportability, it is convenient to make use of selection diagrams.

The structural causal model (SCM) approach to causality offers a second viewpoint on the transportability of treatment effects (Bareinboim and Pearl, 2016). In this framework,

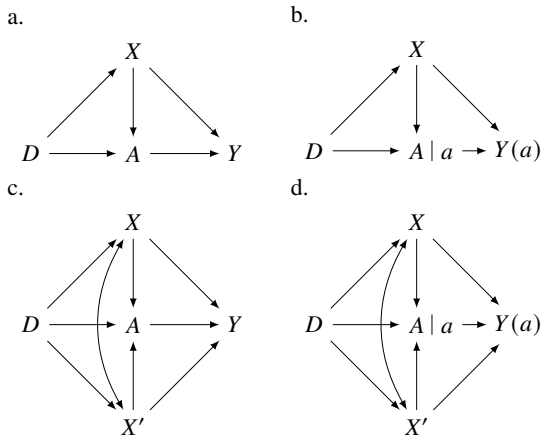

Figure 3.2. Selection diagrams (a and c) and the corresponding SWIGs (b and d) showing when the target-population average treatment effect is identified as (2.9) (a and b) and when identification does not rely on prognostic variables X' (c and d).

each structural equation can be translated into a receiving node and its incoming edges in a causal graph. Directed acyclic graphs (DAGs) are the most commonly used causal graphs, but acyclic directed mixed graphs are also convenient for encoding unobserved common causes between two variables.

Again, we focus on identifiability of the TATE, where we are concerned with transporting the effect from the source population to a different population rather than generalizing it to a super-population. This distinction allows us to represent aspects of the SCMs from two populations using a unified diagram, called the selection diagram (Pearl and Bareinboim, 2014; Bareinboim and Pearl, 2013). In a non-nested sampling scheme, the population is represented by a node D with edges pointing to all other variables whose structural equations differ between the populations or when the innovations differ in distributions. Figure 3.2 displays two selection diagrams.

Suppose we have a clear goal of estimating the TATE and have elicited a selection diagram from domain knowledge. To apply the potential outcome framework, we then use the selection diagram to justify the conditional distribution transportability (2.2), since DAGs do not encode structural assumptions apart from conditional independences. Typically, we turn the diagram into a single-world intervention graph (Richardson and Robins, 2013, SWIG) and check for the d-separation between the nodes $Y(a)$ and D; see Dahabreh et al. (2020b). If the underlying data generating mechanism satisfies the Markov property with respect to the selection diagram Figure 3.2a, validity of assumption (2.2) and thus identifiability of the TATE are established.

Identification in the SCM framework tends to rely on reading conditional independences off of manipulated selection diagrams. In Figure 3.2b, where assumption (2.2) may not hold due to the path $D \to X' \to Y$. If the structural equation of the outcome is such that for some functions f and g and an innovation ϵ,

$$
\begin{aligned}
&Y \leftarrow A f(X) + g(X, X') + \epsilon, \\
&\epsilon \perp\!\!\!\perp \{A, X, X', D\},
\end{aligned}
$$

the weaker transportability of the conditional average treatment effect (2.5) still holds,

and the identification formulas (2.9) and (2.10) apply.

3.4. Relations to unmeasured effect modifiers

The presence of shifted, unobserved shifted prognostic variables for the outcome does not invalidate identifiability of the TATE, as long as these variables are non-effect modifying in the ATE effect measure. However, if there exist shifted, unobserved effect modifiers U, identifiability of the TATE is lost. For the time being, we pretend to have measured U so that latent transportability and overlap conditions hold:

$$E\{Y(1) - Y(0) \mid X, U, D = 1\} = E\{Y(1) - Y(0) \mid X, U, D = 0\}; \tag{3.10}$$

$$\text{pr}(D = 0 \mid X, U) I\{\text{pr}(D = 1 \mid X, U) > 0\} > 0. \tag{3.11}$$

Additionally, we assume proper randomization of the treatment that protects against unmeasured confounders; that is,

$$A \perp\!\!\!\perp \{Y(1), Y(0), U\} \mid \{X, D = 0\},$$
$$0 < e_0(a \mid x) < 1, \quad x \in \mathcal{X}_0,$$

where $e_0(a \mid x) = \text{pr}(A = a \mid X = x, D = 0)$ is the propensity score in the source population. Treating e_0 as known, we can consider the transformed outcome

$$\tilde{Y} = \frac{2A - 1}{e_0(A \mid X)} Y$$

instead of the original outcome. This is because the relevant aspect of the outcome for the TATE is the CATE, which is exactly the conditional mean of the transformed outcome; that is,

$$E(\tilde{Y} \mid X, U, D = 0) = E(Y \mid A = 1, X, U, D = 0) - E(Y \mid A = 0, X, U, D = 0).$$

Latent identifiability of the TATE is then as follows:

$$\theta = E\{E(\tilde{Y} \mid X, U, D = 0) \mid D = 1\} = E\left\{ \frac{1 - D}{\alpha} \frac{\text{pr}(D = 1 \mid X, U)}{\text{pr}(D = 0 \mid X, U)} \tilde{Y} \right\}. \tag{3.12}$$

Without additional information nor reasonable assumptions, identification of θ in the observed data distribution is impossible.

In Manuscript I, we suggest a proximal causal inference approach to handle the existence of unobserved effect modifiers. Comparing the identification formula of the TATE (3.12) to that of the ATE (3.2), it is intuitive to consider two classes of bridge functions which can replace the nuisance parameters

$$E(\tilde{Y} \mid X, U, D = 0) \quad \text{and} \quad \frac{\text{pr}(D = 1 \mid X, U)}{\text{pr}(D = 0 \mid X, U)},$$

after being projected onto the space spanned by the effect modifiers $\{X, U\}$. We may posit the existence of latent bridge functions depending on two separate proxies W and Z:

$$\mathcal{H}_u = \left\{ h_u(x, w) : E\{h_u(X, W) \mid X, U, D = 0\} = E(\tilde{Y} \mid X, U, D = 0) \right\},$$
$$\mathcal{Q}_u = \left\{ q_u(x, z) : E\{q_u(X, Z) \mid X, U, D = 0\} = \frac{\text{pr}(D = 1 \mid X, U)}{\text{pr}(D = 0 \mid X, U)} \right\}.$$

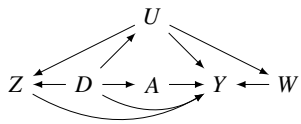

Figure 3.3. A DAG compatible with the definitions of proxies for transportability. The node X is not drawn but may point to any other node but D.

The definition of $\mathcal{H}_u$ does not concern the latent conditional outcome mean

$$E(Y \mid A, X, U, D = 0).$$

This aligns with assumption (3.10), in that the target population treatment-specific means $E\{Y(a) \mid D = 1\}$ are not latently identifiable, and we would generally not be interested in introducing unnecessary complexity to nuisance parameters not directly relevant to the TATE.

In the selection of proxies to combat unmeasured confounding, it is helpful to think about negative controls. When the goal is to capture undesired unmeasured effect modification whose magnitude varies between populations, some intuition can be gained through analogy. First, we would like the proxy W to be invariant in the sense that its distribution should not shift between populations once we control for the shifted effect modifiers $\{X, U\}$. Second, the proxy Z should not contribute to effect modification in the source population if the other effect modifiers are already conditioned on. Third, the proxies Z and W must be (conditionally) independent to disentangle the aspect of covariate shift and effect modification in relation to U. Now we formally state the definition of proxies used in Manuscript I. We require that in the source trial, a reweighting proxy Z and an adjustment proxy W can be obtained, while knowledge of the adjustment proxy suffices in the target trial. Thus, the observed data is an i.i.d. sample of $O = \{(1-D)Y, (1-D)A, X, (1-D)Z, W\}$. In addition to a extended version of randomization

$$A \perp\!\!\!\perp \{U, Z, W\} \mid \{X, D = 0\},$$

the proxies satisfy the conditional independences

$$\begin{aligned} Z &\perp\!\!\!\perp W \mid \{X, U, D = 0\}, \\ D &\perp\!\!\!\perp W \mid \{X, U\}, \end{aligned}$$

and the structural assumption

$$E(\tilde{Y} \mid Z, X, U, D = 0) = E(\tilde{Y} \mid X, U, D = 0).$$

Figure 3.3 displays a DAG compatible with the conditions for the proxies. Comparing Figures 3.1 and 3.3, we see that the edges $Z \to Y$ and $D \to Y$ are allowed to exist, since transportability of the CATE allows for the presence of shifted, unmeasured prognostic variables, and the reweighting proxy Z itself is also allowed to be a prognostic variable.

These conditions on the proxies give rise to the observed data bridge function classes

$$\begin{aligned} \mathcal{H} &= \left\{h(x, w) : E\{h(X, W) \mid X, Z, D = 0\} = E(\tilde{Y} \mid X, Z, D = 0)\right\}, \\ \mathcal{Q} &= \left\{q(x, z) : E\{q(X, Z) \mid X, W, D = 0\} = \frac{\mathrm{pr}(D = 1 \mid X, W)}{\mathrm{pr}(D = 0 \mid X, W)}\right\}. \end{aligned}$$

Under assumptions (3.4) and (3.5), the TATE is identifiable in the observed data distribution as

$$\theta = E\{h(X, W) \mid D = 1\}, \qquad \text{for any } h \in \mathcal{H},$$
$$\theta = E\left\{\frac{1 - D}{\alpha} q(X, Z)\tilde{Y}\right\}, \qquad \text{for any } q \in \mathcal{Q},$$

where $\alpha = \text{pr}(D = 1)$. It is now clear that the adjustment proxy W and the outcome bridge function h yield the proximal g-formula, and the reweighting proxy Z and the sampling odds bridge function q form the proximal sampling odds weighting formula. In Manuscript I, under local regularity conditions guaranteeing that $\mathcal{H}$ and $\mathcal{Q}$ are singletons, we show that an influence function of θ is

$$\frac{1 - d}{\alpha} q(x, z)\{\tilde{y} - h(x, w)\} + \frac{d}{\alpha}\{h(x, w) - \theta\}.$$

Denote the number of observations from the target population by n_1. Given estimators of the bridge functions, we propose the estimator

$$\hat{\theta} = \frac{1}{n_1} \sum_{i:D_i=0} \hat{q}(X_i, Z_i)\{\tilde{Y}_i - \hat{h}(X_i, W_i)\} + \frac{1}{n_1} \sum_{i:D_i=1} \hat{h}(W_i, X_i),$$

which is doubly robust against misspecification of the bridge functions and asymptotically linear under a rate condition similar to (3.9).

4. Multisource data fusion

4.1. Setup and identifiability

We continue to work with a non-nested design as in Chapter 2. To distinguish the current setup from the one previously discussed, we denote the target population by the indicator $G = 1$ and the source population with $G = 0$. In fact, the source population is an artificial super-population comprising a mixture of multiple separate source subpopulations. For simplicity, we work with two source subpopulations denoted by $D = 0, 1$, whereas the general setup is detailed in Manuscript II. In each source subpopulation, we collect a set of baseline covariates X, the treatment assigned A, and the outcome Y of each individual sampled. We only have access to the baseline covariates in the sample from the target population. Therefore, the joint sample from all environments can be conveniently viewed as an i.i.d. sample from the distribution over $O = \{(1-G)Y, (1-G)A, X, (1-G)D, G\}$. Importantly, we assume the outcome to be continuous and unbounded. A partial list of notations used in this chapter is given in Table 4.1 for easier reference.

Let $e_d(a \mid x) = \text{pr}(A = a \mid X = x, D = d)$ denote the propensity score in the source subpopulation d. Consistency, unconfoundedness, and positivity of treatment assignment are assumed to hold within each subpopulation:

$$(1-G)Y(A) = (1-G)Y; \tag{4.1}$$

$$0 < e_d(a \mid x) < 1, \quad x \in \mathcal{X}_d; \tag{4.2}$$

$$Y(a) \perp\!\!\!\perp A \mid \{X, D\}. \tag{4.3}$$

We are interested in the TATE

$$\theta = E\{Y(1) - Y(0) \mid G = 1\}.$$

Table 4.1. *A partial list of notations used in Chapter 4.*

Notation	Quantity
α	$\text{pr}(G = 1)$
$\pi(x)$	$\text{pr}(G = 1 \mid X = x)$
$\eta(d \mid x)$	$\text{pr}(D = d \mid X = x, G = 0)$
$\eta(d \mid a, x)$	$\text{pr}(D = d \mid A = a, X = x, G = 0)$
$e_d(a \mid x)$	$\text{pr}(A = a \mid X = x, D = d)$
$\mu_d(a, x)$	$E(Y \mid A = a, X = x, D = d)$
$e_\bullet(a \mid x)$	$\text{pr}(A = a \mid X = x, G = 0)$
$\mu_\bullet(a, x)$	$E(Y \mid A = a, X = x, G = 0)$
$V_d(a, x)$	$\text{var}(Y \mid A = a, X = x, D = d)$
$\mathcal{X}_0$	$\{x : \pi(x) < 1\}$
$\mathcal{X}_1$	$\{x : \pi(x) > 0\}$

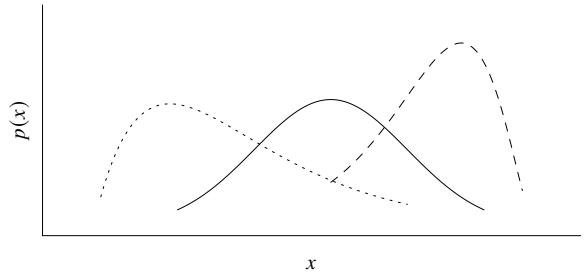

Figure 4.1. Densities p of a covariate within the target population and the source subpopulations. Overlap is fulfilled for the combined source population. Solid: target population density $p(x \mid G = 1)$; Dotted: source subpopulation 1 density $p(x \mid D = 0)$; Dashed: source subpopulation 2 density $p(x \mid D = 1)$.

Let $\pi(x) = \text{pr}(G = 1 \mid X = x)$ denote the conditional probability of being sampled into the target population. Identifiability of θ now hinges on the following overlap condition between the source and the target populations

$$\mathcal{X}_1 \subset \mathcal{X}_0, \tag{4.4}$$

where $\mathcal{X}_0$ and $\mathcal{X}_1$ are the supports of the baseline covariates X in the source and target populations. Although (4.4) appears to be identical to the previously used (2.1), with multisource data, the overlap condition here is much less severe in practice, since we do not require overlap for every pair of source subpopulation and the target population. This means that even if the target population contains very diverse strata, as long as the same participants could have appeared in any of the source subpopulations, the TATE can be recovered. With small source subpopulations, which can be study populations of small clinical trials, we can potentially approximate a large target population. See Figure 4.1 for an illustration.

Next is the transportability assumption. Consider transportability of the conditional potential outcome distribution, of the conditional treatment-specific means (CTSMs), and of the conditional average treatment effect (CATE) from any source subpopulation to the target population, respectively:

$$p\{y(a) \mid X = x, D = d\} = p\{y(a) \mid X = x, G = 1\}; \tag{4.5}$$

$$E\{Y(a) \mid X = x, D = d\} = E\{Y(a) \mid X = x, G = 1\}; \tag{4.6}$$

$$E\{Y(1) - Y(0) \mid X = x, D = d\} = E\{Y(1) - Y(0) \mid X = x, G = 1\} \tag{4.7}$$

for all (d, x) such that

$$\eta(d \mid x)\{\pi(1 - \pi)\}(x) > 0, \tag{4.8}$$

where $\eta(d \mid x) = \text{pr}(D = d \mid X = x, G = 0)$ is the conditional probability of being sampled from subpopulation d within the source population.

Let $\mu_d(a, x) = E(Y \mid A = a, X = x, D = d)$ denote the conditional outcome mean in source subpopulation d. Under assumptions (4.1)–(4.4) and one of assumptions (4.5)–(4.7) above, the TATE is identifiable in the observed data distribution as (Dahabreh

et al., 2023)

$$\theta = E\left[\sum_{d=0,1} \eta(d \mid X)\{\mu_d(1,X) - \mu_d(0,X)\} \,\middle|\, G = 1 \right] = E\left\{ \frac{1-G}{\alpha} \frac{\pi(X)}{1-\pi(X)} \frac{2A-1}{e_D(A \mid X)} Y \right\}. \tag{4.9}$$

In principle, this identifiability result also holds under an even weaker transportability assumption on the CATE like (2.5):

$$E\{Y(1) - Y(0) \mid X = x, G = 0\} = E\{Y(1) - Y(0) \mid X = x, G = 1\}, \quad x \in \mathcal{X}_1.$$

However, this assumption is hardly interpretable because it compares the CATE between the target population and the entire source population, the latter of which is a constructed population that mostly likely has no practical meaning. We will not make further use of this assumption.

4.2. Efficiency bounds in nested models

Under either of the two stronger transportability assumptions (4.5) and (4.6), we can further write the identfication formula as

$$\theta = E\{\mu_\bullet(1,X) - \mu_\bullet(0,X) \mid G = 1\} = E\left\{ \frac{1-G}{\alpha} \frac{\pi(X)}{1-\pi(X)} \frac{2A-1}{e_\bullet(A \mid X)} Y \right\}, \tag{4.10}$$

where $\mu_\bullet(a,x) = E(Y \mid A = a, X = x, G = 0)$ and $e_\bullet(a \mid x) = \mathrm{pr}(A = a \mid X = x, G = 0)$ are the conditional outcome mean and the propensity score within the whole source population. Comparing the identification formulas (4.9) and (4.10) here to (2.9) and (2.10), it is not hard to guess that

$$\frac{1-g}{\alpha} \frac{\pi(x)}{1-\pi(x)} \frac{2a-1}{e_d(a \mid x)} \{y - \mu_d(a,x)\} + \frac{g}{\alpha} \left[\sum_{d'=0,1} \eta(d \mid x)\{\mu_{d'}(1,x) - \mu_{d'}(0,x)\} - \theta \right] \tag{4.11}$$

and

$$\frac{1-g}{\alpha} \frac{\pi(x)}{1-\pi(x)} \frac{2a-1}{e_\bullet(a \mid x)} \{y - \mu_\bullet(a,x)\} + \frac{g}{\alpha} \{\mu_\bullet(1,x) - \mu_\bullet(0,x) - \theta\} \tag{4.12}$$

can be used to construct estimating equations for θ under the respective transportability assumptions. In fact, it can be verified that (4.11) and (4.12) are Neyman orthogonal scores with respect to all nuisance parameters. Therefore, we would expect the Z-estimators solving these estimating equations to have the rate robustness property and possibly robustness against model misspecifications given in Chapter 2. However, a sensible researcher should investigate whether such estimators make efficient use of the data according to the transportability assumptions.

If the observed data distribution is completely free to vary, the scores (4.11) and (4.12) should each be the efficient influence function (EIF) of θ in some sense. In

reality, when working with multisource data like in the current setup, the observed data distribution is rarely left unconstrained. Although transportability assumptions (4.5)–(4.7) are untestable, they may have testable implications in the observed data distribution. Within the intersection of the support of baseline covariates between the source subpopulations, we assume either identical conditional distributions, conditional outcome means, or conditional outcome mean differences; that is, for all x such that $\eta(1 \mid x)\eta(0 \mid x) > 0$:

$$Y \perp\!\!\!\perp D \mid \{A = a, X = x, G = 0\}; \tag{4.13}$$

$$\mu_0(a, x) = \mu_1(a, x); \tag{4.14}$$

$$\mu_0(1, x) - \mu_0(0, x) = \mu_1(1, x) - \mu_1(0, x). \tag{4.15}$$

To use the geometric approach to study the semiparametric efficiency bounds, we need to define the models so that obey the testable implications above. The models should be nested because the testable implications respect the logic:

$$\begin{array}{ccccccc} (4.13) & \Rightarrow & (4.14) & \Rightarrow & (4.15), \\ \mathcal{P}_{\ll} & \subset & \mathcal{P}_{<} & \subset & \mathcal{P}, \end{array}$$

where we use $\mathcal{P}$, $\mathcal{P}_{<}$, and $\mathcal{P}_{\ll}$ to denote the sets of probability distributions over O that satisfy (4.15), (4.14), and (4.13). The tangent spaces $\dot{\mathcal{P}}$, $\dot{\mathcal{P}}_{<}$, and $\dot{\mathcal{P}}_{\ll}$ for the nested models also have a nesting property:

$$\begin{array}{ccccccccc} & & & & \dot{\mathcal{P}} & \subset & L_2^0(P), & \text{for } P \in \mathcal{P}, \\ & & \dot{\mathcal{P}}_{<} & \subset & \dot{\mathcal{P}} & \subset & L_2^0(P), & \text{for } P \in \mathcal{P}_{<}, \\ \dot{\mathcal{P}}_{\ll} & \subset & \dot{\mathcal{P}}_{<} & \subset & \dot{\mathcal{P}} & \subset & L_2^0(P), & \text{for } P \in \mathcal{P}_{\ll}, \end{array}$$

where $L_2^0(P)$ is the Hilbert space of mean-zero square integrable functions with respect to the measure P.

We start with the largest model $\mathcal{P}$, which is also the closest to the nonparametric model with no restriction. Under some local regularity conditions, we show in Manuscript II that the EIF of θ in model $\mathcal{P}$ is given by

$$\begin{aligned} \varphi(o) = \frac{1-g}{\alpha} \frac{\pi(x)}{1-\pi(x)} \frac{w_d(x)}{\sum_{d'=0,1} \eta(d' \mid x) w_{d'}(x)} \frac{2a-1}{e_d(a \mid x)} \{y - \mu_d(a, x)\} \\ + \frac{g}{\alpha} \left[\sum_{d'=0,1} \eta(d \mid X) \{\mu_{d'}(1, x) - \mu_{d'}(0, x)\} - \theta \right] \end{aligned} \tag{4.16}$$

where

$$w_d(x) = \left\{ \sum_{a=0,1} \frac{V_d(a, x)}{e_d(a \mid x)} \right\}^{-1},$$

and $V_d(a, x) = \text{var}(Y \mid A = a, X = x, D = d)$ is the conditional outcome variance under treatment a in source subpopulation d. We refer to the function w_d as the optimal weight function. Replacing it in φ with other weight functions $\tilde{w}_d$ such that $E\{\tilde{w}_D(X) \mid X = x, G = 0\} \neq 0$ still produces valid influence functions (IFs). The score (4.11) is one such IF, and we denote it by ϕ. This is easily seen from the result in Manuscript II that the orthocomplement of the tangent space $\dot{\mathcal{P}}$ is

$$\dot{\mathcal{P}}^{\perp} = \left\{ (1-g) h(x, d) \frac{2a-1}{e_d(a \mid x, d)} \{y - \mu_d(a, x)\} : E\{h(X, D) \mid X, G = 0\} = 0 \right\}.$$

Hence, if we are to construct Z-estimators using the geometric approach, the efficient estimator should solve the estimating equation based on φ. In this case, we would weigh the nonparametric regression residual $Y - \mu_D(A, X)$ in a two-step procedure. The first step uses the propensity score within the source subpopulations to balance the over- and under-represented treatments, yielding

$$\tilde{\epsilon} = \frac{2A - 1}{e_D(A \mid X)}\{Y - \mu_D(A, X)\}.$$

The second step multiplies the transformed residual from the first step with a normalized inverse-variance weight,

$$\frac{\text{var}(\tilde{\epsilon} \mid X, D)}{\text{var}(\tilde{\epsilon} \mid X, G = 0)},$$

such that the conditional variance of the resulting quantity is minimized in the source population.

Now consider $P \in \mathcal{P}_< \subset \mathcal{P}$, then the pathwise derivative of θ along the submodel $\{P_\varepsilon\} \in \mathcal{P}_<$ with score $s_<(o) \in \dot{\mathcal{P}}_< \subset \dot{\mathcal{P}}$ such that $P_\varepsilon|_{\varepsilon=0} = P$ is

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta(P_\varepsilon)\bigg|_{\varepsilon=0} = E\{\varphi(O)s_<(O)\}.$$

Under mild regularity conditions, we show in Manuscript II that

$$\dot{\mathcal{P}}_<^\perp = \left\{(1-g)\frac{2a-1}{e_\bullet(a \mid x)}h(a, x, d)\{y - \mu_\bullet(a, x)\} : E\{h(A, X, D) \mid A, X, G = 0\} = 0\right\}.$$

Then the EIF in the smaller model $\mathcal{P}_<$ nested in larger model $\mathcal{P}$ can be obtained by the projection

$$\begin{aligned}\varphi_< &= \Pi\{\varphi \mid \dot{\mathcal{P}}_<\} = \varphi - \Pi\{\varphi \mid \dot{\mathcal{P}}_<^\perp\} \\ &= \frac{1-g}{\alpha}\frac{\pi(x)}{1-\pi(x)}\frac{2a-1}{e_\bullet(a \mid x)}\frac{V_d^{-1}(a, x)}{\sum_{d'=0,1} V_{d'}^{-1}(a, x)\eta(d' \mid a, x)}\{y - \mu_\bullet(a, x)\} \\ &\qquad + \frac{d}{\alpha}\{\mu_\bullet(1, x) - \mu_\bullet(0, x) - \theta\},\end{aligned}$$

where $\eta(d \mid a, x) = \text{pr}(D = d \mid A = a, X = x, G = 0)$ is the conditional probability of being sampled from the source subpopulation d among those receiving treatment a in the entire source population. Replacing $V_d^{-1}(a, x)$ with 1 in $\varphi_<$, we see that the score (4.12) is an IF of θ, and we denote it by $\phi_<$. The optimal weighting strategy suggested by $\varphi_<$ differs from that by φ. The residual is first weighted by the normalized inverse variance within each treatment group in the source population, and then balanced across the treatment groups. In the context of integrating external controls into randomized trials, Li et al. (2023) assume transportability of the conditional outcome mean under the control treatment. This results in a model similar to $\mathcal{P}_<$, and the EIF of the treatment-specific mean under control they consider resembles $\varphi_<$.

Finally, for $P \in \mathcal{P}_\ll$, we can obtain the EIF $\varphi_\ll$ of θ by projecting, for example, either φ or $\varphi_<$ onto the orthocomplement of the tangent space

$$\begin{aligned}\dot{\mathcal{P}}_\ll^\perp = \{(1-g)h(y, a, x, d) : &E\{h(Y, A, X, D) \mid A, X, D\} = 0, \\ &E\{h(Y, A, X, D) \mid Y, A, X, G = 0\} = 0\}.\end{aligned}$$

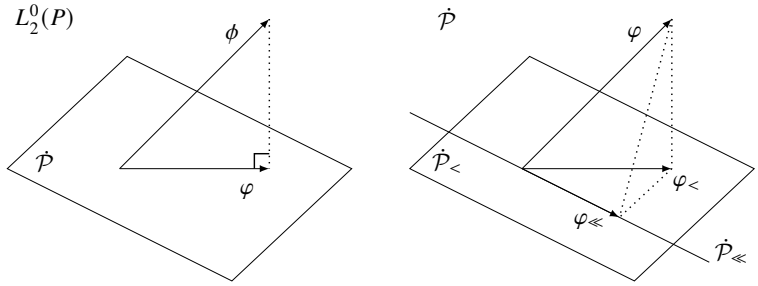

Figure 4.2. Relationship between the EIFs in different models. Left: $P \in \mathcal{P}$. Right: $P \in \mathcal{P}_\ll$. Adapted from Bickel et al. (1993, Figure 2).

Since $\phi_<$ is orthogonal to $\mathcal{P}_\ll^\perp$, it is exactly the EIF. Under model $\mathcal{P}_\ll$, Dahabreh et al. (2023); Wang et al. (2024) study the TATE in the whole target population and in subgroups of the target population. Their estimators are precisely based on the EIF $\varphi_\ll$.

In Figure 4.2, we illustrate the projection technique to find the EIF in each of the nested models by starting with the largest. In the left panel when $P \in \mathcal{P}$, the EIF φ is simply the projection of any IF $\phi \in L_2^0(P)$ onto the tangent space $\dot{\mathcal{P}}$, as discussed in Chapter 2. In the right panel when $P \in \mathcal{P}_\ll$, we may obtain the EIF $\varphi_\ll$ by successively projecting $\varphi \in \dot{\mathcal{P}}$ down to $\dot{\mathcal{P}}_<$, which yields φ_l, and then to $\dot{\mathcal{P}}_\ll$. A direct projection of any ϕ onto the tangent space $\dot{\mathcal{P}}_\ll$ also gives the EIF.

We conclude the section with several remarks. First, in the model $\mathcal{P}$, the estimand

$$E\{\mu_\bullet(1, X) - \mu_\bullet(0, X) \mid D = 1\}$$

is no longer necessarily the TATE (4.9). To see this, simply note that the difference between these estimands is

$$E\left[\sum_{a=0,1}\sum_{d=0,1}(2a-1)\{\eta(d \mid a, x) - \eta(d \mid x)\}\mu_d(a, x)\right].$$

However, if the propensity scores in the source subpopulations are equal,

$$e_1(a \mid x) = e_0(a \mid x),$$

the estimand above will again agree with the TATE. By including this additional restriction into the model $\mathcal{P}$, we can propose robust estimators without having to estimate $\mu_d(a, x)$ within each source subpopulation. However, estimators formed in this fashion will not be efficient because the EIF in this model is still φ. Second, when testable implications on the observed data distribution constrain the mean of the outcome, the EIFs in the models $\mathcal{P}$ and $\mathcal{P}_<$ involve the variance. Similarly, if homoscedasticity was included as part of the model restriction, the resulting EIF would involve the skewness of the outcome. In this respect, the pursuit of efficiency quickly becomes cumbersome when we move away from mean restrictions. Since efficient estimators would likely require the estimation of many nuisance parameters, the precision of estimates may not be better in finite samples.

4.3. Variational dependence and non-collapsibility

So far, we have seen that for an unbounded continuous outcome, transportability of the CATE (4.7) is in some sense the weakest assumption needed to identify the TATE. On one hand, it recognizes possible heterogeneity in the CTSMs by allowing them to differ freely among the populations. In other words, transportability of the CATE leaves the CTSMs variationally independent. Indeed, it is unnecessary to constrain the CTSMs between the source and target populations, when we are only interested in identifying the treatment effect and not the treatment-specific means (TSMs).

On the other hand, there is a neat correspondence between the CATE as a conditional effect and the average treatment effect (ATE) as a marginal effect. To identify the TATE, we simply standardize the CATE in the source population with respect to the baseline covariate distribution in the target population. The standardization step is intuitive and sane, because the target population marginal effect will be a weighted average of the conditional effects. This property of an effect measure is called collapsibility (Huitfeldt et al., 2019).

In the following, we discuss whether these two favorable properties can be maintained for a binary outcome. We follow the notations from previous sections, although most issues raised below occur irrespective of the multisource data setup.

When it comes to applying the identification formula (4.9), there is really no difference between a continuous and a binary outcome. However, transportability on the CATE (4.7) induces variational dependence between the TSMs when the outcome is binary. For example, suppose we observe, for some level of baseline covariates $x_0 \in \mathcal{X}_1$ in the target population, that in the source subpopulation d with $\eta(d \mid x_0) > 0$ the CATE is

$$E\{Y(1) \mid X = x_0, D = d\} - E\{Y(0) \mid X = x_0, D = d\} = -c_0,$$

where $0 \leq c_0 \leq 1$. Then the target population CTSMs must respect the following bounds:

$$\begin{aligned} c_0 \leq E\{Y(0) \mid X = x_0, G = 1\} \leq 1, \\ 0 \leq E\{Y(1) \mid X = x_0, G = 1\} \leq 1 - c_0. \end{aligned}$$

These additional constraints may be falsifiable with prior knowledge of the TSMs in the target population, but they are not verifiable. We should strive to avoid inducing unverifiable relations that are not directly assumed but fall out as ramifications of other assumptions. Moreover, if $\eta(0 \mid x_0)\eta(1 \mid x_0) > 0$, then there is also variation dependence between the conditional outcome means in the observed data distribution.

Because the odds maps a probability to an unbounded positive value, transportability of the conditional causal odds ratio

$$\begin{aligned} (4.17) \quad & \frac{E\{Y(1) \mid X = x, D = d\}}{1 - E\{Y(1) \mid X = x, D = d\}} \frac{1 - E\{Y(0) \mid X = x, D = d\}}{E\{Y(0) \mid X = x, D = d\}} \\ & = \frac{E\{Y(1) \mid X = x, G = 1\}}{1 - E\{Y(1) \mid X = x, G = 1\}} \frac{1 - E\{Y(0) \mid X = x, G = 1\}}{E\{Y(0) \mid X = x, G = 1\}}, \end{aligned}$$

for (d, x) such that (4.8) holds, guarantees variational independence of the CTSMs, as long as the conditional distributions of the potential outcomes are non-degenerate. The problem with the odds ratio is that it is a non-collapsible effect measure (Daniel et al.,

2021; Didelez and Stensrud, 2022). There is no standardization of the conditional causal odds ratios from the source population that would produce the target population causal odds ratio

$$\frac{E\{Y(1) \mid G=1\}}{1-E\{Y(1) \mid G=1\}} \frac{1-E\{Y(0) \mid G=1\}}{E\{Y(0) \mid G=1\}}.$$

Interpretation of this estimand as a weighted average of conditional effects is no longer possible. The target population causal odds ratio could even lie outside the range of conditional effects in the source population (Colnet et al., 2024b).

A separate issue that comes with a binary outcome is that unlike with a continuous outcome, the difference effect measure does not disentangle the treatment effect from the "baseline" conditional risk

$$E\{Y(0) \mid X=x, G=1\}.$$

The lack of disentanglement blurs the distinction between prognostic variables and effect modifiers, hence defying the intuition that shifted effect modifiers are in some sense sufficient for identification of the target population treatment effects defined in collapsible effect measures. Consider a data generating mechanism under which treatment 1 has the same beneficial effect for all subjects, but no harmful effect for anyone (Colnet et al., 2024b; Cinelli and Pearl, 2021). We assume that the resulting distribution admits a straightforward parametrization:

$$\begin{aligned} &\text{pr}\{Y(1)=0 \mid Y(0)=1, X=x, G=1\} = \text{constant}, \\ &\text{pr}\{Y(1)=1 \mid Y(0)=0, X=x, G=1\} = 0. \end{aligned} \tag{4.18}$$

For an individual with baseline characteristics x, the former expression is the probability of having no event under treatment 1, if the individual would suffer from the event under treatment 0. It is constant in x, corresponding to the assumption of no heterogeneity. The latter expression is the individual's risk under treatment 1, given that the individual would be event-free under treatment 0, which is exactly 0, because the treatment cannot harm. Assuming (4.18), the conditional effect

$$\begin{aligned} &\text{pr}\{Y(1)=1 \mid X=x, G=1\} - \text{pr}\{Y(0)=1 \mid X=x, G=1\} \\ &\qquad = -\text{pr}\{Y(1)=0 \mid Y(0)=1, X=x, G=1\}\text{pr}\{Y(0)=1 \mid X=x, G=1\} \end{aligned}$$

may still depend on the individual risk under treatment 0 that differ among populations, making (4.7) invalid. Therefore, transportability of the CATE fails to accommodate this simple setup.

5. Competing risks analysis

5.1. Setup and identifiability

In this chapter, we work with the setup in which the treatment-outcome information is available from the target population. Additionally, we have access to the outcome from a source population where everyone receives the control treatment. A sample from such a source population is called external controls. We discuss transportability assumptions under which the external controls are compatible with the target population sample, as well as how the external controls can be incorporated into the analysis to produce estimates of target population treatment effects with greater precision.

Unlike the setup in previous chapters, we now work with time-to-event outcomes. In Manuscript III, we are specifically concerned with competing risks analysis, where there exist other types of events that preclude the occurrence of the event of interest. For example, if the study population contains frail people, some may experience severe adverse events such as death, so the event of interest may never occur for these people. The event process can be captured by a Markov model where an individual starts from an event-free state and, at some point in time, transitions into one of the event states. A diagram illustrating this model is shown in Figure 5.1.

Time-to-event outcomes are susceptible to right censoring, meaning that the actual event may remain unobserved by being cut off prematurely. The source of censoring varies between studies. In randomized clinical trials, the observation period often ends after a preset number of events have accumulated in the sample, but some subjects may also leave the study or simply lost to follow-up. To focus on the presentation, we assume that there is no censoring. In Manuscript III, we handle censoring with the standard independent censoring assumption.

The outcome is a tuple of the time to event since their entry into the study T and the associated event type J. For simplicity, we assume that there are only two types of events, so that when the event of interest occurs, we observe $J = 1$, and when the competing event occurs, we observe $J = 2$. We also collect binary treatment A and some baseline covariates X. In the source population ($D = 0$), all individuals are exposed to treatment $A = 0$. Under non-nested sampling from the target and source populations, the data can be treated as an independent and identically distributed sample from the underlying observed data distribution over $O = (T, J, DA, X, D)$. A partial list of notations used in this chapter is given in Table 5.1.

Suppose we are interested in the τ-time target population causal cumulative incidence differences

$$\theta_j = \theta_{j1} - \theta_{j0},$$

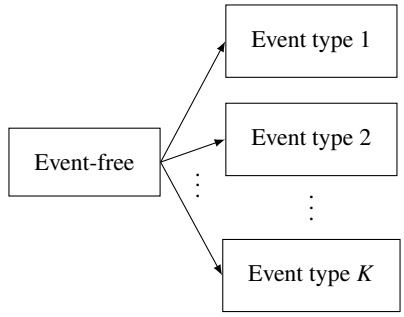

Figure 5.1. Competing risks model with K possible event types. Adapted from Figure I.3.5. in Andersen et al. (1993).

Table 5.1. *A partial list of notations used in Chapter 5.*

Notation	Quantity
α	$\text{pr}(D = 1)$
$\alpha_{1j}(t \mid a, x)$	(5.7)
$\alpha_{0j}(t \mid x)$	(5.8)
$e_1(a \mid x)$	$\text{pr}(A = a \mid X = x, D = 1)$
$F_{1j}(t \mid a, x)$	$\text{pr}(T \leq t, J = j \mid A = a, X = x, D = 1)$
$F_{0j}(t \mid x)$	$\text{pr}(T \leq t, J = j \mid X = x, D = 0)$
$M_{1j}(t \mid a, x)$	$I(T \leq t, J = j) - \int_0^t \text{I}(T \geq s)\alpha_{1j}(s \mid a, x)\text{d}s$
$\pi(x)$	$\text{pr}(D = 1 \mid X = x)$
$S_1(t \mid a, x)$	$\text{pr}(T > t \mid A = a, X = x, D = 1)$
$S_0(t \mid x)$	$\text{pr}(T > t \mid X = x, D = 0)$

where

$$\theta_{ja} = \text{pr}\{T(a) \leq \tau, J(a) = j \mid D = 1\}$$

are the target population treatment-specific cumulative incidences. In fact, these parameters are simply the target population average treatment effects (TATEs) and the target population treatment-specific means (TTSMs) for the binary outcomes $I\{T(a) \leq \tau, J(a) = j\}$, and we will continue to refer to them as such. Identifiability of the TATEs and the TTSMs depends on consistency in both populations, positivity of treatment assignment in the target population, and exchangeability of treatment assignment in the target population:

$$\{T(A), J(A)\} = (T, J); \tag{5.1}$$

$$0 < e_1(a \mid x) < 1; \tag{5.2}$$

$$\{T(a), J(a)\} \perp\!\!\!\perp A \mid \{X, D = 1\}. \tag{5.3}$$

Assumptions (5.2) and (5.3) are generally satisfied in (conditionally) randomized controlled trials. Then θ_{ja} is identifiable in the observed data distribution as

$$\theta_{ja} = E\{F_{1j}(\tau \mid a, X) \mid D = 1\},$$

where $F_{1j}(t \mid a, x) = \text{pr}(T \leq t, J = j \mid A = a, X = x, D = 1)$ is the cause j conditional cumulative incidence function under treatment a in the target population.

5.2. Two transportability assumptions

Incorporation of the external controls into the target population sample may be licensed by transportability assumptions. In this section, we discuss two such assumptions that relate the distribution of the potential event time and potential event type under the control treatment in the source population to that in the target population.

In the setting without competing risks, Lee et al. (2022, 2024) generalize and transport treatment effects in survival analysis with one type of event from a source population to a target population. They work under transportability of the conditional distributions of the potential event times. In the current setup, an analogous assumption is the following:

$$\{T(0), J(0)\} \perp\!\!\!\perp D \mid X. \tag{5.4}$$

In words, this assumption states that the conditional distribution of the potential time to event is transportable given the baseline covariates. Equivalently, we can represent assumption (5.4) as

$$\alpha^0_{0j}(t \mid x) = \alpha^0_{1j}(t \mid x), \quad \text{for } x \in \mathcal{X}_1 \cap \mathcal{X}_0,$$

where

$$\alpha^0_{dj}(t \mid x) = \lim_{\triangle t \downarrow 0} \frac{\text{pr}\{T(0) \in [t, t+\triangle t), J(0) = j \mid T(0) \geq t, X = x, D = d\}}{\triangle t},$$

denotes the conditional cause j hazard of the potential outcome under the control treatment within the target or source population. It requires that between the populations, all shifted prognostic variables for the outcome are contained in the set of baseline covariates X. In some situations, the hazards of the event of interest α^0_{d1} may be aligned between the populations given the observed covariates, but the hazards of the competing event α^0_{d2} may differ. For example, suppose the use of tobacco and the exposure to ultraviolet radiation vary between the populations, but in the data, we only collect smoking habits of study participants. Suppose the competing risks are different causes of death. Then the hazard of cardiovascular death may be transportable conditioning on the frequency of smoking, whereas the hazard of death from skin cancer is much less likely to be transportable without controlling for solarium visits or the use of sunscreen. In such cases, we may consider transportability of the conditional cause 1 hazard under the control treatment

$$\alpha^0_{01}(t \mid x) = \alpha^0_{11}(t \mid x), \quad \text{for } x \in \mathcal{X}_1 \cap \mathcal{X}_0, \tag{5.5}$$

while the conditional cause 2 hazard can differ arbitrarily.

We can also illustrate the difference between assumptions (5.4) and (5.5) in causal diagrams. Recall from Chapter 3 that selection diagrams encode covariate shifts by using an edge from the population indicator to every shifted covariate. Viewing the event time and the type of event $\{T, J\}$ as a single node, we can reproduce a directed acyclic graph and a single-world intervention graph where assumption (5.4) holds, resembling those given in Figure 3.2. However, assumption (5.5) cannot be depicted in a similar fashion, since d-separation implies only conditional independence, while this assumption involves hazards of only one cause of event.

It turns out that (5.5) is an instance of local independence and can thus be represented in a local independence graph (Didelez, 2008; Røysland et al., 2025). One way of

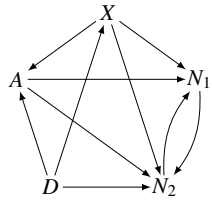

Figure 5.2. A local independence graph compatible with assumption (5.5).

representing the time-to-event outcome is to note whether an event of type j has occurred before time t in the observation period $\mathcal{T} = [0, \tau]$ in the form of

$$N_j(t) = I(T \leq t, J = j), \quad t \in \mathcal{T}.$$

These stochastic processes indexed by time are examples of counting processes. Given assumption (5.3), assumption (5.5) is equivalent to that the counting process N_1 is locally independent of D given $\{N_1, N_2, A = 0, X\}$, written as

$$D \nrightarrow N_1 \mid \{N_1, N_2, A = 0, X\}. \tag{5.6}$$

This local independence means that the intensity of the counting process N_1 with respect to the filtration generated by $\{N_1, N_2, A = 0, X, D\}$ is the same as the intensity generated without D. Informally, consider two individuals with the same baseline covariates under treatment 0, one from the source population and the other from the target population. If neither of them has experienced any event by time t, the instantaneous rate of the event of interest is the same for both, which is exactly the interpretation of assumption (5.5). Graphically, (5.6) can be represented by the so-called δ-separation. In the local independence graph from Figure 5.2, D can be δ-separated from N_1 because all trails (here the same as paths) from D to N_1 are blocked by the set of nodes $\{N_2, A, X\}$. The edge $D \rightarrow N_2$ means that there may be unobserved shifted prognostic variables U with $D \rightarrow U \rightarrow N_2$. If there was no edge between D and N_2, the graph would also be compatible with assumption (5.4).

In Manuscript III, we briefly discuss transportability assumptions on the τ-time conditional causal cumulative incidences, on the subdistribution functions, and on the all-cause survival functions. These assumptions are less intuitive than assumptions (5.4) and (5.5), since they involve the conditional hazards of both event types in specific functional forms.

5.3. Efficiency in three models

Intuitively, the stronger the transportability assumption is, the more compatible the data generating mechanisms will be between the target and source populations, and the higher potential precision gain for the estimates of the TATEs and TTSMs. Given a transportability assumption, it is of interest to propose efficient estimators that make optimal use of external controls to achieve the maximum precision gain. In this section, we present the efficient influence functions (EIFs) of the TATEs and the TTSMs under three nested models. Under suitable regularity conditions, Z-estimators solving the estimating equations formed by the EIFs will be efficient in the respective models.

We adopt a similar style to motivate the efficiency calculations as in Chapter 4. Let the observed data conditional cause-specific hazards in the target and source distributions be

$$\alpha_{1j}(t \mid a,x) = \lim_{\triangle t \downarrow 0} \frac{\text{pr}\{T \in [t, t+\triangle t), J = j \mid T \geq t, A = a, X = x, D = 1\}}{\triangle t}, \tag{5.7}$$

$$\alpha_{0j}(t \mid x) = \lim_{\triangle t \downarrow 0} \frac{\text{pr}\{T \in [t, t+\triangle t), J = j \mid T \geq t, X = x, D = 0\}}{\triangle t}. \tag{5.8}$$

Then the corresponding observed data conditional all-cause survival functions and conditional cumulative incidence functions are

$$S_1(t \mid a,x) = \text{pr}(T > t \mid A = a, X = x, D = 1) = \exp\left\{ -\int_0^t (\alpha_{11} + \alpha_{12})(s \mid a,x)\text{d}s \right\},$$

$$S_0(t \mid x) = \text{pr}(T > t \mid X = x, D = 0) = \exp\left\{ -\int_0^t (\alpha_{01} + \alpha_{02})(s \mid x)\text{d}s \right\},$$

$$F_{1j}(t \mid a,x) = \text{pr}(T \leq t, J = j \mid A = a, X = x, D = 1) = \int_0^t S_1(s \mid a,x)\alpha_{1j}(s \mid a,x)\text{d}s,$$

$$F_{0j}(t \mid x) = \text{pr}(T \leq t, J = j \mid X = x, D = 0) = \int_0^t S_0(s \mid x)\alpha_{0j}(s \mid x)\text{d}s.$$

Under assumptions (5.1)–(5.3), the transportability assumptions (5.4) and (5.5) have the following testable implications in the observed data distribution:

$$\{T, J\} \perp\!\!\!\perp D \mid \{A = 0, X\}; \tag{5.9}$$

$$\alpha_{11}(t \mid 0,x) = \alpha_{01}(t \mid x). \tag{5.10}$$

Then we can define three nested models

$$\mathcal{P}_{\ll} \subset \mathcal{P}_{<} \subset \mathcal{P},$$

where the models are sets of probability distributions over O that satisfy the conditional independence (5.9), the local independence (5.10), and no special constraints, from the smallest to the largest.

Suppose the underlying distribution P stems from the unrestricted model $\mathcal{P}$. Then the tangent space is exactly the entire Hilbert space of square integrable mean-zero functions $L_2^0(P)$. The EIF of the TTSM $\theta_{ja'}$ for treatment a' and cause j is (Rytgaard et al., 2023)

$$\varphi_{ja'}(o) = \frac{d}{\alpha}\frac{I(a = a')}{e_1(a \mid x)} \int_0^\tau \frac{g_{j1}(t,a,x)\text{d}m_{11}(t \mid a,x) + g_{j2}(t,a,x)\text{d}m_{12}(t \mid a,x)}{S_1(t \mid a,x)} + \frac{d}{\alpha}\{F_{1j}(t \mid a',x) - \theta_{ja'}\},$$

where

$$g_{jk}(t,a,x) = I(j = k)S_1(t \mid a,x) - \{F_{1j}(\tau \mid a,x) - F_{1j}(t \mid a,x)\},$$

and $m_{1j}(t \mid a,x)$ are realizations of $M_{1j}(t \mid A,X) = N_j(t) - \int_0^t I(T \geq s)\alpha_{1j}(s \mid A,X)\text{d}s$. In fact, $DM_{1j}(t \mid A,X)$ is a martingale with respect to the filtration representing

the history of events as time passes as well as the information available at baseline. Martingales may be viewed as residuals from nonparametric regressions, so it should be rather intuitive that martingale integrals, such as those appearing in $\varphi_{ja'}$, characterize the form of the tangent spaces associated with the counting processes. Another structure that often arises in the derivation of EIFs of causal parameters with time-to-event data is the weight composed of the propensity score and the at-risk (survival) probability. The weight is often self-normalizing in the sense that it has mean 1. For example, writing out the at-risk indicator $I(T \geq t)$ in the martingale increment, the time-varying weight is

$$\frac{DI(A = a')I(T \geq t)}{\alpha e_1(A \mid X)S_1(t \mid A, X)}.$$

Finally, the functions g_{1j} are tied to the parameters of interest, which in this case are the TTSMs. Because the EIF is multiplied by the indicator of being in the target population, the Z-estimator based on $\varphi_{ja'}$ does not use any external controls.

In the sequel, we only discuss the TTSMs θ_{j0} under the control treatment, since the efficiency bounds of θ_{j1} should not change when there is no restriction on the event time and event type distribution under treatment 1 in the target population. Consider first the model $\mathcal{P}_<$ where the counting process of the event of interest is locally independent of the population given the history under the control treatment. The EIF of θ_{j0} can be found by subtracting from φ_{j0} its projection onto the orthocomplement of the tangent space $\dot{\mathcal{P}}_<$:

$$\dot{\mathcal{P}}_<^{\perp} = \left\{(1-a)\int_0^\tau h_1(t,x,d)\mathrm{d}m_{11}(t \mid 0,x) : E\{h_1(t,X,D) \mid T \geq t, X\} = 0\right\}.$$

Let $\pi(x) = \mathrm{pr}(D = 1 \mid X = x)$ be the probability of being sampled from the target population. The EIF of θ_{j0} is

$$\varphi_{j0,<}(o) = \frac{\pi(x)}{\alpha}(1-a)\int_0^\tau \frac{g_{j1}(t,0,x)\mathrm{d}m_{11}(t \mid 0,x)}{\pi(x)e_1(0 \mid x)S_1(t \mid 0,x) + \{1-\pi(x)\}S_0(t \mid x)}$$
$$\frac{d}{\alpha}\frac{1-a}{e_1(0 \mid x)}\int_0^\tau \frac{g_{j2}(t,a,x)}{S_1(t \mid a,x)}\mathrm{d}m_{12}(t \mid a,x) + \frac{d}{\alpha}\{F_{1j}(t \mid 0,x) - \theta_{j0}\}.$$

Comparing $\varphi_{j0,<}$ to φ_{j0}, it is clear that the martingale integral over $m_{11}(t \mid 0, X)$ now involves observations under treatment 0 from both populations, but the martingale integral over $m_{12}(t \mid 0, X)$ remains unchanged. Accordingly, the weight

$$\frac{\pi(X)}{\alpha}\frac{(1-A)I(T \geq t)}{\pi(X)e_1(0 \mid X)S_1(t \mid 0,X) + \{1-\pi(X)\}S_0(t \mid X)} \tag{5.11}$$

first rescales the residual of the counting process N_1 with the inverse joint probability of surviving until time t and receiving the control treatment

$$\mathrm{pr}(T \geq t, A = 0 \mid X = x)$$

and then standardizes it over the distribution of the baseline covariates in the target population. Previously in Chapter 4, we have seen that for an ordinary continuous outcome, the restriction on the conditional outcome means leads to an EIF involving the conditional variance of the outcome (Li et al., 2023). Here, since the counting process

increment $\mathrm{d}N_1(t)$ is binary and completely characterized by the hazards, the "conditional variances" are also equal and therefore do not appear in (5.11). Heuristically, precision gain via external control data is only possible when there is an overlap in the distribution of baseline covariates; that is, there should at least exist non-null sets of baseline covariates where

$$\pi(x)\{1-\pi(x)\} > 0.$$

Finally, the orthocomplement of the tangent space $\dot{\mathcal{P}}_{\ll}$ of the model $\mathcal{P}_{\ll}$ is

$$\dot{\mathcal{P}}_{\ll} = \left\{(1-a)\int_0^\tau \{h_1(t,x)\mathrm{d}m_{11}(t \mid 0,x) + h_2(t,x)\mathrm{d}m_{12}(t \mid 0,x)\} : \right.$$
$$\left. E\{h_j(t,X) \mid T \geq t, A=0, X\} = 0, j = 1,2\right\}.$$

Hence, it is easily seen that the EIF of θ_{j0} in the model $\mathcal{P}_{\ll}$ is

$$\varphi_{j0,\ll}(o) = \frac{\pi(x)}{\alpha}(1-a)\int_0^\tau \frac{g_{j1}(t,0,x)\mathrm{d}m_{11}(t \mid 0,x) + g_{j2}(t,0,x)\mathrm{d}m_{12}(t \mid 0,x)}{\pi(x)e_1(0 \mid x)S_1(t \mid 0,x) + \{1-\pi(x)\}S_0(t \mid x)}$$
$$+ \frac{d}{\alpha}\{F_{1j}(t \mid 0,x) - \theta_{j0}\}.$$

The difference from $\varphi_{j0,<}$ is that now the martingale integral over $m_{j2}(t \mid 0,x)$ also incorporates observations from the external control population. This corresponds to the fact that (5.9) is (5.10) plus the restriction on the cause 2 hazards

$$\alpha_{12}(t \mid 0,x) = \alpha_{02}(t \mid x).$$

In Manuscript III, we consider event time distributions that may not be absolutely continuous and are allowed to have masses on a countable set of timepoints. The more general setup also accommodates discrete-time competing risks analysis, where the martingale integrals in the EIFs reduce to summations over timesteps. Since the transportability assumptions (5.4) and (5.5) are not specific to any effect measure, other estimands in competing risks analysis may also be adopted. Apart from the τ-time cumulative incidence differences, we also consider the differences in τ-time restricted mean times lost (Andersen, 2013)

$$E(I\{J(1)=j\}[\tau - \{T(1)\wedge\tau\}] \mid D=1) - E(I\{J(0)=j\}[\tau - \{T(0)\wedge\tau\}] \mid D=1)$$

as an alternative effect measure. The restricted mean time lost to cause j is a generalization of the restricted mean survival time $E\{T(a)\wedge\tau \mid D=1\}$ in the absence of competing risks. Its interpretation follows directly from the decomposition

$$\tau - E\{T(a)\wedge\tau \mid D=1\} = \sum_{j=1,2} E(I\{J(a)=j\}[\tau - \{T(a)\wedge\tau\}] \mid D=1).$$

The efficiency calculations for cumulative incidences apply directly to restricted mean times lost, because the latter estimands are simply integrals of the former.

6. Summary of manuscripts

6.1. Manuscript I

In Manuscript I, we provide a new method for indirect comparison of two treatments, when direct evidence of the treatment effect is unavailable, but the treatments are researched separately in two randomized controlled trials (RCTs) in possibly different populations. We are in interested in the average treatment effect (ATE) comparing these treatments in the target population, which is the study population from one of the two RCTs. We consider the anchored indirect comparison setting, where the RCTs share a common treatment. To establish the desired treatment effect, current methods rely on transportability of the conditional average treatment effect (CATE) between the source and target RCTs. However, if some shifted effect modifiers are unmeasured, the transportability assumption will be violated. Thus, application of the estimators from Chapter 2 adjusting for observed covariates may lead to biased results.

Through the use of proxy variables, we establish a new proximal identification result that relies on the existence of the so-called bridge functions. We assume that transportability of the CATE is restored when further conditioning on the unobserved effect modifiers. We require that a pair of proxy variables are collected in the source RCT, whereas only one of the two needs to be collected in the target RCT. As explained in Chapter 3, the bridge functions are functions of observed variables only, but they mimic the underlying nuisance functions containing unobserved variables. For estimation, we propose a doubly robust estimator of the target population average treatment effect (TATE), which depends on two bridge functions.

Using data from the STEP-2 and SCALE clinical trials, we estimated the weight loss effect, comparing once-weekly semaglutide 2.4 mg and once-daily liraglutide 3.0 mg at week 44 among the study population of STEP-2. The outcome was the percentage change of body weight from baseline to week 44. The two groups of proxy variables were lab measurements of blood sugar levels and lipid levels. A complete-case analysis of the data with the proximal estimator demonstrateed a significant effect at -3.82 percentage points [95%-confidence interval (CI): $(-4.73, -2.90)$] in favor of semaglutide. Applying an estimator that did not account for unobserved effect modifiers gave -3.80 percentage points [95%-CI: $(-4.59, -3.01)$]. The negligible difference between the estimates may have been a consequence of insufficiently informative proxies.

6.2. Manuscript II

In Manuscript II, we consider efficient estimation of the TATE with multisource data. By "multisource", we mean that treatment and outcome information is available from

more than one data source, but the information is unavailable in the target population. Normally, when only a single source population contains this information, identifiability of the TATE requires that the baseline characteristics of the individuals in this population be as diverse as the characteristics in the target population. Using several RCTs, this overlap condition is instead required between the combined source subpopulations and the target population. To identify the TATE, we assume transportability of the CATE across data sources, including the target population. The effect-measure transportability holds when all shifted effect modifiers are collected.

This transportability assumption has testable implications that place restrictions on the observed data distributions. By studying the space of influence functions of the TATE, we propose a class of doubly robust estimators. We also provide a construction for an efficient estimator that hinges on the estimation of the optimal weight function; see also Chapter 4. In addition to the TATE, we consider the target population projected CATE, which is defined as the best linear approximation of the CATE in the target population by a pre-specified finite-dimensional basis of the effect modifiers. We propose estimators for the projected CATE with associated pointwise and uniform confidence intervals.

To illustrate the method, we split the STEP-1 clinical trial into four regions: the United States (US), the United Kingdom, as well as countries of continental Europe and East Asia. We used treatment and outcome information from the three regions excluding the US to estimate the weight loss effect comparing once-weekly semaglutide 2.4 mg against placebo in the study population within the US. The outcome was the percentage change in body weight from baseline to week 68. A complete-case analysis with the estimator using the optimal weight function yielded a TATE estimate of -12.57 percentage points [95% CI: $(-14.03, -11.10)$]. In this data example, we also computed the ATE in the US study population with an augmented inverse probability estimator, which led to an effect of -13.21 percentage points [95% CI: $(-14.50, -11.92)$]. The discrepancy between these estimates might have been due to unmeasured effect modifiers, suggesting the need for sensitivity analysis.

6.3. Manuscript III

In Manuscript III, the aim is to improve the precision of cumulative incidence estimates by incorporating external controls into an RCT where competing risks are present. For example, mortality is a common competing risk that precludes the event of interest. We utilize external control data by assuming that the conditional hazard of the event of interest under the control treatment is transportable between the external population and the RCT study population. This transportability assumption holds even when there exist unobserved, shifted prognostic variables that directly alter the hazard of the competing risks. In other words, we allow the observed conditional hazards of the competing risks to differ between the two populations; see Chapter 5 for details.

The target parameters are the causal (or standardized) cumulative incidence differences in the RCT. Under assumed transportability, we derive the efficient influence functions (EIFs) of the target parameters. The structure of the EIFs prompts the construction of efficient estimators which turn out to be triply robust against model misspecification. Specifically, the external controls are integrated into these estimators through martingale integrals associated with the event of interest. Moreover, in contrast to existing fusion estimators for an ordinary continuous outcome, the samples receive time-varying

weights related to the inverse probability of being at risk in the pooled population.

To illustrate the estimators, we re-analyzed data from the LEADER and SUSTAIN-6 clinical trials. The event of interest was defined as the composite event of nonfatal myocardial infarction and nonfatal stroke, and the competing event was the all-cause mortality. Controls from LEADER were used as external controls to estimate the causal cumulative incidence differences in SUSTAIN-6 at weeks 26, 52, 78, and 104, comparing once-weekly semaglutide 1.0 mg to placebo. See the manuscript for the full display of results. The point estimates computed with external controls were fairly close to those computed without them. The fusion estimates had standard errors around 9% smaller than the RCT-only estimators.

7. Discussion and outlook

Sensitivity analysis for transportability. When the outcome is not observed in the target population, correct identification of target population treatment effects often depend on the validity of transportability assumptions. The concern for the violation of such assumptions is partially addressed in Manuscript I, where we allow for the existence of unobserved, shifted effect modifiers. Though meant for point identification, the proximal estimators can also be used in sensitivity analyses for unmeasured effect modifiers based on proxies. We may alter the proxies of each type to produce a range of possible target population treatment effects. Current sensitivity analysis methods for transportability borrow ideas from sensitivity tools for unmeasured confounding, including parametric bias functions (Nguyen et al., 2017), the exponential tilting model (Dahabreh et al., 2022), marginal sensitivity models (Nie et al., 2021), and the omitted variable bias framework (Huang, 2024). Further research may extend recent developments for unmeasured confounding to transportability bias, such as doubly sharp and valid partial identification (Dorn et al., 2024) and nonparametric robustness values (Chernozhukov et al., 2024).

Bias-robust data combination. When transportability assumptions have testable implications in the observed data distribution, these implications can be used to falsify the assumptions. In Manuscripts II and III, we do not make such considerations, but if the transportability assumptions fail to hold, incorporating external controls may introduce substantial bias to the estimators. Especially when the goal is to improve precision of estimates, the reference estimator without data combination may not be efficient but is guaranteed to be bias-free. The bias-variance trade-off between the reference estimator and the possibly biased but efficient estimator can be handled, for example, by linear interpolation (Oberst et al., 2023), a test-then-pool procedure (Yang et al., 2023), or sieve model selection (van der Laan et al., 2025). These methods easily adapt to the settings of Manuscripts II and III. An alternative approach is automatic bias detection and data fusion on a subset of data (Gao et al., 2024).

Optimal adjustment set for transportability. It is well-known that the average treatment effect is overidentified in randomized controlled trials when baseline covariates are observed. Thus, valid estimators can be formed both with and without adjustment of baseline covariates (Yang and Tsiatis, 2001). In the model consisting of distributions compatible with a directed acyclic graph (DAG), there exists a graphical criterion for identifying the optimal adjustment set, if the average treatment effect is identifiable (Henckel et al., 2022; Rotnitzky and Smucler, 2020). For augmented inverse probability weighting estimators that are also asymptotically linear, the asymptotic variance of the estimator adjusting for the optimal set lower-bounds the asymptotic variance of estimators adjusting for other sets. In certain DAGs, the optimal adjustment set characterizes

the semiparametric efficiency bound under the aforementioned model. In transportability of causal effects, similar optimality results are not yet known. Since DAGs do not encode assumptions from structural equations, extensions on existing graphical models may be necessary to capture the effect measure in question. In a class of reweighting estimators for the target-population average treatment effect, Colnet et al. (2024a) note that controlling for shifted but non-effect-modifying covariates increases estimator variance.

Bibliography

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Andersen, P. K. (2013). Decomposition of number of life years lost according to causes of death. *Statistics in Medicine*, 32(30):5278–5285.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. Springer.

Andrews, D. W. K. (2017). Examples of L^2-complete and boundedly-complete distributions. *Journal of Econometrics*, 199(2):213–220.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85.

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

Bareinboim, E. and Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, 1(1):107–134.

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

Bennett, A., Kallus, N., Mao, X., Newey, W., Syrgkanis, V., and Uehara, M. (2023). Inference on strongly identified functionals of weakly identified functions. *arXiv*:2208.08291v3.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Springer.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(4):1193–1209.

Burgess, S., Small, D. S., and Thompson, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, 26(5):2333–2355.

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

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., and Syrgkanis, V. (2024). Long story short: Omitted variable bias in causal machine learning. *arXiv*:2112.13398v5.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022a). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022b). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

Cinelli, C. and Pearl, J. (2021). Generalizing experimental results by leveraging knowledge of mechanisms. *European Journal of Epidemiology*, 36(2):149–164.

Cole, S. R., Edwards, J. K., Breskin, A., Rosin, S., Zivich, P. N., Shook-Sa, B. E., and Hudgens, M. G. (2023). Illustration of 2 fusion designs and estimators. *American Journal of*

Epidemiology, 192(3):467–474.

Cole, S. R. and Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, 172(1):107–115.

Colnet, B., Josse, J., Varoquaux, G., and Scornet, E. (2022). Causal effect on a target population: A sensitivity analysis to handle missing covariates. *Journal of Causal Inference*, 10(1):372–414.

Colnet, B., Josse, J., Varoquaux, G., and Scornet, E. (2024a). Re-weighting the randomized controlled trial for generalization: Finite-sample error and variable selection. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnae043.

Colnet, B., Josse, J., Varoquaux, G., and Scornet, E. (2024b). Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize? *arXiv*:2303.16008v3.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2024c). Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1):165–191.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2024). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359.

Dahabreh, I. J. (2024). Invited commentary: Combining information to answer epidemiologic questions about a target population. *American Journal of Epidemiology*, 193(5):741–750.

Dahabreh, I. J., Haneuse, S. J.-P. A., Robins, J. M., Robertson, S. E., Buchanan, A. L., Stuart, E. A., and Hernán, M. A. (2021). Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology*, 190(8):1632–1642.

Dahabreh, I. J., Robertson, S. E., Petito, L. C., Hernán, M. A., and Steingrimsson, J. A. (2023). Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics*, 79(2):1057–1072.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernán, M. A. (2020a). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39(14):1999–2014.

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

Dahabreh, I. J., Robins, J. M., Haneuse, S. J.-P. A., Robertson, S. E., Steingrimsson, J. A., and Hernán, M. A. (2022). Global sensitivity analysis for studies extending inferences from a randomized trial to a target population. *arXiv*:2207.09982v1.

Dahabreh, I. J., Robins, J. M., and Hernán, M. A. (2020b). Benchmarking observational methods by comparing randomized trials and their emulations. *Epidemiology*, 31(5):614.

Daniel, R., Zhang, J., and Farewell, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, 63(3):528–557.

Davi, R., Mahendraratnam, N., Chatterjee, A., Dawson, C. J., and Sherman, R. (2020). Informing single-arm clinical trials with external controls. *Nature Reviews Drug Discovery*, 19(12):821–822.

Davies, M., Færch, L., Jeppesen, O. K., Pakseresht, A., Pedersen, S. D., Perreault, L., Rosenstock, J., Shimomura, I., Viljoen, A., Wadden, T. A., and Lingvay, I. (2021). Semaglutide 2.4 mg once a week in adults with overweight or obesity, and type 2 diabetes (STEP 2): A randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. *The Lancet*, 397(10278):971–984.

Davies, M. J., Bergenstal, R., Bode, B., Kushner, R. F., Lewin, A., Skjøth, T. V., Andreasen, A. H., Jensen, C. B., DeFronzo, R. A., and for the NN8022-1922 Study Group (2015). Efficacy of liraglutide for weight loss among patients with type 2 diabetes: The SCALE diabetes randomized clinical trial. *JAMA*, 314(7):687–699.

Degtiar, I. and Rose, S. (2023). A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10(1):501–524.

D'Haultfœuille, X. (2011). On the completeness condition in nonparametric instrumental prob-

lems. *Econometric Theory*, 27(3):460–471.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):245–264.

Didelez, V. and Stensrud, M. J. (2022). On the logic of collapsibility for causal effect measures. *Biometrical Journal*, 64(2):235–242.

Dorn, J., Guo, K., and Kallus, N. (2024). Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*. Advance online publication.

Gao, C., Yang, S., Shan, M., Y E, W., Lipkovich, I., and Faries, D. (2024). Improving randomized controlled trial analysis via data-adaptive borrowing. *Biometrika*. In press.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Henckel, L., Perković, E., and Maathuis, M. H. (2022). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hotz, V. J., Imbens, G. W., and Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1):241–270.

Huang, M. Y. (2024). Sensitivity analysis for the generalization of experimental results. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187(4):900–918.

Huitfeldt, A., Stensrud, M. J., and Suzuki, E. (2019). On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging Themes in Epidemiology*, 16(1).

Kallus, N., Mao, X., and Uehara, M. (2022). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv*: 2103.14029v4.

Kennedy, E. H. (2024). Semiparametric doubly robust targeted double machine learning: A review. In *Handbook of Statistical Methods for Precision Medicine*. Chapman and Hall/CRC.

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

Kompa, B., Bellamy, D. R., Kolokotrones, T., Robins, J. M., and Beam, A. L. (2022). Deep learning methods for proximal inference via maximum moment restriction. In *NIPS '22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 11189–11201. Curran Associates.

Kress, R. (2014). *Linear integral equations*, volume 82 of *Applied Mathematical Sciences*. Springer.

Lee, D., Gao, C., Ghosh, S., and Yang, S. (2024). Transporting survival of an HIV clinical trial to the external target populations. *Journal of Biopharmaceutical Statistics*, 34(6):922–943.

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D., and Cai, J. (2023). Improving trial generalizability using observational studies. *Biometrics*, 79(2):1213–1225.

Lee, D., Yang, S., and Wang, X. (2022). Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *Journal of Causal Inference*, 10(1):415–440.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., and Cole, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, 28(4):553–561.

Li, X., Miao, W., Lu, F., and Zhou, X.-H. (2023). Improving efficiency of inference in clinical

trials with external control data. *Biometrics*, 79(1):394–403.

Lipsitch, M., Tchetgen Tchetgen, E., and Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388.

Liu, J., Park, C., Li, K., and Tchetgen Tchetgen, E. J. (2024). Regression-based proximal causal inference. *American Journal of Epidemiology*. In press.

Marso, S. P., Bain, S. C., Consoli, A., Eliaschewitz, F. G., Jódar, E., Leiter, L. A., Lingvay, I., Rosenstock, J., Seufert, J., Warren, M. L., Woo, V., Hansen, O., Holst, A. G., Pettersson, J., and Vilsbøll, T. (2016a). Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine*, 375(19):1834–1844.

Marso, S. P., Daniels, G. H., Brown-Frandsen, K., Kristensen, P., Mann, J. F., Nauck, M. A., Nissen, S. E., Pocock, S., Poulter, N. R., Ravn, L. S., Steinberg, W. M., Stockner, M., Zinman, B., Bergenstal, R. M., and Buse, J. B. (2016b). Liraglutide and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, 375(4):311–322.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7512–7523. PMLR.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Nguyen, T. Q., Ackerman, B., Schmid, I., Cole, S. R., and Stuart, E. A. (2018). Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PLoS ONE*, 13(12):e0208795.

Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., and Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, 11(1):225–247.

Nie, X., Imbens, G., and Wager, S. (2021). Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv*:2112.04723v1.

Oberst, M., D'Amour, A., Chen, M., Wang, Y., Sontag, D., and Yadlowsky, S. (2023). Understanding the risks and rewards of combining unbiased and possibly biased estimators, with applications to causal inference. *arXiv*: 2205.10467v2.

O'Muircheartaigh, C. and Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(2):195–210.

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

Pfanzagl, J. (1982). *Contributions to a general asymptotic statistical theory*. Number 13 in Lecture Notes in Statistics. Springer. With the assistance of W. Wefeleyer.

Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R., and Welton, N. J. (2018). Methods for population-adjusted indirect comparisons in health technology appraisal. *Medical Decision Making*, 38(2):200–211.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. Working Paper 128, Center for Statistics and the Social Sciences, University of Washington.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rotnitzky, A. and Smucler, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86.

Røysland, K., C. Ryalen, P., Nygård, M., and Didelez, V. (2025). Graphical criteria for the

identification of marginal causal effects in continuous-time survival and event-history analyses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(1):74–97.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188.

Rubino, D. M., Greenway, F. L., Khalid, U., O'Neil, P. M., Rosenstock, J., Sørrig, R., Wadden, T. A., Wizert, A., Garvey, W. T., and STEP 8 Investigators (2022). Effect of weekly subcutaneous semaglutide vs daily liraglutide on body weight in adults with overweight or obesity without diabetes: The STEP 8 randomized clinical trial. *JAMA*, 327(2):138–150.

Rudolph, K. E. and van der Laan, M. J. (2017). Robust estimation of encouragement design intervention effects transported across sites. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(5):1509–1525.

Rytgaard, H. C. W., Eriksson, F., and van der Laan, M. J. (2023). Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*, 79(4):3038–3049.

Severini, T. A. and Tripathi, G. (2012). Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, 170(2):491–498.

Shi, X., Miao, W., Nelson, J. C., and Tchetgen Tchetgen, E. J. (2020a). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):521–540.

Shi, X., Miao, W., and Tchetgen, E. (2020b). A selective review of negative control methods in epidemiology. *Current Epidemiology Reports*, 7(4):190–202.

Singh, R. (2023). Kernel methods for unobserved confounding: Negative controls, proxies, and instruments.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen Tchetgen, E. J. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science*, 31(3):348–361.

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

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. (2024). An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3):239–266.

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., and Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, 7(1):114–135.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer Series in Statistics. Springer.

van der Laan, M., Qiu, S., Tarp, J. M., and van der Laan, L. (2025). Adaptive-TMLE for the average treatment effect based on randomized controlled trial augmented with real-world data. *arXiv*:2405.07186v2.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (2023). *Weak convergence and empirical processes: With applications to statistics*. Springer Series in Statistics. Springer.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., and Thompson, L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54.

Vo, T.-T., Porcher, R., Chaimani, A., and Vansteelandt, S. (2019). A novel approach for identifying

and addressing case-mix heterogeneity in individual participant data meta-analysis. *Research Synthesis Methods*, 10(4):582–596.

Wang, G., Levis, A., Steingrimsson, J., and Dahabreh, I. (2024). Efficient estimation of subgroup treatment effects using multi-source data. *arXiv:* 2402.02684v1.

Westreich, D., Edwards, J. K., Lesko, C. R., Stuart, E., and Cole, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, 186(8):1010–1014.

Wilding, J. P., Batterham, R. L., Calanna, S., Davies, M., Van Gaal, L. F., Lingvay, I., McGowan, B. M., Rosenstock, J., Tran, M. T., Wadden, T. A., Wharton, S., Yokote, K., Zeuthen, N., and Kushner, R. F. (2021). Once-weekly semaglutide in adults with overweight or obesity. *New England Journal of Medicine*, 384(11):989–1002.

Yang, L. and Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician*, 55(4):314–321.

Yang, Q., Yang, Z., Cai, X., Zhao, H., Jia, J., and Sun, F. (2024). Advances in methodologies of negative controls: A scoping review. *Journal of Clinical Epidemiology*, 166:111228.

Yang, S., Gao, C., Zeng, D., and Wang, X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596.

Zhang, J., Li, W., Miao, W., and Tchetgen Tchetgen, E. (2023). Proximal causal inference without uniqueness assumptions. *Statistics & Probability Letters*, 198:109836.

Zhang, Z. (2009). Covariate-adjusted putative placebo analysis in active-controlled clinical trials. *Statistics in Biopharmaceutical Research*, 1(3):279–290.

Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted learning: Causal inference for observational and experimental data*, Springer Series in Statistics, pages 459–474. Springer.

Manuscripts

Manuscript I

Proximal indirect comparison

Zehao Su
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
zehao.su@sund.ku.dk

Helene C. Rytgaard
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
hely@sund.ku.dk

Henrik Ravn
Novo Nordisk A/S,
Vandtårnsvej 108, DK-2860 Søborg, Denmark
hnrv@novonordisk.com

Frank Eriksson
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
eriksson@sund.ku.dk

Abstract

We consider the problem of indirect comparison, where a treatment arm of interest is absent by design in one randomized controlled trial (RCT) but available in the other. The former is the target RCT, and the latter is the source RCT. The identifiability of the target population average treatment effect often relies on conditional transportability assumptions. However, it is a common concern whether all relevant effect modifiers are measured and controlled for. We give a new proximal identification result in the presence of shifted, unobserved effect modifiers based on proxies: an adjustment proxy in both RCTs and an additional reweighting proxy in the source RCT. We propose an estimator which is doubly-robust against misspecifications of the so-called bridge functions and asymptotically normal under mild consistency of estimators for the bridge functions. We use two weight management trials as a context to illustrate selection of proxies and apply our method to compare the weight loss effect of active treatments from these trials.

Keywords: Indirect comparison; Meta-analysis; Transportability; Proximal causal inference.

1. Introduction

Indirect comparison is the contrast of treatments that are not compared in head-to-head randomized controlled trials (RCTs). An important application of indirect comparison in health technology assessment is the comparison of a new treatment and an existing treatment in the health system, when both treatments are only studied in placebo-controlled RCTs. Indirect comparison can be viewed as an instance of transportability in causal inference. The transportability of the effect of the existing treatment versus placebo (or the lack of such transportability) from the source to the target RCT determines whether the effect between the new and the existing treatment can be established. It is therefore hardly surprising that current methods for indirect comparison require effect-measure transportability, adjusting for shifted effect modifiers (Colnet et al., 2024), that is, the effect modifiers that do not follow the same distribution across the RCTs. However, when there are unobserved shifted effect modifiers, transportability cannot be established by controlling for the observed baseline variables. The lack of transportability jeopardizes the external validity of treatment effects in an indirect comparison. If the treatments of interest come from RCTs which are conducted with a considerable time gap apart, there may be changes in the standard of care that could affect the treatment effects. Social determinants of health, which are often unmeasured in RCTs, can also change the magnitude of treatment effects.

In this paper, we use negative controls, or proxies, to minimize external validity bias. In observational studies, negative controls are known to help detect unmeasured confounding (Lipsitch et al., 2010). Recently, a family of methods called proximal causal inference has shown how appropriately selected proxies may rectify confounding bias. Miao et al. (2018) demonstrated a nonparametric identification formula for the counterfactual distribution of outcomes with a pair of complementary proxies in the presence of unmeasured confounders. Under a nearly identical design, Cui et al. (2024) proposed a proximal doubly robust estimator of the average treatment effect based on two identification strategies.

In indirect comparisons, randomization eliminates the threat of confounding, and we use proxies to tackle bias arising from shifted, unobserved effect modifiers. We propose a novel method that extends proximal causal inference to indirect comparison when individual patient data (IPD) is available in all RCTs. Our proposed estimator also relies on a pair of proxies, namely a reweighting proxy and an adjustment proxy. The proxies correspond to identification strategies that mirror participation odds weighting and the g-formula for transporting causal effects. We remark that while both proxies are required in the source RCT, only the adjustment proxy needs to be collected in the target RCT. Our proposed estimator handles both continuous and binary outcomes. We show that the estimator is robust against misspecifications of the required nuisance functions of the proxies.

2. Indirect comparison with unmeasured shifted effect modifiers

Consider two treatment pairs, $A \in \{0, 1\}$ and $A \in \{0, -1\}$, where the former is the treatments investigated in the source RCT $S = 1$ and the latter is those in the target RCT $S = 0$. This corresponds to the situation where there is a treatment shared by the two RCTs, in this case $A = 0$, so that the comparison of the treatments $A = 1$

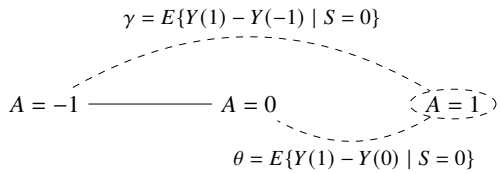

Figure 1. Indirect comparison parameters defined on the target RCT population $S = 0$. Treatment $A = 1$ is unavailable in the target RCT.

and $A = -1$ may be made with the help of the common treatment arm. Indirect comparison of this kind is called anchored comparison (Phillippo et al., 2018). The setup described here is a subset of network meta-analysis, which usually places interest on the causal effects comparing at least three different pairs of interventions. The treatment $A = 1$ that we wish to emulate in $S = 1$, also referred to as the missing treatment, can be a placebo or an active treatment. Additionally in each RCT a set of baseline covariates X is measured. Let $Y(a)$ denote the potential outcome under the intervention $A = a \in \{-1, 0, 1\}$. In this paper we consider as the target parameter θ the average treatment effect (ATE) in the target population comparing treatments $A = 1$ and $A = 0$; that is, $\theta = E\{Y(1) - Y(0) \mid S = 0\}$. The indirect comparison parameter $\gamma = E\{Y(1) - Y(-1) \mid S = 0\}$ is the difference between θ and the ATE in the target RCT, $E\{Y(-1) - Y(0) \mid S = 0\}$. Figure 1 contains a diagram describing the parameters γ and θ. For the identifiability of θ, the natural effect measure for transportability is the conditional average treatment effect (CATE). The mean scale is a common choice in various problems where the outcome is continuous (e.g., body weight) or binary (e.g., occurrence of a cardiac arrest).

The fundamental problem of indirect comparison is that the treatment $A = 1$ is never observed in the target RCT $S = 0$. In order to establish identifiability for the target parameter in the observed data, we need a set of plausible assumptions which justify the transportability of the treatment-specific mean from the source population to the target population. The existing indirect comparison and network meta-analysis literature tend to assume effect-measure transportability (Phillippo et al., 2018, 2020), which does not hold when conditioning on the observed baseline covariates cannot completely account for differences in the effect between RCTs, because there exist unmeasured shifted effect modifiers. In hope of capturing all effect modifiers, it may be tempting to adjust for as many baseline or pre-treatment variables as possible in order to achieve valid effect-measure transportability. However, this strategy is not foolproof, as it may instead result in additional M-bias (Cinelli et al., 2024). In Supplementary Material §S1, we give examples of data generating mechanisms where not only does conditional transportability not hold by means of adjustment, but the bias is further amplified by adjustment.

We propose a weaker version of CATE transportability conditioning on both the baseline covariates X and the unobserved effect modifiers U.

Assumption 1. $\text{pr}(X, U \mid S = 0)$-almost surely:

(i) (Consistency) $Y(a) = Y$ whenever $A = a \in \{0, 1\}$;

(ii) (Randomization) $\{Y(1), Y(0), U\} \perp\!\!\!\perp A \mid (X, S = 1)$;

(iii) (CATE transportability) $E\{Y(1) - Y(0) \mid X, U, S = 0\} = E\{Y(1) - Y(0) \mid X, U, S = 1\}$;

(iv) (Positivity) $\text{pr}(S = 1 \mid X, U)\text{pr}(A = a \mid X, S = 1) > 0$.

Assumption 1(iv) is crucial for the transportability of causal effects from the source population $S = 1$ to the target population $S = 0$. It requires that all values of $\{X, U\}$ observable in the target population must also be observable in the source population. The positivity assumption on trial participation essentially guarantees no extrapolation of information from the source trial. As we will see in §3, violation of positivity not only compromises the identifiability of the target parameter but also has great implications on its estimation with observed data.

The causal assumptions give rise to identifiability of the target parameter in the full data distribution. Let $\alpha = \text{pr}(S = 0)$ be the probability of an individual in the joined population belonging to the target RCT. We denote the propensity score in the source trial by $e(a \mid X) = \text{pr}(A = a \mid X, S = 1)$ for $a \in \{0, 1\}$. In the source RCT, we also let $\tilde{Y} = (2A - 1)Y/e(A \mid X)$ denote a transformed outcome, whose behavior is comparable to that of the contrast $Y(1) - Y(0)$, in the sense that their conditional expectations on $\{X, U\}$ are equal.

Proposition 1 (Latent identifiability). *Suppose Assumption 1 holds. The target parameter is identifiable as*

$$\theta = E\big\{E(\tilde{Y} \mid X, U, S = 1) \,\big|\, S = 0\big\}. \tag{1}$$

Equivalently,

$$\theta = \frac{1}{\alpha} E\left\{S \frac{\text{pr}(S = 0 \mid X, U)}{\text{pr}(S = 1 \mid X, U)} \tilde{Y}\right\}. \tag{2}$$

These identification formulae of θ are very similar to the transportability results in Theorem 1 of Dahabreh et al. (2023), where the outcome model and the participation odds depend on the baseline covariates X only. The distinction mainly lies in their assumptions for the exchangeability of trial participation and subsequently the exchangeability of treatment assignment, which hold without the additional unobserved covariates U.

Proposition 1 suggests that in order to identify the target parameter θ, we need at least the knowledge of either the mean outcome difference $E(\tilde{Y} \mid X = x, U = u, S = 1)$ or the trial participation odds $\text{pr}(S = 0 \mid X = x, U = u)/\text{pr}(S = 1 \mid X = x, U = u)$. However, both quantities depend on the unobserved effect modifiers U, and neither would be identifiable without further assumptions and/or extra information.

3. Reweighting and adjustment proxies

In this section, we extend the proximal causal inference framework to handle scenarios in indirect comparison where CATE transportability fails to hold conditionally on the baseline covariates X. We refer to this approach as proximal indirect comparison. The core idea of proximal indirect comparison is that the knowledge of a pair of proxies (Z, W) helps to learn the underlying dependence of $Y(1) - Y(0)$ on U, thereby restoring the identifiability of the target parameter using observed data. We call Z the reweighting proxy and W the adjustment proxy. The intuition is that Z will be used to reweight the samples from the source RCT to match the composition of the target RCT like in (2),

and W will appear in the adjustment formula for transportability in (1), as if we had adjusted for U.

3.1. *Proximal identifiability*

Before discussing identifiability of the parameter, we formally introduce the variables at our disposition. Specifically, we observe $O^1 = (A, X, Y, W, Z)$ in the source RCT and $O^0 = (X, W)$ in the target RCT, so the observed variables can be represented by $O = (S, SA, X, SY, W, SZ)$. We describe the following distinction between the observed data and the full data. Denote the full data distribution over (O, U) by P^{U}. We can marginalize it to a measure P over O. Further, P^1 stands for the conditional probability measure over O^1 such that $P^1(O^1 \in \cdot) = P(O^1 \in \cdot, S = 1)/P(S = 1)$.

We require that the proxies satisfy a set of conditional independences.

Assumption 2 (Adjustment and reweighting proxies). In P^{U}:
(i) $\{Z, W, U\} \perp\!\!\!\perp A \mid (X, S = 1)$;
(ii) $Z \perp\!\!\!\perp W \mid (X, U, S = 1)$;
(iii) $Z \perp\!\!\!\perp Y \mid (A, X, U, S = 1)$;
(iv) $W \perp\!\!\!\perp S \mid (X, U)$.

This assumption describes the relations between valid proxies and other variables. In the source RCT, Assumptions 2(i)–2(iii) essentially require that randomization does not depend on the proxies, that the proxies are not associated through other unmeasured variables besides U, and that Z has no causal effect on Y, unless it is mediated through $\{X, U\}$. Also, if Z and W occur after randomization, they must be negative control outcomes in the sense that they can share many causes with Y but may not be caused by the treatment A. Across the two RCTs, Assumption 2(iv) requires that the difference in the distribution of W is totally accounted for by $\{X, U\}$. A directed acyclic graph (DAG) encoding conditional independences which are compatible with Assumption 2 is displayed in Fig. 2. Since the full data distribution can satisfy this assumption in various ways, we list additional examples of compatible graphs in Fig. S2 in the Supplementary Material. Some differences to Fig. 2 include Z being a cause of U ($Z \to U$) and W sharing unmeasured causes with Y ($W \leftrightarrow Y$). However, no direct edge may exist between Z and $\{A, Y, W\}$ or between W and $\{A, S, Z\}$.

Problem-specific effect measures allow for the relaxation of the distribution-level Assumption 2(iii). In this work, we work with the CATE transportability Assumption 1(iii), and the target parameter is an ATE. It suffices to assume that

$$E(Y \mid A = 1, Z, X, U, S = 1) - E(Y \mid A = 0, Z, X, U, S = 1) \\ = E(Y \mid A = 1, X, U, S = 1) - E(Y \mid A = 0, X, U, S = 1).$$

That is, we allow $Z \not\perp\!\!\!\perp Y \mid (A, X, U, S = 1)$, as long as Z is not an effect modifier of A on Y on the CATE scale after adjusting for $\{X, U\}$.

Heuristically, we want the adjustment proxy W to emulate the effect of U in the mean outcome difference model. Similarly, the reweighting proxy Z should take the place of U in the participation odds model. To this end, we introduce two sets of functions involving the proxies, which are defined on the full data distribution:

$$\mathcal{H}^{\mathrm{U}} = \left\{h^{\mathrm{U}}(w, x) \in L_2(W, X; P^1) : E\{\tilde{Y} - h^{\mathrm{U}}(W, X) \mid X, U, S = 1\} = 0\right\},$$

$$\mathcal{Q}^{\mathrm{U}} = \left\{ q^{\mathrm{U}}(z,x) \in L_2(Z,X;P^1) : E\{q^{\mathrm{U}}(Z,X) \mid X,U,S=1\} = \frac{P^{\mathrm{U}}(S=0 \mid X,U)}{P^{\mathrm{U}}(S=1 \mid X,U)} \right\},$$

where the conditions inside the sets hold $P^{\mathrm{U}}(X,U \mid S=0)$-almost surely, and $L_2(V;P^1)$ is the space of square integrable functions of V with respect to the probability measure P^1.

We refer to the elements of $\mathcal{H}^{\mathrm{U}}$ and $\mathcal{Q}^{\mathrm{U}}$, if they exist, as outcome bridge functions and participation bridge functions, respectively. These bridge functions recover the unobserved nuisance functions in Proposition 1 after being projected onto a subspace of the full data distribution.

Assumption 3. Either $\mathcal{H}^{\mathrm{U}}$ or $\mathcal{Q}^{\mathrm{U}}$ is nonempty.

A sufficient condition for $\mathcal{H}^{\mathrm{U}}$ or $\mathcal{Q}^{\mathrm{U}}$ to be nonempty is a relevance assumption for the proxy W or Z, stating that the proxy should at least be correlated with U after controlling for the baseline covariates. If the bridge functions do not exist, measurements of the proxies will not grant more information on U than what can be inferred from adjusting for X, making them ineffective in mimicking the dependence of U on S and of Y on U. See the discussions in Kallus et al. (2022), Examples 3 and 4. Some examples of potential unobserved effect modifiers are the standard of care and the social determinants of health among the RCT participants. These are often high-dimensional covariates that are not available from the experimental setting. Nonetheless, it may be reasonable to posit that only a low-dimensional subset of these covariates strongly contributes to effect modification, such as concomitant medication and employment stability. If this is true, then the number of proxies does not have to be large for bridge functions to exist.

On the observed data distribution, we define two sets of observed data bridge functions:

$$\mathcal{H} = \left\{ h(w,x) \in L_2(W,X;P^1) : E\{\tilde{Y} - h(W,X) \mid Z,X,S=1\} = 0 \right\},$$
$$\mathcal{Q} = \left\{ q(z,x) \in L_2(Z,X;P^1) : E\{q(Z,X) \mid W,X,S=1\} = \frac{P(S=0 \mid W,X)}{P(S=1 \mid W,X)} \right\},$$

where the condition inside $\mathcal{H}$ holds on the set $\{(z,x) : P(S=1 \mid Z=z, X=x)P(S=0 \mid X=x) > 0\}$, and the condition inside $\mathcal{Q}$ holds $P(W,X \mid S=0)$-almost surely.

Under the conditional independences in Assumption 2, we can relate the outcome bridge functions $h^{\mathrm{U}}(w,x)$ to the reweighting proxy Z and the participation bridge functions $q^{\mathrm{U}}(z,x)$ to the adjustment proxy W. The existence of the bridge functions on the full data distribution implies their existence on the observed data distribution under proper proxy assumptions.

Lemma 1. *If Assumptions 2(i)–2(iii) hold, then $\mathcal{H}^{\mathrm{U}} \subset \mathcal{H}$. If Assumptions 2(ii) and 2(iv) hold, then $\mathcal{Q}^{\mathrm{U}} \subset \mathcal{Q}$.*

The sets of observed bridge functions give rise to the following identifiability result.

Proposition 2 (Identifiability). *Suppose Assumptions 1–3 hold. If $\mathcal{H} \neq \emptyset$ and $\mathcal{Q} \neq \emptyset$, then the target parameter θ is identifiable in the observed data distribution P. For any $h \in \mathcal{H}$,*

$$\theta = E\{h(W,X) \mid S=0\},$$

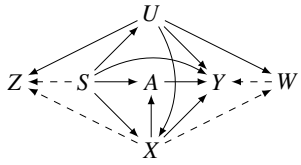

Figure 2. A DAG compatible with Assumption 2. Dashed edges can be removed.

and for any $q \in \mathcal{Q}$,

$$\theta = \frac{1}{\alpha} E\{Sq(Z, X)\tilde{Y}\}.$$

If there exist solutions to both of the equations in $\mathcal{H}$ and $\mathcal{Q}$, then they can essentially be regarded as the unobservable bridge functions in $\mathcal{H}^{\mathrm{U}}$ and $\mathcal{Q}^{\mathrm{U}}$ in the identification of the target parameter.

Rather than CATE transportability, an indirect comparison can also assume mean transportability $E\{Y(1) \mid X, U, S = 1\} = E\{Y(1) \mid X, U, S = 0\}$. An analogous assumption to the existence of an outcome difference bridge function $h^{\mathrm{U}} \in \mathcal{H}^{\mathrm{U}}$ is the existence of functions $\tilde{h}^{\mathrm{U}}(w, x)$ such that $E\{Y - \tilde{h}^{\mathrm{U}}(W, X) \mid X, A = 1, U, S = 1\} = 0$. This is referred to as unanchored indirect comparison (Phillippo et al., 2018). Unanchored indirect comparison is valid only when all shifted prognostic variables, observed and unobserved, are taken into account. From a practical point of view, the existence of bridge functions is related to the explanatory power of the proxies relative to the unobserved variables. The set of proxies required for identifiability of the causal parameter $E\{Y(1) \mid S = 0\}$ via the outcome bridge $\tilde{h}^{\mathrm{U}}$ is potentially much larger than what is needed for the identifiability of θ via the outcome difference bridge h^{U}.

3.2. *Connections to other proximal causal inference approaches*

Ghassami et al. (2022) described a method for estimating long-term treatment effects exploiting the internal validity of experimental data while using proxies to account for confoundedness in observational data. The target population is the one defined by the observational data, and the short-term outcome plays the role of outcome-inducing proxy W. The treatment bridge function q in their work is intended to emulate the inverse of the propensity score in the confounded data, which depends on unobserved confounders U. Imbens et al. (2024) also employed proximal causal inference in their attempt to estimate long-term treatment effect via data fusion. The selection bridge function assumed in their work has a similar form as the participation bridge in $\mathcal{Q}$, in that they both capture the variability of the unobserved variables U between two data sources.

However, the problem studied in these two articles is fundamentally different from ours, since the authors make a "data combination" assumption, which is Assumption 3 in Ghassami et al. (2022) and Assumptions 3 and 10 in Imbens et al. (2024). The assumption is that the unobserved variables are independent of the RCT indicator S, possibly conditioning on some baseline covariates X, but we do not assume this. In proximal indirect comparison, we wish to capture the difference in distributions of unobserved variables between the study populations through the use of proxies. Additionally, both bridge functions in these works are assumed on the target population, whereas

the bridge functions $\mathcal{H}$ and $\mathcal{Q}$ in proximal indirect comparison are posited on the source population. In particular, the outcome bridge cannot be defined nor learned on the target population due to the impossibility of observing certain treatments of interest.

In the appendix of their work, Imbens et al. (2024) provide an additional identification result, allowing unobserved variables to vary between experimental population and observational population. Instead, they require that the treatment is completely randomized and that latent transportability holds on treatment-specific means. Unlike their setup, we avoid making these assumptions, since our bridge functions do not involve specific treatments. Our Proposition 2 remains valid under stratified randomization, which is common in RCTs.

Ghassami et al. (2024) proposed a closely related approach for causal mediation analysis and front-door adjustment in the presence of hidden mediators. For the identifiability of controlled and uncontrolled treatment-specific means, they assumed a treatment bridge function describing the difference between the unobserved variables under different interventions (their Assumption 9). Despite the resemblance of the bridge functions in our works, we are interested in relaxing the CATE transportability for the estimation of the treatment effect with at least one treatment arm that is not present in the target population. The unmeasured variables in indirect comparison cannot be a mediator in that the CATE transportability in Assumption 1(iii) is conditional on U. In addition, the RCT indicator S is not an intervention nor exposure, so conceptually U cannot be considered as a mediator between S and Y.

We make an important remark that for data fusion, it is often not necessary to observe one group of the proxies in both populations. For example, in Ghassami et al. (2022) the treatment-inducing proxy is only required in the observational regime. The short-term outcome S_1 in Imbens et al. (2024), which enjoys similar properties as a treatment-inducing proxy, appears also in the observational data only. In proximal indirect comparison, no reweighting proxy Z needs to be observed in the target population.

4. Asymptotic theory for target parameter estimation

Before presenting an estimator for the target parameter, we make a connection between the sampling scheme and the probability model. In the source trial $S_i = 1$, the observed data is an independent and identically distributed (i.i.d.) sample $(A_i, X_i, W_i, Z_i, Y_i)$, $i = 1, 2, \ldots, n_1$. In the target trial $S_i = 0$, the observed data is an i.i.d. sample of (X_i, W_i), $i = n_1 + 1, n_1 + 2, \ldots, n_1 + n_0$. For the asymptotic arguments, we require that the ratio n_0/n approach the fixed number α between zero and one when n goes to infinity, where $n = n_0 + n_1$ is the total number of observations. Then we can consider the observed data as an i.i.d. sample from P_0, the true data generating mechanism. In this section, we use subscript 0 to indicate dependence on P_0. We write $\tilde{Y}_0 = (2A-1)Y/e_0(A \mid X)$. We deal with the true parameter θ_0, which has a causal interpretation under the assumptions in Proposition 2.

4.1. *A nonparametric influence function*

We study the target parameter with tools from semiparametric estimation theory under the following regularity conditions on the true data distribution P_0. Let the linear transformation $T_0 : L_2(W, X; P_0^1) \to L_2(Z, X; P_0^1)$ be such that $(T_0 h)(z, x) =$

$E_{P_0}\{h(W,X) \mid Z = z, X = x, S = 1\}$. Moreover, let F_0 be the linear transformation $(F_0 g)(z,x) = E_{P_0}[\{\tilde{Y}_0 - h_0(W,X)\}g(Y,Z,W,A,X) \mid Z = z, X = x, S = 1]$ where $g \in L_2(Y,Z,W,A,X;P_0^1)$.

Assumption 4 (Regularity conditions).

(i) $E_{P_0}(\tilde{Y}_0 \mid Z = z, X = x, S = 1) \in L_2(Z,X;P_0^1)$, $P_0(S = 1 \mid W = w, X = x)/P_0(S = 0 \mid W = w, X = x) \in L_2(W,X;P_0^1)$;

(ii) T_0 is bijective;

(iii) $\text{range}(F_0) \subset L_2(Z,X;P_0^1)$.

(iv) $(Z,W) \perp\!\!\!\perp_{P_0} A \mid (X, S = 1)$ and $e_0(a \mid x)$ is known.

Define the model

$$\mathcal{P} = \{P \in \mathcal{M} : (Z,W) \perp\!\!\!\perp A \mid (X, S = 1), e(a \mid x) \text{ known}, \mathcal{H} \neq \emptyset\},$$

where $\mathcal{M}$ is the set of all probability measures over O. Assumptions 4(i)–4(iii) are not necessary for proposing regular and asymptotically linear estimators for θ_0, but they enable the characterization of all such estimators of θ_0 under $\mathcal{P}$. We do not consider the propensity score $e_0(a \mid x)$ as a nuisance function, since the treatments in RCTs are usually administered according to a predetermined protocol.

It is clear that under Assumptions 4(i)–(ii), we have $\mathcal{H}_0 = \{h_0\}$ and $\mathcal{Q}_0 = \{q_0\}$. This observation has two implications. First, $\mathcal{H}_0$ and $\mathcal{Q}_0$ are simultaneously nonempty. Working under boundedness conditions related to the projected variance of the bridge functions, Zhang et al. (2023) show that this condition is necessary for the observed data functional θ_0 to be $n^{1/2}$-estimable. Second, the bridge functions h_0 and q_0 are unique. If other bridge functions exist, the target parameter will be a uniquely identified functional on possibly nonunique nuisance parameters (Zhang et al., 2023; Bennett et al., 2023). Although this poses no difficulty to identification of the target parameter, it largely hinders the study of estimators constructed using estimates of the bridge functions. In Supplementary Material §S3, we give sufficient conditions for establishing the existence and uniqueness of the observed data bridge functions, namely completeness assumptions and more regularity conditions on T_0 and its adjoint.

We are now in a position to present a useful characterization of the target parameter by an influence function.

Proposition 3. *Suppose Assumption 4 holds. An influence function of the observed data target parameter θ_0 under the model $\mathcal{P}$ is*

$$\phi_0(o) = \frac{s}{\alpha_0} q_0(z,x)\{\tilde{y}_0 - h_0(w,x)\} + \frac{1-s}{\alpha_0}\{h_0(w,x) - \theta_0\}.$$

In fact, we can find many influence functions of θ_0 under the model restrictions specified in Assumption 4. We purposefully choose to present ϕ_0 over the efficient one, which involves an additional nuisance parameter $E_{P_0}(Y \mid Z = z, W = w, A = a, X = x, S = 1)$. The efficient influence function can be found in the proof of the proposition in Supplementary Material §S6. In the sequel, we construct an estimator for θ_0 from ϕ_0, which is preferred because we only need to estimate the bridge functions h_0 and q_0. While we do not pursue the possibility here, we remark that asymptotically efficient estimators, which attain the semiparametric efficiency bound, may be constructed from the efficient influence function.

4.2. *Doubly robust estimation*

In general, we do not have the knowledge of the bridge functions. Therefore, we resort to two-stage data-adaptive estimation of the target parameter. In the first stage, we use the observed data to obtain estimates of the bridge functions, which are nuisance functions to the estimation problem. In the second stage, we plug in the estimated bridge functions to some valid estimating equation for the target parameter, and an estimate of the target parameter is obtained as the solution to the estimating equation.

Suppose $\hat{h}$ and $\hat{q}$ are nonparametric or semiparametric estimators intended for the true, unique bridge functions h_0 and q_0. Let $\hat{\alpha} = n_0/n$ be the proportion of samples from the target RCT. Propositions 2 and 3, together with the form of the influence function ϕ_0, suggest a natural estimator of the target parameter θ_0 as

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{S_i}{\hat{\alpha}}\hat{q}(Z_i, X_i)\{\tilde{Y}_i - \hat{h}(W_i, X_i)\} + \frac{1-S_i}{\hat{\alpha}}\hat{h}(W_i, X_i)\right].$$

The estimator $\hat{\theta}$ solves the estimating equation based on ϕ_0. We now characterize its asymptotic behavior.

Assumption 5 (Regularity conditions).

(i) The function class

$$\mathcal{G}_0 = \left\{ g(o) = \frac{s}{\alpha'}q'\{\tilde{y} - h'\} + \frac{1-s}{\alpha'}h' : \right.$$
$$\left. \alpha' \in [0,1], h' \in L_2(W, X; P_0^1), q' \in L_2(Z, X; P_0^1) \right\}$$

is P_0-Donsker;

(ii) There exists a universal constant $M > 1$ such that $\alpha_0 \geq M^{-1}$, $\hat{\alpha} \geq M^{-1}$, $e_0 \geq M^{-1}$, $|h_0| \leq M$, $|\hat{q}| \leq M$, $P_0(S = 1 \mid W, X) \geq M^{-1}$, and $E_{P_0}(Y^2 \mid Z, X, S = 1) \leq M$.

Theorem 1. *Suppose Assumptions 4–5 hold and that* $\|\hat{h} - \bar{h}\|_{P_0^1} = o_{P_0}(1)$, $\|\hat{q} - \bar{q}\|_{P_0^1} = o_{P_0}(1)$ *for some nonrandom functions* $\bar{h}(w, x)$, $\bar{q}(z, x)$ *in* $L_2(P_0^1)$. *Then:*

1. *The estimator* $\hat{\theta}$ *is consistent for* θ_0, *if either* $\bar{h} = h_0$ *or* $\bar{q} = q_0$.
2. *The estimator* $\hat{\theta}$ *is asymptotically linear with influence function* ϕ_0, *if* $\bar{h} = h_0$, $\bar{q} = q_0$, *and* $\|\hat{q} - q_0\|_{P_0^1}\|\hat{h} - h_0\|_{P_0^1} = o_{P_0}(n^{-1/2})$.

The Donsker class condition on $\mathcal{G}_0$ can be relaxed by applying cross-fitting to the estimation of the bridge functions. When the bridge functions are estimated using minimax criteria, their convergence in the $L_2(P_0^1)$-norm can be established following the arguments from Kallus et al. (2022). A key assumption is the equivalence of the $L_2(P_0^1)$-norm of the bridge functions and that of the projected bridge functions, the latter of which is easier to bound. For the outcome bridge function, it amounts to $\|\hat{h} - h_0\|_{P_0^1} = O_{P_0}\{\|T_0(\hat{h} - h_0)\|_{P_0^1}\}$. The bijectivity of T_0 from Assumption 4(ii) is sufficient for the norm equivalence, if T_0 is also a bounded operator.

The estimator $\hat{\theta}$ is doubly robust in the sense that it is consistent if either the outcome bridge function h_0 or the participation odds bridge function q_0 is correctly estimated.

Moreover, the estimator is asymptotically normal under the observed data model $\mathcal{P}$, when both bridge functions converge sufficiently fast to the ground truth, for example, both at the subparametric $o_{P_0}(n^{-1/4})$-rate. In this case, the squared empirical $L_2(P_0)$-norm $n^{-1}\sum_{i=1}^n \hat{\phi}_0^2(O_i)$ is a consistent estimator of the asymptotic variance of the estimator, where we obtain $\hat{\phi}_0$ by plugging the nuisance estimates into ϕ_0.

5. Numerical results

In this section, we posit parametric models for the bridge functions to illustrate the method of proximal indirect comparison in numerical studies. Note however, in general the parametric assumptions are not necessary, and the asymptotic properties of the estimator $\hat{\theta}$ in §4 hold for nonparametric estimators for the bridge functions. Following Cui et al. (2024), the finite-dimensional parameters in the bridge functions are estimated by the generalized method of moments motivated by their influence functions. The details are available in Supplemantary Material §S2.

5.1. *Simulated data example*

We generate the full data (X, U, S, SA, SZ, W, SY) sequentially from the distributions described below. The baseline covariates $X \sim \Phi[\text{Normal}\{(0,0,0)^{\mathrm{T}}, \Sigma\}]$, where $\Phi(\cdot)$ is the standard normal distribution function and the covariance matrix is

$$\Sigma = \begin{pmatrix} 1 & 0.25 & 0.25 \\ 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 1 \end{pmatrix},$$

and thus are bounded between 0 and 1. The rest of the variables are obtained in the following way:

$$\begin{aligned}
U &\sim \text{Uniform}([-1,0]\times[-1,0]\times[-1,0]),\\
S \mid (X,U) &\sim \text{Bernoulli}\{\text{expit}(-0.625 + 0.5X^{\mathrm{T}}1 + 0.5U^{\mathrm{T}}1)\},\\
A \mid (X, S=1) &\sim \text{Bernoulli}(0.5),\\
Z \mid (U, X, S=1) &\sim \text{Normal}(U+X, 0.25\text{Id}),\\
W \mid (U, X) &\sim \text{Normal}(U+X, 0.25\text{Id}),\\
Y \mid (W, A, X, U, S=1) &\sim \text{Normal}(0.5 - A + U^{\mathrm{T}}1 + AU^{\mathrm{T}}1 + X^{\mathrm{T}}1 + W^{\mathrm{T}}1 + AW^{\mathrm{T}}1, 0.5^2).
\end{aligned}$$

The parameters are selected so that the probability α is close to 0.65. In this data generating mechanism, U is an effect modifier and the target parameter is $\theta = E\{E(Y \mid A=1, X, U, S=1) - E(Y \mid A=0, X, U, S=1) \mid S=0\}$.

Let $b(z,x) = (1, z^{\mathrm{T}}, x^{\mathrm{T}})^{\mathrm{T}}$, $c(w,x) = (1, w^{\mathrm{T}}, x^{\mathrm{T}})^{\mathrm{T}}$. As we show in Supplementary Material §S4, the underlying bridge functions are unique and have the closed forms $h_{\eta_0}(w,x) = \eta_0^{\mathrm{T}} c(w,x)$ and $q_{\xi_0}(z,x) = \exp\{\xi_0^{\mathrm{T}} b(z,x)\}$, where η_0 and ξ_0 are nuisance parameter vectors of appropriate dimensions. We compare three estimators, namely $\hat{\theta}$ proposed in §4, as well as

$$\hat{\theta}_h = \frac{1}{n_0}\sum_{i:S_i=0} h_{\hat{\eta}}(W_i, X_i),$$

$$\hat{\theta}_q = \frac{1}{n_0} \sum_{i:S_i=1} q_{\hat{\xi}}(Z_i, X_i)\tilde{Y}_i.$$

The nuisance parameter estimators with correctly specified h_η and q_ξ were obtained on the full sample via the generalized method of moments:

$$\hat{\eta} = \arg\min_{\eta} \left\| \frac{1}{n_1} \sum_{i:S_i=1} b(Z_i, X_i)\{\tilde{Y}_i - h_\eta(W_i, X_i)\} \right\|^2,$$

$$\hat{\xi} = \arg\min_{\xi} \left\| \frac{1}{n} \sum_{i=1}^{n} \{c(W_i, X_i)\}^3 \{S_i q_\xi(Z_i, X_i) - (1 - S_i)\} \right\|^2.$$

The cubic of the function $c(w, x)$ makes sure that the estimators $\hat{\theta}$ and $\hat{\theta}_q$ are numerically distinguishable, as the true bridge function $h_\eta(w, x)$ is linear. We present details of the claim in Supplementary Material §S4.

To contrast the behavior of the estimators under model misspecifications, we considered configurations where neither h nor q was misspecified (experiment 1), where q was misspecified (experiment 2), where h was misspecified (experiment 3), and where both h and q were misspecified (experiment 4). The misspecified models were fitted by replacing W and Z with $|W|^{1/2}$ and $|Z|^{1/2}$ wherever appropriate. Summary statistics of the estimators from 1000 repeated samples of size $n \in \{1000, 2000\}$ are displayed in Table 1, where the reference Monte-Carlo target parameter was calculated by static interventions of A when $S = 0$. The standard errors of the estimators for $\hat{\theta}_h$ and $\hat{\theta}_q$ were obtained by plugging in the nuisance parameter estimates in the theoretical asymptotic variances shown in Supplementary Material §S2. The estimators $\hat{\theta}_h$ and $\hat{\theta}_q$ showed little bias only when h and q were correctly estimated, respectively. This was constrasted by $\hat{\theta}$, which exhibited the double robustness property as expected. The influence-function-based standard error for $\hat{\theta}$ also showed robustness against model misspecification. In Supplementary Material §S4, we present additional simulations under alternative data generating mechanisms to investigate the behavior of the proximal estimators, including invalid proxies, nonunique bridge functions, weak proxies and near violation of positivity. In particular, we found that Tikhonov regularization on the parameters η and ξ recovered valid inference for the doubly robust estimator when the bridge functions were nonuniquely defined.

5.2. *Real data example*

Proximal indirect comparison allows for treatment effect estimation via transportability in the presence of unobserved effect modifiers. For a real data application of our method, we make use of the individual-level patient data from two global weight management RCTs, namely SCALE [clinicaltrials.gov ID NCT03552757, Davies et al. (2015)] and STEP-2 [clinicaltrials.gov ID NCT01272232, Davies et al. (2021)]. While these trials are inherently longitudinal, we ignore this structure for the sole purpose of illustrating our method. Whenever a subject deviates from the predetermined protocol at randomization, we treat the subsequent weight measurements as missing.

The active treatments are once-daily liraglutide, 3.0 mg or 1.8 mg in SCALE and once-weekly semaglutide, 2.4 mg or 1.0 mg in STEP-2, injected subcutaneously, both of which are glucagon-like pepetide-1 (GLP-1) agonists. Both RCTs are placebo-controlled with placebo administration matched to their respective active treatments.

Table 1. *Simulation results of experiments 1–4.*

n	Experiment	Estimator	Mean	Bias	RMSE	SE	Coverage
1000	1	$\hat{\theta}_h$	−2.64	1.03	3.17	3.21	95.7
		$\hat{\theta}_q$	−2.65	−9.37	3.96	3.99	95.4
		$\hat{\theta}$	−2.64	3.28	3.70	3.75	95.4
	2	$\hat{\theta}_h$	−2.64	1.03	3.17	3.21	95.7
		$\hat{\theta}_q$	−2.39	251.28	3.95	3.16	86.2
		$\hat{\theta}$	−2.64	5.69	3.42	3.08	92.7
	3	$\hat{\theta}_h$	−2.39	257.16	3.91	3.05	86.6
		$\hat{\theta}_q$	−2.65	−9.37	3.96	3.99	95.4
		$\hat{\theta}$	−2.65	−3.10	4.07	3.84	93.8
	4	$\hat{\theta}_h$	−2.39	257.16	3.91	3.05	86.6
		$\hat{\theta}_q$	−2.39	251.28	3.95	3.16	86.2
		$\hat{\theta}$	−2.39	251.68	3.93	3.16	87.1
2000	1	$\hat{\theta}_h$	−2.65	−5.83	2.28	2.25	94.3
		$\hat{\theta}_q$	−2.66	−17.61	2.69	2.68	95.3
		$\hat{\theta}$	−2.65	−9.46	2.55	2.55	94.8
	2	$\hat{\theta}_h$	−2.65	−5.83	2.28	2.25	94.3
		$\hat{\theta}_q$	−2.40	243.48	3.22	2.14	79.0
		$\hat{\theta}$	−2.65	−5.31	2.40	2.11	92.3
	3	$\hat{\theta}_h$	−2.40	249.97	3.25	2.08	77.3
		$\hat{\theta}_q$	−2.66	−17.61	2.69	2.68	95.3
		$\hat{\theta}$	−2.66	−14.11	2.76	2.59	94.1
	4	$\hat{\theta}_h$	−2.40	249.97	3.25	2.08	77.3
		$\hat{\theta}_q$	−2.40	243.48	3.22	2.14	79.0
		$\hat{\theta}$	−2.40	243.56	3.22	2.14	78.6

Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-1}; SE: average of standard error estimates, 10^{-1}; Coverage: 95% confidence interval coverage, %. Experiment 1: h and q correctly specified; experiment 2: q misspecified; experiment 3: h misspecified, experiment 4: h and q misspecified.

Table 2. *Percentages of missing body weight measurements at week* 44.

	SCALE		STEP-2	
	Liraglutide 3.0 mg	Placebo	Semaglutide 2.4 mg	Placebo
N	402	205	397	393
Missing (%)	21.39	43.41	9.82	10.69

These superiority trials are designed to show the efficacy of semaglutide and liraglutide for weight loss among overweight or obese adults with type-2 diabetes. However, the study populations in the RCTs can differ in practice due to the sampling of study participants. Since the studies were conducted 5 to 6 years apart, a concern for the transportability of the treatment effect is a potential drift in social determinants of health which are unmeasured in both RCTs. The main objective of the statistical analysis is to provide a head-to-head comparison of the treatments liraglutide versus semaglutide in the study population of STEP-2, taking into account the unobserved social determinants.

The outcome Y is chosen as the percentage change from baseline (week 0) to week 44 in body weight. This is the timepoint closest to the end of treatment where body weight is measured in both RCTs. In both RCTs, we imputed the body weight at week 44 with the last-observation-carried-forward principle. In order to perform an anchored comparison, we make the assumption that the placebos used in these studies do not have any meaningful difference in their effect on the outcome, despite the differences in the frequency of administration and the volume of injection. We restate the parameter $\theta = E\{Y(1) - Y(0) \mid S = 0\}$, where $A = 1$ means liraglutide, 3.0 mg, $A = 0$ means placebo, and $S = 0$ indicates the STEP-2 trial. To balance the study populations, we adjust for a set of baseline adjustment variables $X = \{$baseline body weight, age, sex, body-mass index, race, region, waist circumference, smoking status, duration of diabetes$\}$.

Our method further requires the selection of appropriate negative controls to account for these unobserved effect modifiers. For the adjustment proxy W, we choose the percentage of glycated hemoglobin (HbA1c), the fasting plasma glucose (FPG) level and the fasting insulin level at baseline. In a review of the impact of social determinants among type-2 diabetic patients in the United States, Walker et al. (2014) pointed out that many studies support the link between social determinants on glycemic control measured in HbA1c. For the reweighting proxy Z, we select the baseline low-density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol levels as well as the baseline level of triglycerides. Cholesterol level has previously been found to be linked to health systems factors and economic development in many countries worldwide (Venkitachalam et al., 2012). Furthermore, there is no evidence on the existence of causal pathways from the current lipid level of a person to the future body weight. Additional assumptions on the proxies are illustrated by the causal graph in Fig. 3. Note that we assume the levels of HbA1c, FPG and fasting insulin differ between the study populations only because the social determinants and possibly the baseline adjustment variables are distributed differently.

To diminish skewness, the measurements for FPG, fasting insulin, HLDL cholesterol, VLDL cholesterol and triglycerides were log-transformed. Specifically for the estimation of the bridge functions, the numerical variables among X were transformed into an orthogonal cubic basis, and ridge regularization was applied to the linear parameters. We compared the multiply robust proximal indirect comparison estimator to the standard doubly robust estimator proposed by Dahabreh et al. (2020), where CATE

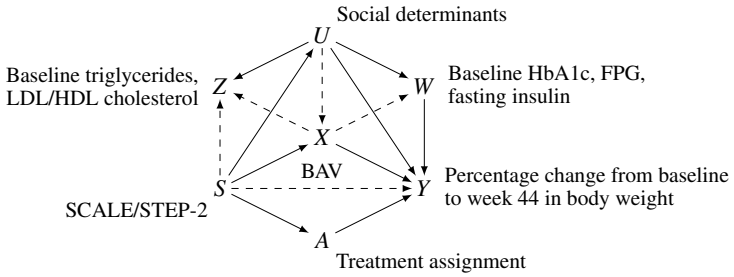

Figure 3. Hypothesized DAG of the observed and unobserved variables in the data example. The dashed arrows may or may not be present. BAV: baseline adjustment variables.

Table 3. *Results from the indirect comparison analysis with SCALE and STEP-2.*

Estimand	Standard	95%-CI	Proximal	95%-CI
$\theta = E\{Y(1) - Y(0) \mid S = 0\}$	-3.80	$(-4.59, -3.01)$	-3.82	$(-4.73, -2.90)$
$\gamma = E\{Y(1) - Y(-1) \mid S = 0\}$	-3.09	$(-4.20, -1.98)$	-3.08	$(-4.28, -1.87)$

Standard: estimators without the use of proxies detailed in Supplementary Material §S5; proximal: proximal indirect comparison estimators; CI: confidence interval.

transportability holds conditional on X. The subjects with missing measurements for X, Z or W and those with no body weight measurements beyond baseline were removed from the analysis, adding up to $41/846$ in SCALE and $26/1210$ in STEP-2. Table 2 shows that the percentages of missing outcomes are drastically different between the subjects randomized to the liraglutide, 3.0 mg arm and the placebo arm in the SCALE trial. The difference is less pronounced in the STEP-2 trial, where missingness in outcome measurements is much less frequent. To obtain estimates of the indirect comparison estimand θ, we further estimated the average treatment effect $E\{Y(-1) - Y(0) \mid S = 0\}$ within STEP-2 using the standard doubly robust estimator (Bang and Robins, 2005), where $A = -1$ stands for the once-weekly semaglutide, 2.4 mg treatment. A detailed description of the estimators and postulated nuisance models can be found in Supplementary Material §S5. Standard errors for all estimators were calculated as empirical L_2-norms of the corresponding influence functions.

The estimates and the corresponding 95% confidence intervals are displayed in Table 3. The proximal estimate and the standard estimate of θ show a similar weight loss effect of liraglutide at week 44. If the modelling assumptions hold, we may postulate that the unobserved social determinants have not altered the effect of the GLP-1 agnonists. The proximal estimate of the indirect comparison parameter γ is -3.08% versus the standard estimate of -3.09%, both indicating a stronger weight loss effect of semaglutide in the study population of the STEP-2 trial. Note, however, that the confidence interval based on the proximal estimator is slightly wider than that based on the standard estimator.

6. Discussion

In this article, we propose a novel method for indirect comparison in the presence of unmeasured, shifted effect modifiers. We tailor the proximal causal inference framework to the problem of indirect comparison, where the treatment of interest is not observed among the target population. We require IPD and in particular the existence of a pair of proxies in the source RCT and that of an adjustment proxy in the target RCT. The proximal indirect comparison estimator can be bias-free even when the CATE transportability fails to hold conditioning on the observed data. Despite the fact that the target parameter can be treated as a functional of the observed data, its interpretability depends on the underlying full data distribution.

A particular challenge for applying proximal indirect comparison is the selection of proxies in RCTs. For example, safety measurements and vital signs, which are routinely collected, usually do not affect the outcome, but they tend to be suboptimal proxy candidates because their distributions do not vary much between populations after controlling for baseline covariates. Data linkage would allow subjects from RCTs to be identified in the health registry, thereby providing far more potential proxies to choose from. Besides, the collection of proxy variables can be extended before and after the running period of the RCTs. When medical history is treated as a proxy, data linkage also helps avoid the use of self-reported data from questionnaires.

There are many interesting directions for future research. Throughout the development of the article we have assumed the availability of IPD in both RCTs. However, if only aggregate data can be obtained in one of the RCTs, the data likelihood changes and the bridge functions cannot be estimated with the same integral equations. A possible solution follows the calibration approach (Josey et al., 2021, 2022) to balance the moments of baseline covariates and proxies between RCTs. In longitudinal studies like SCALE and STEP-2 described in §5.2, subjects sometimes deviate from the treatment plan or drop out before the end of the studies. The extension of proximal indirect comparison to estimating the full compliance effect is straightforward, if one is willing to assume CATE transportability at baseline (Breskin et al., 2021) and no unmeasured time-varying confounding within RCTs. Treating noncompliance as a form of missingness at random, we extend our estimator to this case in Supplementary Material §S7. A more general transportability framework for observational longitudinal data under weaker causal assumptions can build on Ying et al. (2023). Finally beyond indirect comparisons, network meta-analyses may compare more than two active treatments from different studies. In a dense network, direct evidence can often be strengthened by indirect evidence. The formulation of proximal causal inference for data fusion is left for future work.

Acknowledgement

The authors thank Marie Thi Dao Tran from Novo Nordisk A/S for valuable input and discussions on the choices of proxies in the clinical trials SCALE and STEP-2.

Conflict of interest

Zehao Su is funded by a research gift from Novo Nordisk A/S to the Section of Biostatistics, University of Copenhagen. Henrik Ravn is employed by Novo Nordisk A/S.

Supplementary material

The Supplementary Material contains proofs, details of both the simulated data example and the real data example, additional simulations, and an extension of the method to missing outcomes.

References

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

Bennett, A., Kallus, N., Mao, X., Newey, W., Syrgkanis, V., and Uehara, M. (2023). Inference on strongly identified functionals of weakly identified functions. In *Proceedings of Thirty Sixth Conference on Learning Theory*, page 2265. PMLR.

Breskin, A., Cole, S. R., Edwards, J. K., Brookmeyer, R., Eron, J. J., and Adimora, A. A. (2021). Fusion designs and estimators for treatment effects. *Statistics in Medicine*, 40(13):3124–3137.

Carrasco, M., Florens, J.-P., and Renault, E. (2007). Chapter 77: Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6, pages 5633–5751. Elsevier.

Cinelli, C., Forney, A., and Pearl, J. (2024). A crash course in good and bad controls. *Sociological Methods & Research*, 53(3):1071–1104.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2024). Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1):165–191.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2024). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359.

Dahabreh, I. J., Robertson, S. E., Petito, L. C., Hernán, M. A., and Steingrimsson, J. A. (2023). Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics*, 79(2):1057–1072.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernán, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39(14):1999–2014.

Davies, M., Færch, L., Jeppesen, O. K., Pakseresht, A., Pedersen, S. D., Perreault, L., Rosenstock, J., Shimomura, I., Viljoen, A., Wadden, T. A., and Lingvay, I. (2021). Semaglutide 2.4 mg once a week in adults with overweight or obesity, and type 2 diabetes (STEP 2): A randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. *The Lancet*, 397(10278):971–984.

Davies, M. J., Bergenstal, R., Bode, B., Kushner, R. F., Lewin, A., Skjøth, T. V., Andreasen, A. H., Jensen, C. B., DeFronzo, R. A., and for the NN8022-1922 Study Group (2015). Efficacy of liraglutide for weight loss among patients with type 2 diabetes: The SCALE diabetes randomized clinical trial. *JAMA*, 314(7):687–699.

Ghassami, A., Yang, A., Richardson, D., Shpitser, I., and Tchetgen Tchetgen, E. (2022). Combining experimental and observational data for identification and estimation of long-term causal effects. *arXiv*: 2201.10743v3.

Ghassami, A., Yang, A., Shpitser, I., and Tchetgen Tchetgen, E. (2024). Causal inference with hidden mediators. *Biometrika*. Advance online publication.

Imbens, G., Kallus, N., Mao, X., and Wang, Y. (2024). Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. Advance online publication.

Josey, K. P., Berkowitz, S. A., Ghosh, D., and Raghavan, S. (2021). Transporting experimental results with entropy balancing. *Statistics in Medicine*, 40(19):4310–4326.

Josey, K. P., Yang, F., Ghosh, D., and Raghavan, S. (2022). A calibration approach to transportability and data-fusion with observational data. *Statistics in Medicine*, 41(23):4511–4531.

Kallus, N., Mao, X., and Uehara, M. (2022). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv*: 2103.14029v4.

Kress, R. (2014). *Linear integral equations*, volume 82 of *Applied Mathematical Sciences*. Springer.

Lipsitch, M., Tchetgen Tchetgen, E., and Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R., and Welton, N. J. (2018). Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making*, 38(2):200–211.

Phillippo, D. M., Dias, S., Ades, A. E., Belger, M., Brnabic, A., Schacht, A., Saure, D., Kadziola, Z., and Welton, N. J. (2020). Multilevel network meta-regression for population-adjusted treatment comparisons. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(3):1189–1210.

Venkitachalam, L., Wang, K., Porath, A., Corbalan, R., Hirsch, A. T., Cohen, D. J., Smith, S. C., Ohman, E. M., Steg, P. G., Bhatt, D. L., and Magnuson, E. A. (2012). Global variation in the prevalence of elevated cholesterol in outpatients with established vascular disease or 3 cardiovascular risk factors according to national indices of economic development and health system performance. *Circulation*, 125(15):1858–1869.

Walker, R. J., Smalls, B. L., Campbell, J. A., Strom Williams, J. L., and Egede, L. E. (2014). Impact of social determinants of health on outcomes for type 2 diabetes: A systematic review. *Endocrine*, 47(1):29–48.

Ying, A., Miao, W., Shi, X., and Tchetgen Tchetgen, E. J. (2023). Proximal causal inference for complex longitudinal studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):684–704.

Zhang, J., Li, W., Miao, W., and Tchetgen Tchetgen, E. (2023). Proximal causal inference without uniqueness assumptions. *Statistics & Probability Letters*, 198:109836.

Supplementary material for "Proximal indirect comparison"

Zehao Su
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
zehao.su@sund.ku.dk

Helene C. Rytgaard
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
hely@sund.ku.dk

Henrik Ravn
Novo Nordisk A/S,
Vandtårnsvej 108, DK-2860 Søborg, Denmark
hnrv@novonordisk.com

Frank Eriksson
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
eriksson@sund.ku.dk

S1. M-bias in transportability

Consider a family of structural equation models parameterized by k ($k \neq 0$) as follows:

$$\begin{aligned} S &\leftarrow \epsilon_S, \\ U &\leftarrow k^{-1}\epsilon_U, \\ A &\leftarrow S\epsilon_{A,1} - (1-S)\epsilon_{A,2}, \\ X &\leftarrow S + kU + \epsilon_X, \\ Y(a) &\leftarrow aX + aU + X + U + S + \epsilon_{Y(a)}, \quad a \in \{-1, 0, 1\}, \end{aligned}$$

where the random errors are mutually independent,

$$\begin{aligned} \epsilon_S, \epsilon_{A,1}, \epsilon_{A,2} &\sim \text{Bernoulli}(0.5), \\ \epsilon_U, \epsilon_X, \epsilon_{Y(a)} &\sim \text{Normal}(0, 1). \end{aligned}$$

In this data model, X and U are effect modifiers with respect to the average treatment effect (ATE).

Without conditioning on $\{X, U\}$, $E\{Y(1) - Y(0) \mid S = s\} = E(X + U + \epsilon_{Y(1)} - \epsilon_{Y(0)} \mid S = s) = E(X \mid S = s) = s$. If we were to wrongly assume external validity of the ATE from the source randomized control trial (RCT) and mistake the value of $\theta_1^{\text{mis}} = E\{Y(1) - Y(0) \mid S = 1\}$ for $\theta = E\{Y(1) - Y(0) \mid S = 0\}$, the bias would be $\theta_1^{\text{mis}} - \theta = 1$. When we also observe baseline covariates X, it is tempting to hypothesize conditional average treatment effect (CATE) transportability instead and attempt to identify θ by way of this assumption. The density of U conditioning on $\{X, S\}$ is

$$p(u \mid x, s) \propto p(x \mid u, s)p(u)$$

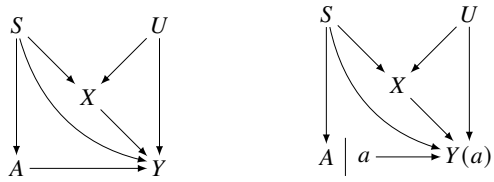

Figure S1. A DAG (left) and its corresponding SWIG (right) illustrating possible M-bias in transportability.

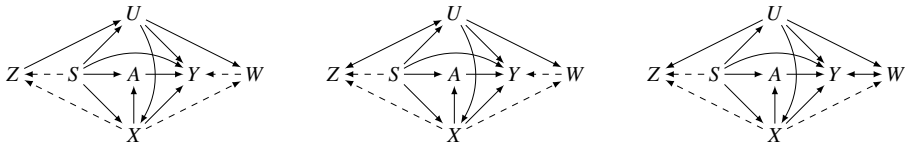

Figure S2. Acyclic directed mixed graphs compatible with Assumption 2.

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2}(x-s-ku)^2\right\}\exp\left(-\frac{k^2u^2}{2}\right)\\
&= \exp\left\{-\frac{1}{2(2k^2)^{-1}}\left(u^2+\frac{s-x}{k}u\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2(2k^2)^{-1}}\left(u-\frac{x-s}{2k}\right)^2\right\},
\end{aligned}
$$

so $U \mid (X = x, S = s) \sim \text{Normal}\{(x-s)/2k, (2k^2)^{-1}\}$. We have the expression for the CATE $E\{Y(1)-Y(0) \mid X = x, S = s\} = x + E(U + \epsilon_{Y(1)} - \epsilon_{Y(0)} \mid X = x, S = s) = x + E(U \mid X = x, S = s) = x + (x-s)/(2k)$. In this case, the CATE is clearly not transportable. However, if we were to mistakenly use the value of $E\{Y(1)-Y(0) \mid X = x, S = 1\}$ in lieu of $E\{Y(1)-Y(0) \mid X = x, S = 0\}$, we would have a wrongly identified parameter θ_2^{mis} and the bias would be $\theta_2^{\text{mis}} - \theta = -1/(2k)$. Therefore, when $|k| > 1/2$, the absolute bias is amplified due to the adjustment for X. Note also that the direction of the bias is flipped when $k > 0$.

In Fig. S1, we present a directed acyclic graph (DAG) and its corresponding single-world intervention graph (SWIG) that are compatible with the family of structural equations above. Since, S and U are d-separated, there is no marginal dependence between S and U; that is, the effect modifier U is not shifted. However, when we condition on X, the paths $S \to X \leftarrow U \to Y$ and $S \to X \leftarrow U \to Y(a)$ are opened. CATE transportability conditioning on X is unlikely to hold given potential association between S and Y (or $Y(a)$) through effect modifiers $\{X, U\}$.

S2. Asymptotic theory for estimators of bridge functions

In this section, we present useful results on the asymptotics of the proximal indirect comparison estimators in the simulation study. In particular, we will derive their asymptotic variances under the assumption of parametric bridge functions. Without loss of generality, we assume the parametric components have the same dimension.

Assumption S1 (Parametric bridge functions). For every $P \in \mathcal{P}$, $\mathcal{H} = \{h_{\eta_0}\}$ and $\mathcal{Q} = \{q_{\xi_0}\}$ are singletons, where $\eta_0, \xi_0 \in \mathbb{R}^k$ are Euclidean parameters.

We choose some basis expansions $b_*(z,x)$ and $c_*(w,x)$ in $\mathbb{R}^k$ and define the functions

$$\psi_\eta^h(o) = s b_*(z,x)\{\tilde{y} - h_\eta(w,x)\},$$

$$\psi^q_\xi(o) = c_*(w,x)\{(1-s) - sq_\xi(z,x)\}.$$

The Z-estimators $\hat{\eta}$ and $\hat{\xi}$ are such that $\|P_n\psi^h_{\hat{\eta}}\| = o_P(n^{-1/2})$ and $\|P_n\psi^q_{\hat{\xi}}\| = o_P(n^{-1/2})$. Note that the estimation of ξ_0 need not involve nuisance models for the participation odds $\mathrm{pr}(S=0 \mid W,X)/\mathrm{pr}(S=1 \mid W,X)$. A similar point was raised in the estimation of parametric treatment bridge functions to account for unmeasured confounding Cui et al. (2024).

Assumption S2 (Regularity conditions).

(i) For all $\eta \in \mathbb{R}^k$, the map $\eta \mapsto h_\eta(w,x)$ is differentiable with derivative $\dot{h}_\eta$ for all (w,x), so that $\dot{\psi}^h_\eta = -sb_*\dot{h}_\eta$. $E\sup_{\eta\in K}\|\dot{\psi}^h_\eta\|^2 < \infty$ for every compact $K \subset \mathbb{R}^k$. $\|E\psi^h_\eta\|$ has a unique zero point at η_0, and $E\dot{\psi}^h_{\eta_0}$ is invertible.

(ii) For all $\xi \in \mathbb{R}^k$, the map $\xi \mapsto q_\xi(w,x)$ is differentiable with derivative $\dot{q}_\xi$ for all (z,x), so that $\dot{\psi}^q_\xi = -sc_*\dot{q}_\xi$. $E\sup_{\xi\in K}\|\dot{\psi}^q_\xi\|^2 < \infty$ for every compact $K \subset \mathbb{R}^k$. $\|E\psi^q_\xi\|$ has a unique zero point at ξ_0, and $E\dot{\psi}^q_{\xi_0}$ is invertible.

Proposition S1. *Suppose Assumptions S1–S2 hold. Then:*

1. *The estimator $\hat{\eta}$ is asymptotically normal with influence function $-(E\dot{\psi}^h_{\eta_0})^{-1}\psi^h_{\eta_0}$. The estimator $\hat{\theta}_h$ is asymptotically normal with influence function*

$$\frac{1-S}{\alpha}\{h_{\eta_0} - \theta\} - \frac{1}{\alpha}E(S\dot{h}^{\mathrm{T}}_{\eta_0})(E\dot{\psi}^h_{\eta_0})^{-1}\psi^h_{\eta_0}.$$

2. *The estimator $\hat{\xi}$ is asymptotically normal with influence function $-(E\dot{\psi}^q_{\xi_0})^{-1}\psi^q_{\xi_0}$. The estimator $\hat{\theta}_q$ is asymptotically normal with influence function*

$$\frac{1}{\alpha}\left\{sq_{\xi_0}\tilde{y} - (1-s)\theta\right\} - \frac{1}{\alpha}E(S\tilde{Y}\dot{q}^{\mathrm{T}}_{\xi_0})(E\dot{\psi}^q_{\xi_0})^{-1}\psi^q_{\xi_0}.$$

Proof. A first-order Taylor expansion of the estimating equations around the true parameters η_0 and ξ_0 shows their influence functions. The influence functions of $\hat{\theta}_h$ and $\hat{\theta}_q$ can be obtained by the same technique, with the exception that $\hat{\theta}_q$ requires an additional expansion around α. □

S3. Existence and uniqueness of bridge functions

To make statistical inference on the target parameter, we rely on existence and uniqueness of the bridge functions h_0 and q_0. For the sake of completness, we state sufficient conditions for this assumption.

We make use of the following completeness conditions on the observed data distribution, under which $\mathcal{H}$ and $\mathcal{Q}$ must be either empty sets or singletons.

Assumption S3 (Completeness). P_1-almost surely:

(i) $E\{g(W,X) \mid Z,X,S=1\} = 0$ implies $g(W,X) = 0$;

(ii) $E\{g(Z,X) \mid W,X,S=1\} = 0$ implies $g(Z,X) = 0$.

Similar completeness assumptions appear in many works on proximal causal inference (Cui et al., 2024; Tchetgen Tchetgen et al., 2024). They use completeness assumptions to translate observed data parameters into causal parameters. In our setup, this corresponds to viewing the identification formulas in Proposition 2 directly as parameters of interest. Then, with Assumption 2 and completeness assumptions on the densities $p(u \mid W=w, X=x, S=1)$ and $p(u \mid Z=z, X=x, S=1)$, we arrive at $\mathcal{H} = \mathcal{H}^{\mathrm{U}}$ and $\mathcal{Q} = \mathcal{Q}^{\mathrm{U}}$. That is, under these alternative assumptions, the two parameters of interest will have the right causal interpretation.

We first introduce some notations. Let $P_1|_{X=x}$ denote the conditional probability measure $P_1(\cdot \mid X = x)$. For every fixed x, define the linear transformation $T_x : L_2(W; P_1|_{X=x}) \to L_2(Z; P_1|_{X=x})$ such that $(T_x h)(z) = E\{h(W) \mid Z = z, X = x\}$. The adjoint of T_x is then $T_x^* : L_2(Z; P_1|_{X=x}) \to L_2(W; P_1|_{X=x})$ such that $(T_x^* q)(w) = E\{q(Z) \mid W = w, X = x\}$. When T_x and T_x^* are compact operators, there exist orthonormal sequences (f_k) in $L_2(Z; P_1|_{X=x})$ and (g_k) in $L_2(W; P_1|_{X=x})$ and a positive sequence of real numbers (σ_k) such that $T_x g_k = \sigma_k f_k$ and $T_x^* f_k = \sigma_k g_k$ for all positive integers k (Kress, 2014, Theorem 15.16).

Assumption S4 (Regularity conditions). For all x:

(i) T_x and T_x^* are compact operators;
(ii) $E(\tilde{Y} \mid Z, X = x, S = 1) \in L_2(Z; P_1|_{X=x})$;
(iii) $\{\text{pr}(S = 1 \mid W, X = x)\}^{-1}\text{pr}(S = 0 \mid W, X = x) \in L_2(W; P_1|_{X=x})$;
(iv) $\sum_{k=1}^{\infty} \sigma_k^{-2} \left|\langle E(\tilde{Y} \mid Z, X = x, S = 1), f_k\rangle_{L_2(Z;P_1|_{X=x})}\right|^2 < \infty$;
(v) $\sum_{k=1}^{\infty} \sigma_k^{-2} \left|\left\langle \{\text{pr}(S = 1 \mid W, X = x)\}^{-1}\text{pr}(S = 0 \mid W, X = x), g_k\right\rangle_{L_2(W;P_1|_{X=x})}\right|^2 < \infty$.

Proposition S2 (Identifiability of bridge functions).

1. *Under Assumptions S3(i), S4(i), S4(ii), and S4(iv), $\mathcal{H}$ is nonempty and a singleton.*
2. *Under Assumptions S3(ii), S4(i), S4(iii), and S4(v), $\mathcal{Q}$ is nonempty and a singleton.*

Proof. We only show the proof of the first index of Proposition S2. The singular value decomposition of T_x exists by Assumption S4(i). By definition, the nonemptiness of $\mathcal{H}$ follows from the existence of a solution to the linear integral equation $(T_x h)(z) = E(\tilde{Y} \mid Z = z, X = x, S = 1)$ in the Hilbert space $L_2(Z; P_1|_{X=x})$ for every x. Applying Picard's Theorem (Theorem 15.18 in Kress (2014)), the equation for h has a solution due to Assumption S4(ii) and S4(iv). We will use proof by contradiction to show the second part of the statement. Suppose the contrary that there exist two solutions $h \neq h'$ to the equation in $\mathcal{H}$, such that $E(h - h' \mid Z, X = x, S = 1) = 0$ holds $P_1(Z \mid X = x)$-almost surely. Then by AssumptionS3(i), we must have $h = h'$ almost surely. The proof for the second index can be similarly obtained using the assumptions shown in Proposition S2. □

In general, the operators T_x and T_x^* are not compact operators. A sufficient condition for Assumption S4(i) is $\iint p(w \mid z, X, S = 1)p(z \mid w, X, S = 1)\text{d}w\text{d}z < \infty$, $P_1(X)$-almost surely [see Example 2.3 in Carrasco et al. (2007)]. In this case, T_x is a Hilbert-Schmidt operator, which is guaranteed to be compact.

S4. Details of the simulated data example

S4.1. *Underlying bridge functions*

The baseline covariates X, the unobserved variables U, the negative control outcomes W, and the negative control treatments Z are multivariate. The rest of the variables are univariate. We use b, β, and B for scalar, vector, and matrix coefficients. Their dimensions should be clear from the context. We generated the data sequentially according to (excluding X and U, which can follow arbitrary joint distributions):

$$
\begin{aligned}
S \mid (X, U) &\sim \text{Bernoulli}\{\text{expit}(b_s + \beta_{sx}^{\mathrm{T}} X + \beta_{su}^{\mathrm{T}} U)\},\\
A \mid (X, U, S = 1) &\sim \text{Bernoulli}(b_a), 0 < b_a < 1,\\
Z \mid (X, U, S = 1) &\sim \text{Normal}(\beta_z + B_{zu} U + B_{zx} X, \Sigma_z),\\
W \mid (X, U) &\sim \text{Normal}(\beta_w + B_{wu} U + B_{wx} X, \Sigma_w),\\
Y \mid (W, A, X, U, S = 1) &\sim \text{Normal}(b_y + b_{ya} A + \beta_{yu}^{\mathrm{T}} U + \beta_{y(au)}^{\mathrm{T}} (AU) + \beta_{yx}^{\mathrm{T}} X
\end{aligned}
$$

$$+ \beta_{yw}^{\mathrm{T}} W + \beta_{y(aw)}^{\mathrm{T}} (AW), \sigma_y^2).$$

The conditional distribution $W \mid (A, X, U, S = 1)$ is the same as the conditional distribution of $W \mid (X, U, S = 1)$. The conditional expectation

$$\begin{aligned}
&E(Y \mid A = 1, X, U, S = 1) - E(Y \mid A = 0, X, U, S = 1) \\
&\quad = E\{E(Y \mid W, A = 1, X, U, S = 1) - E(Y \mid W, A = 0, X, U, S = 1) \mid X, U, S = 1\} \\
&\quad = b_{ya} + \beta_{y(au)}^{\mathrm{T}} U + \beta_{y(aw)}^{\mathrm{T}} E(W \mid X, U, S = 1) \\
&\quad = b_{ya} + \beta_{y(aw)}^{\mathrm{T}} \beta_w + (\beta_{y(au)}^{\mathrm{T}} + \beta_{y(aw)}^{\mathrm{T}} B_{wu}) U + \beta_{y(aw)}^{\mathrm{T}} B_{wx} X.
\end{aligned}$$

Let the outcome bridge function $h(W, X) = \eta_0 + \eta_w^{\mathrm{T}} W + \eta_x^{\mathrm{T}} X$, then

$$\begin{aligned}
E\{h(W, X) \mid X, U, S = 1\} &= \eta_0 + \eta_x^{\mathrm{T}} X + \eta_w^{\mathrm{T}} E(W \mid X, U, S = 1) \\
&= \eta_0 + \eta_w^{\mathrm{T}} \beta_w + (\eta_x^{\mathrm{T}} + \eta_w^{\mathrm{T}} B_{wx}) X + \eta_w^{\mathrm{T}} B_{wu} U.
\end{aligned}$$

Comparing the coefficients in the two expressions, we have the following system of equations:

$$\begin{aligned}
\eta_0 + \eta_w^{\mathrm{T}} \beta_w &= b_{ya} + \beta_{y(aw)}^{\mathrm{T}} \beta_w \\
\eta_x^{\mathrm{T}} + \eta_w^{\mathrm{T}} B_{wx} &= \beta_{y(aw)}^{\mathrm{T}} B_{wx} \\
\eta_w^{\mathrm{T}} B_{wu} &= \beta_{y(au)}^{\mathrm{T}} + \beta_{y(aw)}^{\mathrm{T}} B_{wu},
\end{aligned}$$

so the parameters of the bridge function are

$$\begin{aligned}
\eta_0 &= b_{ya} - \beta_w^{\mathrm{T}} B_{wu}^{-\mathrm{T}} \beta_{y(au)}, \\
\eta_x &= -B_{wx}^{\mathrm{T}} B_{wu}^{-\mathrm{T}} \beta_{y(au)}, \\
\eta_w &= B_{wu}^{-\mathrm{T}} \beta_{y(au)} + \beta_{y(aw)}.
\end{aligned}$$

The probability ratio

$$\frac{\mathrm{pr}(S = 0 \mid X, U)}{\mathrm{pr}(S = 1 \mid X, U)} = \exp(-b_s - \beta_{sx}^{\mathrm{T}} X - \beta_{su}^{\mathrm{T}} U).$$

Let the participation bridge function be

$$q(Z, X) = \exp(\xi_0 + \xi_z^{\mathrm{T}} Z + \xi_x^{\mathrm{T}} X),$$

then

$$\begin{aligned}
E\{q(Z, X) \mid X, U, S = 1\} &= \exp(\xi_0 + \xi_x^{\mathrm{T}} X) E\{\exp(\xi_z^{\mathrm{T}} Z) \mid X, U, S = 1\} \\
&= \exp(\xi_0 + \xi_x^{\mathrm{T}} X) \exp\left\{\xi_z^{\mathrm{T}} (\beta_z + B_{zu} U + B_{zx} X) + \frac{1}{2} \xi_z^{\mathrm{T}} \Sigma_z \xi_z\right\} \\
&= \exp\left\{\xi_0 + \xi_z^{\mathrm{T}} \beta_z + \frac{1}{2} \xi_z^{\mathrm{T}} \Sigma_z \xi_z + (\xi_x^{\mathrm{T}} + \xi_z^{\mathrm{T}} B_{zx}) X + \xi_z^{\mathrm{T}} B_{zu} U\right\},
\end{aligned}$$

where we have used the moment generating function of the conditional distribution $Z \mid (X, U, S = 1)$. Comparing the coefficients in the two expressions, we have the following system of equations:

$$\begin{aligned}
\xi_0 + \xi_z^{\mathrm{T}} \beta_z + \frac{1}{2} \xi_z^{\mathrm{T}} \Sigma_z \xi_z &= -b_s \\
\xi_x^{\mathrm{T}} + \xi_z^{\mathrm{T}} B_{zx} &= -\beta_{sx}^{\mathrm{T}} \\
\xi_z^{\mathrm{T}} B_{zu} &= -\beta_{su}^{\mathrm{T}},
\end{aligned}$$

so the parameters of the bridge function are

$$\begin{aligned}
\xi_0 &= -b_s + \beta_z^{\mathrm{T}} B_{zu}^{-\mathrm{T}} \beta_{su} - \frac{1}{2} \beta_{su}^{\mathrm{T}} B_{zu}^{-1} \Sigma_z B_{zu}^{-\mathrm{T}} \beta_{su}, \\
\xi_x &= -\beta_{sx} + B_{zx}^{\mathrm{T}} B_{zu}^{-\mathrm{T}} \beta_{su}, \\
\xi_z &= -B_{zu}^{-\mathrm{T}} \beta_{su}.
\end{aligned}$$

S4.2. *Justification for the cubic of basis*

Suppose we obtain an estimator of ξ_0 without using the cubic of the basis $c(w,x)$, so that

$$\tilde{\xi} = \arg\min_{\xi} \left\| \frac{1}{n} \sum_{i=1}^{n} c(W_i, X_i)\{S_i q_{\xi}(Z_i, X_i) - (1 - S_i)\} \right\|^2.$$

The first-order condition of the optimization problem gives

$$2\left\{ \frac{1}{n} \sum_{i=1}^{n} S_i q_{\tilde{\xi}}(Z_i, X_i) c(W_i, X_i) c^{\mathrm{T}}(W_i, X_i) \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} c(W_i, X_i)\{S_i q_{\tilde{\xi}}(Z_i, X_i) - (1 - S_i)\} \right\} = 0.$$

Since the matrix in the first pair of braces is almost surely nonsingular, we have that with probability 1,

$$\frac{1}{n} \sum_{i=1}^{n} c(W_i, X_i)\{S_i q_{\tilde{\xi}}(Z_i, X_i) - (1 - S_i)\} = 0. \tag{S1}$$

We compare the two estimators

$$\tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{S_i}{\hat{\alpha}} q_{\tilde{\xi}}(Z_i, X_i)\{\tilde{Y}_i - h_{\hat{\eta}}(W_i, X_i)\} + \frac{1 - S_i}{\hat{\alpha}} h_{\hat{\eta}}(W_i, X_i) \right],$$

$$\tilde{\theta}_q = \frac{1}{n_0} \sum_{i:S_i=1} q_{\tilde{\xi}}(Z_i, X_i)\tilde{Y}_i,$$

which are the same as $\hat{\theta}$ and $\hat{\theta}_q$, except $\hat{\xi}$ is replaced by $\tilde{\xi}$. Their difference is

$$\tilde{\theta}_q - \tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\hat{\alpha}} \{S_i q_{\tilde{\xi}}(Z_i, X_i) - (1 - S_i)\} h_{\hat{\eta}}(W_i, X_i).$$

Because $h_{\hat{\eta}}(w,x) = \hat{\eta}^{\mathrm{T}} c(w,x)$ is a linear combination of $c(w,x)$, the observation in (S1) shows that $\tilde{\theta}_q = \tilde{\theta}$ almost surely. However, this can be circumvented by using a nonlinear transformation of $c(w,x)$ as the basis function to estimate ξ_0, as was done in the simulation study, where we raised $c(w,x)$ to the third power elementwise.

S4.3. *Additional experiments under assumption violations*

In experiment 5, we investigated the behaviour of proximal indirect comparison estimators in the absence of unmeasured effect modifiers, where U was set as a zero vector. To understand the impact of the violation of the proxy assumptions (Assumption 2(iii) and 2(iv) in the main text), we replaced the conditional distribution of Y with $Y \mid (Z, W, A, U, S = 0) \sim \mathrm{Normal}(0.5 - A + U^{\mathrm{T}}1 + AU^{\mathrm{T}}1 + X^{\mathrm{T}}1 + W^{\mathrm{T}}1 + AW^{\mathrm{T}}1 + Z^{\mathrm{T}}1 + AZ^{\mathrm{T}}1, 0.5^2)$ in experiment 7 and the condition distribution of W with $W \mid (X, S) \sim \mathrm{Normal}(S1 + X + U, 0.25\mathrm{Id})$ in experiment 8. In experiment 9, we simulated $U \sim \mathrm{Uniform}([-1, 0])$ as a scalar-valued random variable but maintained the proxies as vectors, so that the bridge functions are no longer uniquely identified. The importance of the existence of the bridge functions was studied in experiment 11 by simulating Z from $Z \mid (U, X, S = 0) \sim \mathrm{Normal}(0.05U + X, 0.25\mathrm{Id})$, making Z nearly uncorrelated with U given X in the source RCT. Likewise in experiment 12, we simulated W from $W \mid (U, X) \sim \mathrm{Normal}(0.05U + X, 0.25\mathrm{Id})$ so that W is nearly uncorrelated with U given X. In experiment 13, we examined the effect on near violation of positivity by changing the conditional probability $\mathrm{pr}(S = 1 \mid U, X)$ to $\mathrm{expit}(-0.675 + 0.5X^{\mathrm{T}}1 + 2.5U^{\mathrm{T}}1)$ so that the participation odds $\mathrm{pr}(S = 0 \mid U, X)/\mathrm{pr}(S = 1 \mid U, X)$ is large, as the coefficient of U is dispropotionally large. Finally in experiments 6 and 10, the data

Table S1. *Additional simulation results of experiments 5–13 with sample size n = 1000.*

n	Experiment	Estimator	Mean	Bias	RMSE	SE	Coverage
1000	5	$\hat{\theta}_h$	0.40	−3.29	60.24	585.37	100.00
		$\hat{\theta}_q$	0.38	−25.55	9.79	1318.56	97.39
		$\hat{\theta}$	0.17	−232.71	140.78	65.29	98.59
	6	$\hat{\theta}_h$	0.40	−8.56	3.39	63.76	99.80
		$\hat{\theta}_q$	0.39	−12.37	3.36	42.13	99.50
		$\hat{\theta}$	0.39	−13.32	3.36	3.31	94.70
	7	$\hat{\theta}_h$	−2.84	−93.51	4.87	4.82	95.30
		$\hat{\theta}_q$	−2.87	−116.60	6.41	6.27	95.70
		$\hat{\theta}$	−2.84	−93.01	5.83	5.78	94.80
	8	$\hat{\theta}_h$	−5.45	−2802.40	52.29	44.69	90.60
		$\hat{\theta}_q$	−2.01	631.36	33.84	1.65	5.48
		$\hat{\theta}$	−5.13	−2487.30	53.51	28.60	55.01
	9	$\hat{\theta}_h$	−1.83	−169.06	43.99	379.44	100.00
		$\hat{\theta}_q$	−1.70	−37.91	5.68	150.45	96.70
		$\hat{\theta}$	−1.86	−206.71	56.19	28.78	97.49
	10	$\hat{\theta}_h$	−1.65	3.76	2.39	28.39	99.70
		$\hat{\theta}_q$	−1.66	−5.54	2.56	27.70	99.50
		$\hat{\theta}$	−1.66	−0.08	2.51	2.47	94.20
	11	$\hat{\theta}_h$	−2.73	−79.91	72.12	725.79	100.00
		$\hat{\theta}_q$	−2.52	121.77	10.02	24.72	90.67
		$\hat{\theta}$	−2.64	8.94	73.71	41.25	95.49
	12	$\hat{\theta}_h$	−1.09	85.97	60.62	582.12	100.00
		$\hat{\theta}_q$	−1.10	84.36	6.68	947.45	96.40
		$\hat{\theta}$	−1.31	−134.55	70.33	40.68	99.00
	13	$\hat{\theta}_h$	−2.67	−111.00	31.82	44.00	94.10
		$\hat{\theta}_q$	−2.22	336.45	19.65	7.44	44.10
		$\hat{\theta}$	−2.43	130.62	29.32	18.92	83.47

Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-1}; SE: average of standard error estimates, 10^{-1}; Coverage: 95% confidence interval coverage, %.

were simulated as in experiments 5 and 9, whereas the bridge functions were estimated with ridge regularization. That is, the fitted bridge functions were $h_{\hat{\eta}_\lambda}(W, X)$ and $q_{\hat{\xi}_\lambda}(Z, X)$ where

$$\hat{\eta}_\lambda = \arg\min_{\eta'} \left\| \frac{1}{n_1} \sum_{i:S_i=1} b(Z_i, X_i)\{\tilde{Y}_i - h_{\eta'}(W_i, X_i)\} \right\|^2 + \lambda_h (\eta')^{\mathrm{T}} D_h \eta',$$

$$\hat{\xi}_\lambda = \arg\min_{\xi'} \left\| \frac{1}{n} \sum_{i=1}^{n} \{c(W_i, X_i)\}^3 \{S_i q_{\xi'}(Z_i, X_i) - (1 - S_i)\} \right\|^2 + \lambda_q (\xi')^{\mathrm{T}} D_q \xi',$$

with fixed regularization parameters $\lambda_h = \lambda_q = 10^{-4}$ and D_h and D_q being identity matrices of appropriate dimensions with their upper left corners changed to zero, so that the intercept is unpenalized. All results from the additional simulation studies are displayed in Tables S1 and S2.

S5. Details of the real data example

S5.1. *Description of bridge estimators and standard estimators*

Let $m(a, x, s) = E(Y \mid A = a, X = x, S = s)$ and by an abuse of notation $p(x) = \mathrm{pr}(S = 1 \mid X = x)$, $e(a \mid x, s) = \mathrm{pr}(A = a \mid X = x, S = s)$. We assumed linear models for m and $\log\{p/(1-p)\}$. Let $\tilde{X}$ denote the design vector without intercept from the baseline adjusting

Table S2. *Additional simulation results of experiments 5–13 with sample size n = 2000.*

n	Experiment	Estimator	Mean	Bias	RMSE	SE	Coverage
2000	5	$\hat{\theta}_h$	0.56	154.81	37.69	367.80	100.00
		$\hat{\theta}_q$	0.40	−2.34	6.72	59.25	98.99
		$\hat{\theta}$	0.42	9.49	91.75	39.15	98.79
	6	$\hat{\theta}_h$	0.42	11.12	2.45	84.90	99.80
		$\hat{\theta}_q$	0.41	6.37	2.42	33.26	99.80
		$\hat{\theta}$	0.41	4.56	2.42	2.32	92.80
	7	$\hat{\theta}_h$	−2.85	−104.03	3.59	3.37	93.80
		$\hat{\theta}_q$	−2.87	−124.49	4.40	4.19	95.40
		$\hat{\theta}$	−2.86	−110.04	4.08	3.90	94.50
	8	$\hat{\theta}_h$	−5.60	−2955.11	42.89	31.01	84.30
		$\hat{\theta}_q$	−2.35	293.72	39.12	2.15	4.41
		$\hat{\theta}$	−5.29	−2649.47	51.14	32.02	62.07
	9	$\hat{\theta}_h$	−1.58	77.82	20.55	144.31	99.70
		$\hat{\theta}_q$	−1.68	−24.05	4.03	548.32	97.00
		$\hat{\theta}$	−1.67	−16.70	26.86	17.81	98.30
	10	$\hat{\theta}_h$	−1.66	−3.03	1.64	18.46	99.20
		$\hat{\theta}_q$	−1.67	−8.20	1.78	24.66	99.40
		$\hat{\theta}$	−1.66	−7.20	1.73	1.74	95.20
	11	$\hat{\theta}_h$	−2.78	−131.67	75.44	528.18	100.00
		$\hat{\theta}_q$	−2.51	136.13	8.89	1529.49	92.30
		$\hat{\theta}$	−2.59	51.16	67.84	27.79	92.29
	12	$\hat{\theta}_h$	−1.08	98.20	43.26	293.16	100.00
		$\hat{\theta}_q$	−1.11	71.73	4.91	494.91	97.29
		$\hat{\theta}$	−0.93	249.01	68.86	29.61	98.99
	13	$\hat{\theta}_h$	−2.58	−17.43	12.05	11.62	94.80
		$\hat{\theta}_q$	−2.61	−53.22	17.31	11.60	51.69
		$\hat{\theta}$	−2.66	−99.47	17.48	15.17	88.39

Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-1}; SE: average of standard error estimates, 10^{-1}; Coverage: 95% confidence interval coverage, %.

Table S3. *Additional estimates from the indirect comparison analysis with SCALE and STEP-2.*

Estimand	Estimate	95%-CI
$E\{Y(-1) - Y(0) \mid S = 0\}$	−6.90	(−7.68, −6.11)
$E\{Y(1) - Y(0) \mid S = 1\}$	−3.97	(−4.71, −3.22)

The estimands are direct comparisons of the treatments administered in the respective trials.

variables X with numerical variables transformed into the orthogonal cubic basis and categorical variables transformed into dummy variables. The bridge functions were assumed to follow the parametric forms $h(w,x) = h_\eta(w,x) = \eta^{\mathrm{T}}\tilde{c}(w,x)$ and $q(z,x) = q_\xi(z,x) = \xi^{\mathrm{T}}\tilde{b}(z,x)$, where $\tilde{c}(w,x) = (1,w,\tilde{x}^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{b}(z,x) = (1,z,\tilde{x}^{\mathrm{T}})^{\mathrm{T}}$.

The linear parameters in the bridge functions are fitted using the ridge-regularized generalized method of moment such that

$$\hat{\eta}_\lambda = \arg\min_{\eta} \left\| \frac{1}{n_1} \sum_{i:S_i=1} \tilde{b}(Z_i, X_i)\{\tilde{Y}_i - h_{\eta'}(W_i, X_i)\} \right\|^2 + \lambda_{h,n}\eta^{\mathrm{T}} D_h \eta,$$

$$\hat{\xi}_\lambda = \arg\min_{\xi'} \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{c}(W_i, X_i)\{S_i q_{\xi'}(Z_i, X_i) - (1 - S_i)\} \right\|^2 + \lambda_{q,n}\xi^{\mathrm{T}} D_q \xi,$$

where D_h and D_q are identity matrices of appropriate dimensions with their upper left corners changed to zero. The data-adaptive regularization factors $\lambda_{h,n}$ and $\lambda_{q,n}$ are chosen with 10-fold cross validation from a prespecified grid. The models for $m(a,\cdot,s)$ are fitted separately on the SCALE and STEP-2 samples as well as on each treatment arm to allow for full interaction between $\{A, S\}$ and the other variables. Then the modified target population ATE estimator from Dahabreh et al. (2020) for θ is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{S_i}{\hat{\alpha}}\frac{1-\hat{p}(X_i)}{\hat{p}(X_i)}\frac{(2A_i-1)}{e(A_i \mid S_i)}\{Y_i - \hat{m}(A_i, X_i, S_i)\} + \frac{1-S_i}{\hat{\alpha}}\{\hat{m}(1, X_i, S_i) - \hat{m}(0, X_i, S_i)\}\right],$$

and the modified standard doubly robust ATE estimator from Bang and Robins (2005) for the parameter $E\{Y(-1) - Y(0) \mid S = 0\}$ is

$$\frac{1}{n_0}\sum_{i:S_i=0}\left[\frac{(-2A_i-1)}{e(A_i \mid 0)}\{Y_i - \hat{m}(A_i, X_i, 0)\} + \{\hat{m}(-1, X_i, 0) - \hat{m}(0, X_i, 0)\}\right].$$

For comparison, we also computed the ATE estimate for $E\{Y(1) - Y(0) \mid S = 1\}$ by

$$\frac{1}{n_1}\sum_{i:S_i=1}\left[\frac{(2A_i-1)}{e(A_i \mid 1)}\{Y_i - \hat{m}(A_i, X_i, 1)\} + \{\hat{m}(1, X_i, 1) - \hat{m}(0, X_i, 1)\}\right].$$

The last two estimates are reported along with 95% confidence intervals in Table S3 for reference.

S5.2. *PRISMA-IPD checklist*

In the following we present the PRISMA-IPD checklist for the reporting of meta-analysis with IPD (Stewart et al., 2015).

1. Title: Indirect comparison of once-weekly semaglutide 2.4 mg and once-daily liraglutude 3.0 mg in patients with type 2 diabetes from STEP-2 weight management trial.
2. Structured summary: Not applicable.
3. Rationale: The effect of weight loss from semaglutide and liraglutide has never been compared head-to-head on the obese population with type-2 diabetes. The only direct evidence available from an RCT is from the STEP-8 trial (Rubino et al., 2022) on a nondiabetic population with a relatively small sample size.

4. Objectives: The average treatment effect on change from baseline (week 0) to week 44 in body weight (%), in the study population of the STEP-2 trial, comparing once-weekly semaglutide 2.4 mg and once-daily liraglutude 3.0 mg.
5. Protocol and registration: No applicable.
6. Eligibility criteria: All subjects from the SCALE and STEP-2 trials who were randomized are considered eligible, excluding those with no post-baseline body weight measurement and those with missing measurements in the baseline variables specified in point 11.
7. Identifying studies-information sources: Not applicable.
8. Identifying studies-search: Not applicable.
9. Study selection processes: The inclusion-exclusion criteria used in the SCALE and STEP-2 trials are virturally identical.
10. Data collection processes: The IPD were retrieved from the internal trial database in Novo Nordisk A/S by agreement.
11. Data items: The baseline variables included in the indirect comparison, except the randomized treatment, are one of the three groups: the hypothesized treatment effect modifiers of liraglutide 3.0 mg versus placebo, the adjustment proxies and the reweighting proxies. The definitions of the proxies can be found in §3 of the manuscript, and the rationale for selecting the proxies is explicated in §5.2. The collected baseline variables from the STEP-2 trial include: randomized treatment (categorical, liraglutide 3.0 mg, liraglutide 1.8 mg and placebo), body weight (kg, continuous), body-mass index ($\mathrm{kg}\cdot\mathrm{m}^{-2}$, continuous), smoking status (categorical, current smoker, previous smoker, never smoked), duration of diabetes (years, continuous), waist circumference (cm, continuous), age (years, continuous), sex (binary), race (categorical, white, black and others), region of the clinic (categorical, Europe, North America and others), hemoglobin A1c ($\mathrm{mmol}\cdot\mathrm{mol}^{-1}$, continuous), fasting plasma glucose ($\mathrm{mmol}\cdot\mathrm{L}^{-1}$, continuous) and fasting serum insulin ($\mathrm{pmol}\cdot\mathrm{L}^{-1}$, continuous). The baseline variables from the SCALE trial are those from the STEP-2 trial plus serum high-density lipoprotein ($\mathrm{mmol}\cdot\mathrm{L}^{-1}$, continuous), serum very low-density lipoprotein ($\mathrm{mmol}\cdot\mathrm{L}^{-1}$, continuous) and serum triglycerides ($\mathrm{mmol}\cdot\mathrm{L}^{-1}$, continuous), where the randomized treatment is one of semaglutide 2.4 mg, semaglutide 1.0 mg and placebo. The outcome from both trials is the change from baseline to week 44 in body weight. Imputation for missing measurements of week 44 body weight was performed by last observation carried forward (LOCF).

A1. IPD integrity: See Davies et al. (2015) and Davies et al. (2021). No additional checking was performed.

12. Risk of bias assessment in individual studies: Subjects' deviation from protocol was observed in both RCTs, resulting in missing body weight measurements at week 44. LOCF imputation is likely to yield conservative effects, although bias in the other direction is also possible.
13. Specification of outcomes and effect measures: The estimand is the change from baseline to week 44 body weight in the study population of STEP-2, comparing once-weekly semaglutide, 2.4 mg and once-daily liraglutide, 3.0 mg. The effect measure is the difference in changes in body weight.
14. Synthesis methods: Refer to §5.2 and earlier expositions within this section for details.

A2. Exploration of variation in effects: No subgroup analysis was performed.

15. Risk of bias across studies: The proximal indirect comparison estimator is susceptible to bias from the invalidity of proxies (Assumption 2) or the nonexistence of bridge functions. The target population ATE estimator in Dahabreh et al. (2020) may be biased if there are unobserved, shifted effect modifiers. Both estimators are susceptible to violation of positivity or overlap of the study populations of SCALE and STEP-2.
16. Additional analyses: No additional analysis was performed.
17. Study selection and IPD obtained: The studies SCALE and STEP-2 were selected specifically for the comparison of the two GLP-1 receptor agonists for weight management

among overweight patients with type II diabetes. The IPD were available from Novo Nordisk.

18. Study characteristics: See Davies et al. (2015) and Davies et al. (2021).

A3. IPD integrity: See Davies et al. (2015) and Davies et al. (2021). No issue was identfied.

19. Risk of bias within studies: No bias assessment was conducted.
20. Results of individual studies: See Table S3.
21. Results of syntheses: See Table 3.
22. Risk of bias across studies: No bias assessment was conducted.
23. Additional analyses: No additional analysis was performed.
24. Summary of evidence: See §5.2.
25. Strengths and limitations: Not evaluated.
26. Conclusions: See §5.2.

A4. Implications: Not evaluated.

27. Funding: See conflict of interest after the main text. The IPD were supplied by Novo Nordisk A/S.

S6. Proofs

S6.1. *Proof of Proposition 1*

The conditional average treatment effect is

$$
\begin{aligned}
&E\{Y(1) - Y(0) \mid X, U, S = 0\} \\
&\quad = E\{Y(1) - Y(0) \mid X, U, S = 1\} && \text{[Assumption 1(iii)]} \\
&\quad = E\{Y(1) \mid A = 1, X, U, S = 1\} - E\{Y(0) \mid A = 0, X, U, S = 1\} \\
&\qquad && \text{[Assumption 1(ii) and 1(iv)]} \\
&\quad = E(Y \mid A = 1, X, U, S = 1) - E(Y \mid A = 0, X, U, S = 1). && \text{[Assumption 1(i)]}
\end{aligned}
$$

Proceeding from the equation above, it is immediate that the target parameter is

$$
\begin{aligned}
\theta &= E\{Y(1) - Y(0) \mid S = 0\} \\
&= E[E\{Y(1) - Y(0) \mid X, U, S = 0\} \mid S = 0] \\
&= E\{E(Y \mid A = 1, X, U, S = 1) - E(Y \mid A = 0, X, U, S = 1) \mid S = 0\} \\
&= E\left[E\left\{\frac{(2A-1)Y}{\text{pr}(A \mid X, U, S = 1)} \,\middle|\, X, U, S = 1\right\} \,\middle|\, S = 0\right] \\
&= E\left[E\left\{\frac{(2A-1)Y}{e(A \mid X)} \,\middle|\, X, U, S = 1\right\} \,\middle|\, S = 0\right] && \text{[Assumption 1(ii)]} \\
&= E\{E(\tilde{Y} \mid X, U, S = 1) \mid S = 0\}.
\end{aligned}
$$

Then we show the inverse probability weighting representation via the identification formula above. Starting from the g-formula representation, we write

$$
\begin{aligned}
&E\{E(\tilde{Y} \mid X, U, S = 1) \mid S = 0\} \\
&\quad = E\left[E\left\{\frac{S}{\text{pr}(S = 1 \mid X, U)}\tilde{Y} \,\middle|\, X, U\right\} \,\middle|\, S = 0\right] \\
&\quad = \iint E\left\{\frac{S}{\text{pr}(S = 1 \mid X, U)}\tilde{Y} \,\middle|\, X = x, U = u\right\} p(u, x \mid S = 0)\mathrm{d}u\mathrm{d}x,
\end{aligned}
$$

and by a simple manipulation of probability densities, showing that $p(u, x \mid S = 0) = \text{pr}(S = 0 \mid X = x, U = u)p(x, u)/\text{pr}(S = 0)$,

$$= \iint \frac{1}{\alpha} E\left\{ S \frac{\text{pr}(S = 0 \mid X, U)}{\text{pr}(S = 1 \mid X, U)} \tilde{Y} \,\middle|\, X = x, U = u \right\} p(u, x) \mathrm{d}u \mathrm{d}x$$
$$= \frac{1}{\alpha} E\left\{ S \frac{\text{pr}(S = 0 \mid X, U)}{\text{pr}(S = 1 \mid X, U)} \tilde{Y} \right\}.$$

S6.2. *Proof of Lemma 1*

By Assumption 2(ii), 2(iii) and 2(i) for any $h^{\text{U}} \in \mathcal{H}^{\text{U}}$,

$$E(\tilde{Y} \mid Z, X, U, S = 1) = E\{h^{\text{U}}(W, X) \mid Z, X, U, S = 1\}.$$

The first result of the lemma is immediate after integrating both sides of the equation with respect to the conditional density of U given $(Z, X, S = 1)$.

By Assumption 2(ii), for any $q^{\text{U}} \in \mathcal{Q}^{\text{U}}$,

$$E\{q^{\text{U}}(Z, X) \mid W, X, U, S = 1\} = E\{q^{\text{U}}(Z, X) \mid X, U, S = 1\} = \frac{\text{pr}(S = 0 \mid X, U)}{\text{pr}(S = 1 \mid X, U)}.$$

The second result of the lemma follows by integrating both sides with respect to the density $p(u \mid W, X, S = 1)$,

$$\begin{aligned}
&E\{q^{\text{U}}(Z, X) \mid X, W, S = 1\} \\
&\quad = \int \frac{\text{pr}(S = 0 \mid X, u)}{\text{pr}(S = 1 \mid X, u)} p(u \mid W, X, S = 1) \mathrm{d}u \\
&\quad = \int \frac{p(u \mid S = 0, X)\text{pr}(S = 0 \mid X)}{p(u \mid S = 1, X)\text{pr}(S = 1 \mid X)} \frac{p(W \mid u, X, S = 1)p(u \mid X, S = 1)}{p(W \mid X, S = 1)} \mathrm{d}u \\
&\quad = \frac{\text{pr}(S = 0 \mid X)}{\text{pr}(S = 1 \mid X)p(W \mid S = 1, X)} \int p(W \mid u, X, S = 1)p(u \mid S = 0, X) \mathrm{d}u \\
&\quad = \frac{\text{pr}(S = 0 \mid X)}{\text{pr}(S = 1 \mid X)p(W \mid S = 1, X)} \int p(W \mid u, X, S = 0)p(u \mid S = 0, X) \mathrm{d}u \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[Assumption 2(iv)]} \\
&\quad = \frac{\text{pr}(S = 0 \mid X)p(W \mid S = 0, X)}{\text{pr}(S = 1 \mid X)p(W \mid S = 1, X)} \\
&\quad = \frac{\text{pr}(S = 0 \mid W, X)}{\text{pr}(S = 1 \mid W, X)}.
\end{aligned}$$

S6.3. *Proof of Proposition 2*

Consider the case where $\mathcal{H}^{\text{U}} \neq \emptyset$. The g-formula identification given any $h^{\text{U}} \in \mathcal{H}^{\text{U}}$ is

$$\begin{aligned}
\theta &= E\{E(\tilde{Y} \mid X, U, S = 1) \mid S = 0\} && \text{(Proposition 1)} \\
&= E[E\{h^{\text{U}}(W, X) \mid X, U, S = 1\} \mid S = 0] \\
&= E[E\{h^{\text{U}}(W, X) \mid X, U, S = 0\} \mid S = 0] && \text{[Assumption 2(iv)]} \\
&= E\{h^{\text{U}}(W, X) \mid S = 0\}, \\
&= \frac{1}{\alpha} E\left\{ S \frac{\text{pr}(S = 0 \mid W, X)}{\text{pr}(S = 1 \mid W, X)} h^{\text{U}}(W, X) \right\}
\end{aligned}$$

and for any $q \in \mathcal{Q} \neq \emptyset$, we can write the parameter as

$$= \frac{1}{\alpha} E[SE\{q(Z, X) \mid W, X, S = 1\} h^{\text{U}}(W, X)]$$

$$= \frac{1}{\alpha} E\{Sq(Z,X)h^{\mathrm{U}}(W,X)\}$$
$$= \frac{1}{\alpha} E[Sq(Z,X)E\{h^{\mathrm{U}}(W,X) \mid Z,X,S=1\}]$$

which by Lemma 1 is

$$= \frac{1}{\alpha} E[Sq(Z,X)E\{h(W,X) \mid Z,X,S=1\}]$$
$$= E\{h(W,X) \mid S=0\},$$

for any $h \in \mathcal{H}$.

Now consider the case where $\mathcal{Q}^{\mathrm{U}} \neq \emptyset$. The inverse-probability identification given any $q^{\mathrm{U}} \in \mathcal{Q}^{\mathrm{U}}$ is

$$\theta = \frac{1}{\alpha} E\left\{ \frac{S\mathrm{pr}(S=0 \mid X,U)}{\mathrm{pr}(S=1 \mid X,U)} \tilde{Y} \right\} \qquad \text{(Proposition 1)}$$
$$= \frac{1}{\alpha} E[S\tilde{Y}E\{q^{\mathrm{U}}(Z,X) \mid X,U,S=1\}]$$
$$= \frac{1}{\alpha} E[S\tilde{Y}E\{q^{\mathrm{U}}(Z,X) \mid Y,A,X,U,S=1\}] \qquad \text{[Assumptions 2(iii) and 2(i)]}$$
$$= \frac{1}{\alpha} E\{Sq^{\mathrm{U}}(Z,X)\tilde{Y}\},$$
$$= \frac{1}{\alpha} E\{Sq^{\mathrm{U}}(Z,X)E(\tilde{Y} \mid Z,X,S=1)\},$$

and for any $h \in \mathcal{H} \neq \emptyset$, we can write the parameter as

$$= \frac{1}{\alpha} E[Sq^{\mathrm{U}}(Z,X)E\{h(W,X) \mid Z,X,S=1\}]$$
$$= \frac{1}{\alpha} E\{Sq^{\mathrm{U}}(Z,X)h(W,X)\}$$
$$= \frac{1}{\alpha} E[SE\{q^{\mathrm{U}}(Z,X) \mid W,X,S=1\}h(W,X)],$$

which by Lemma 1 is

$$= \frac{1}{\alpha} E[SE\{q(Z,X) \mid W,X,S=1\}h(W,X)]$$
$$= \frac{1}{\alpha} E\{Sq(Z,X)\tilde{Y}\},$$

for any $q \in \mathcal{Q}$.

Therefore, the parameter is identified when either (i) $\mathcal{H}^{\mathrm{U}}$ (hence $\mathcal{H}$) and $\mathcal{Q}$ are nonempty or (ii) $\mathcal{Q}^{\mathrm{U}}$ (hence $\mathcal{Q}$) and $\mathcal{H}$ are nonempty. This is equivalent to the statement in the proposition.

S6.4. *Proof of Proposition 3*

We state a useful theorem in functional analysis. The following is an adapted version of Theorem 1.3.1 from Kesavan (2022).

Theorem S1 (Implicit function). *Let $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}$ be Banach spaces and let $\Omega \in \mathcal{B}_1 \times \mathcal{B}_2$ be an open subset. Let $G : \Omega \to \mathcal{B}$ be a mapping such that:*

(i) G is continuous on Ω;
(ii) For every $(b_1, b_2) \in \Omega$, $(\partial/\partial b_2)G(b_1, b_2)$ exists and is continuous on Ω;
(iii) $G(c_1, c_2) = 0$, $(\partial/\partial b_2)G(b_1, b_2)|_{b_1=c_1, b_2=c_2}$ is bijective.

Then there exists an open neighborhood $\Omega_1 \times \Omega_2 \subset \Omega$ of c_1, c_2 such that for each $b_1 \in \Omega_1$, there exists a unique, continuous function $\rho : \Omega_1 \to \Omega_2$ satisfying $G\{b_1, \rho(b_1)\} = 0$. Moreover, if G is differentiable at (c_1, c_2), then ρ is differentiable at c_1 with derivative

$$\frac{\mathrm{d}}{\mathrm{d}b_1}\rho(b_1)\bigg|_{b_1=c_1} = -\left\{\frac{\partial}{\partial b_2}G(b_1,b_2)\bigg|_{b_1=c_1,b_2=c_2}\right\}^{-1}\frac{\partial}{\partial b_1}G(b_1,b_2)\bigg|_{b_1=c_1,b_2=c_2}$$

Without loss of generality, assume P_0 has a density p_0 with respect to a dominating measure ν. The density p_0 can be factorized as

$$p_0(o) = \left\{p_0(y \mid a,z,w,x,S=1)e_0(a \mid x)p_0(w \mid z,x,S=1)p_0(z \mid x,S=1)\right\}^s \\ p_0(w \mid x,S=0)^{(1-s)}p_0(x,s),$$

due to the fact that $(Z,W) \perp\!\!\!\perp_{P_0} A \mid (X,S=1)$. Let $L_2^0(P_0)$ denote the Hilbert space of $L_2(P_0)$ functions with zero mean under P_0. The tangent space $\dot{\mathcal{P}}_0$ at $P_0 \in \mathcal{P}$ is the subset of $L_2^0(P_0)$ which is the linear closure of the score functions of parametric submodels in $\mathcal{P}$ that contain P_0. We claim that the tangent space is $\dot{\mathcal{P}}_0 = L_2^0(P_0) \setminus \Lambda$, where $\Lambda = \{sc(z,w,x)\{a - e_0(1 \mid x)\} : c(z,w,x) \in L_2(P_0^1)\}$. Note that $\dot{\mathcal{P}}_0$ is maximal.

In order to verify the claim, we find a dense subset of $\dot{\mathcal{P}}_0$ by constructing appropriate parametric submodels. Let $\kappa(x) = 2\{1+\exp(-2x)\}^{-1}$, so that $\kappa(0) = 1$, and $(\mathrm{d}\kappa/\mathrm{d}x)(0) = 1$. Consider the probability measure P_ε with density $p_\varepsilon(o)$ with respect to ν factorizable as

$$p_\varepsilon(o) = \{p_\varepsilon(y \mid a,z,w,x,S=1)e_0(a \mid x)p_\varepsilon(w \mid z,x,S=1)p_\varepsilon(z \mid x,S=1)\}^s \\ p_\varepsilon(w \mid x,S=0)^{(1-s)}p_\varepsilon(x,s),$$

where

$$p_\varepsilon(y \mid a,z,w,x,S=1) = \frac{p_0(y \mid a,z,w,x,S=1)\kappa\{\varepsilon g(y,a,z,w,x)\}}{\int p_0(y \mid a,z,w,x,S=1)\kappa\{\varepsilon g(y,a,z,w,x)\}\mathrm{d}\nu(y)},$$
$$p_\varepsilon(w \mid z,x,S=1) = \frac{p_0(w \mid z,x,S=1)\kappa\{\varepsilon g(w,z,x)\}}{\int p_0(w \mid z,x,S=1)\kappa\{\varepsilon g(w,z,x)\}\mathrm{d}\nu(w)},$$
$$p_\varepsilon(z \mid x,S=1) = \frac{p_0(z \mid x,S=1)\kappa\{\varepsilon g(z,x)\}}{\int p_0(z \mid x,S=1)\kappa\{\varepsilon g(z,x)\}\mathrm{d}\nu(z)},$$
$$p_\varepsilon(w \mid x,S=0) = \frac{p_0(w \mid x,S=0)\kappa\{\varepsilon g(w,x)\}}{\int p_0(w \mid x,S=0)\kappa\{\varepsilon g(w,x)\mathrm{d}\nu(w)},$$
$$p_\varepsilon(x,s) = \frac{p_0(x,s)\kappa\{\varepsilon g(x,s)\}}{\int p_0(x,s)\kappa\{\varepsilon g(x,s)\}\mathrm{d}\nu(x,s)},$$

that

$$E_{P_0}\{g(Y,Z,W,A,X) \mid Z,W,A,X,S=1\} = 0,$$
$$E_{P_0}\{g(W,Z,X) \mid Z,X,S=1\} = 0,$$
$$E_{P_0}\{g(Z,X) \mid X,S=1\} = 0,$$
$$E_{P_0}\{g(W,X) \mid X,S=0\} = 0,$$
$$E_{P_0}\{g(X,S)\} = 0,$$

and that $sg(y,z,w,a,x)$, $sg(w,z,x)$, $sg(z,x)$, $(1-s)g(w,x)$ and $g(x,s)$ are $L_2^0(P_0)$-functions. By construction, $P_\varepsilon|_{\varepsilon=0} = P_0$. We can find an open neighborhood Γ around zero such that $\{P_\varepsilon : \varepsilon \in \Gamma\}$ is a one-dimensional curve parametrized by ε. The propensity score $e_0(a \mid x)$ is left out of the parameterization, since it is assumed to be known.

We now show how $\{P_\varepsilon : \varepsilon \in \Gamma\}$ can be made into a regular parametric submodel. We invoke Theorem S1. Let $\mathcal{B}_1 = \Gamma$, $\mathcal{B}_2 = L_2(W, X; P_0^1)$, $\mathcal{B} = L_2(Z, X; P_0^1)$, $c_1 = 0$, and $c_2 = h_0$. Then the mapping

$$G(\varepsilon, h) = E_{P_\varepsilon}\{\tilde{Y}_0 - h(W, X) \mid Z = z, X = x, S = 1\}$$

fulfills the conditions in Theorem S1, with $G(0, h_0) = 0$ from the definition of h_0, and the derivative is $(\partial/\partial h)G(\varepsilon, h) = -T_\varepsilon$, where $(T_\varepsilon b_2)(z, x) = E_{P_\varepsilon}\{b_2(W, X) \mid Z = z, X = x, S = 1\}$, and $(\partial/\partial h)G(\varepsilon, h)|_{\varepsilon=0, h=h_0} = -T_\varepsilon|_{\varepsilon=0} = -T_0$, which is bijective by Assumption 4(ii). It follows from Theorem S1 that there exists a unique, continuous function $h_\varepsilon(w, x)$ on an open subset $\tilde{\Gamma} \subset \Gamma$ such that

$$E_{P_\varepsilon}[\{\tilde{Y}_0 - h_\varepsilon(W, X)\} \mid Z = z, X = x, S = 1] = 0.$$

Therefore, $\mathcal{H}_\varepsilon = \{h_\varepsilon\}$ is nonempty, which shows that $\{P_\varepsilon : \varepsilon \in \tilde{\Gamma}\}$ is a submodel in $\mathcal{P}$.

Furthermore, since

$$\frac{\partial}{\partial \varepsilon} G(\varepsilon, h)\bigg|_{\varepsilon=0, h=h_0} = E_{P_0}[\{\tilde{Y}_0 - h_0(W, X)\} \{g(Y, Z, W, A, X) + g(W, Z, X)\} \mid Z = z, X = x, S = 1]$$

exists and that its range is contained in $L_2(Z, X; P_0^1)$ by Assumption 4(iii), the function h_ε is differentiable at $\varepsilon = 0$ with derivative such that

$$\frac{\partial}{\partial \varepsilon} h_\varepsilon(w, x) = \big(T_0^{-1} E_{P_0}[\{\tilde{Y}_0 - h_0(W, X)\} \{g(Y, Z, W, A, X) + g(W, Z, X)\} \mid Z = z, X = x, S = 1]\big)(w, x) \quad \text{(S2)}$$

The submodel $P_\varepsilon(g)$ constructed in this fashion has score

$$g(o) = s\{g(y, z, w, a, x) + g(w, z, x) + g(z, x)\} + (1 - s)g(w, x) + g(x, s),$$

where the dependence on g is written out. The union of the tangent sets of the submodels $\{P_\varepsilon(g) : \varepsilon \in \tilde{\Gamma}\}$ obtained by varying g is

$$\begin{aligned}\dot{\mathcal{P}}_0 = \big\{ & g(o) = s\{g(y, z, w, a, x) + g(w, z, x) + g(z, x)\} + (1 - s)g(w, x) + g(x, s) : \\ & E_{P_0}\{g(Y, Z, W, A, X) \mid Z, W, A, X, S = 1\} = 0, E\{g(W, Z, X) \mid Z, X, S = 1\} = 0, \\ & E_{P_0}\{g(Z, X) \mid X, S = 1\} = 0, E_{P_0}\{g(W, X) \mid X, S = 0\} = 0, E_{P_0}\{g(X, S)\} = 0, \\ & sg(y, z, w, a, x), sg(w, z, x), sg(z, x), (1 - s)g(w, x), g(x, s) \in L_2^0(P_0)\big\}.\end{aligned}$$

Any function in $L_2^0(P_0) \setminus \Lambda$ can be orthogonalized by successive projections, and each of the spaces is a closed subspace of $L_2^0(P_0)$ corresponding to the individual functions comprising $g(o)$. Therefore, $\dot{\mathcal{P}}_0 = L_2^0(P_0) \setminus \Lambda$, and since this tangent set is maximal, it must be the tangent space of the model $\mathcal{P}$ at P_0.

The target parameter is $\theta_0 = E_{P_0}\{h_0(W, X) \mid S = 0\}$. The Gateaux derivative of $\theta_\varepsilon = E_{P_\varepsilon}\{h_\varepsilon(W, X) \mid S = 0\}$ at $\varepsilon = 0$ along the submodel $\{P_\varepsilon : \varepsilon \in \tilde{\Gamma}\}$ is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_\varepsilon\bigg|_{\varepsilon=0} &= \frac{\partial}{\partial \varepsilon} \iint h_\varepsilon(w, x) p_\varepsilon(w, x \mid S = 0) \mathrm{d}\nu(w, x)\bigg|_{\varepsilon=0} \\ &= E_{P_0}\left\{ \frac{\partial}{\partial \varepsilon} h_\varepsilon(W, X)\bigg|_{\varepsilon=0} \,\middle|\, S = 0 \right\} + \iint h_0(w, x) \frac{\partial}{\partial \varepsilon} p_\varepsilon(w, x \mid S = 0)\bigg|_{\varepsilon=0} \mathrm{d}\nu(w, x).\end{aligned} \quad \text{(S3)}$$

We now study the two terms separately. The first term in (S3) is

$$\frac{1}{\alpha_0} E_{P_0}\left\{ S \frac{P_0(S = 0 \mid W, X)}{P_0(S = 1 \mid W, X)} \frac{\partial}{\partial \varepsilon} h_\varepsilon(W, X)\bigg|_{\varepsilon=0} \right\}$$

which, after substituting the identification equation of q_0, is

$$
\begin{aligned}
&= \frac{1}{\alpha_0} E_{P_0}\left[S E_{P_0}\{q_0(Z,X) \mid W,X,S=1\} \frac{\partial}{\partial \varepsilon} h_\varepsilon(W,X)\Big|_{\varepsilon=0} \right] \\
&= \frac{1}{\alpha_0} E_{P_0}\left[S q_0(Z,X) E_{P_0}\left\{ \frac{\partial}{\partial \varepsilon} h_\varepsilon(W,X)\Big|_{\varepsilon=0} \,\middle|\, Z,X,S=1 \right\} \right].
\end{aligned}
$$

Inserting the conditional expectation of the derivative from (S2), we develop the term further as

$$
\begin{aligned}
&= \frac{1}{\alpha_0} E_{P_0}\big(S q_0(Z,X) E_{P_0}[\{\tilde{Y}_0 - h_0(W,X)\} \\
&\qquad\qquad \{g(Y,Z,W,A,X) + g(W,Z,X)\} \mid Z,X,S=1] \big) \\
&= \frac{1}{\alpha_0} E_{P_0}\Big(S q_0(Z,X)\{\tilde{Y}_0 - h_0(W,X)\} \\
&\qquad\qquad [S\{g(Y,Z,W,A,X) + g(W,Z,X) + g(Z,X)\} + (1-S)g(W,X) + g(X,S)] \Big) \\
&= E_{P_0}\left[\frac{S}{\alpha_0} q_0(Z,X)\{\tilde{Y}_0 - h_0(W,X)\} g(O) \right].
\end{aligned}
$$

In the second to last step we added the scores $Sg(Z,X)$ and $g(X,S)$, which is valid because their products with the leading factor all have zero mean due to the identification equation of h_0. The score $(1-S)g(W,X)$ was added, which is allowed, since the factor S renders their product zero.

The second term in display (S3) is

$$
\begin{aligned}
&\frac{1}{\alpha_0} \iint \sum_{s\in\{0,1\}} (1-s) h_0(w,x) \frac{\partial}{\partial \varepsilon} \{p_\varepsilon(w \mid x, S=0) p_\varepsilon(s,x)\}\Big|_{\varepsilon=0} \mathrm{d}\nu(w,x) \\
&\quad - \frac{1}{\alpha_0^2} \iint \sum_{s\in\{0,1\}} (1-s) \frac{\partial}{\partial \varepsilon} \{p_\varepsilon(w \mid x, S=0) p_\varepsilon(s,x)\}\Big|_{\varepsilon=0} \mathrm{d}\nu(w,x) E_{P_0}\{(1-S)h_0(W,X)\} \\
&\quad = \frac{1}{\alpha_0} \iint \sum_{s\in\{0,1\}} (1-s) h_0(w,x)\{g(w \mid x) + g(s,x)\} p_0(w \mid x, S=0) p_0(x,s) \mathrm{d}\nu(w,x) \\
&\qquad - \frac{1}{\alpha_0} \iint \sum_{s\in\{0,1\}} (1-s)\theta_0\{g(w \mid x) + g(s,x)\} p_0(w \mid x, S=0) p_0(x,s) \mathrm{d}\nu(w,x) \\
&\quad = E_{P_0}\left[\frac{1-S}{\alpha_0} \{h_0(W,X) - \theta_0\}\{(1-S)g(W,X) + g(X,S)\} \right] \\
&\quad = E_{P_0}\Big[\frac{1-S}{\alpha_0} \{h_0(W,X) - \theta_0\} \\
&\qquad\qquad [S\{g(Y,Z,W,A,X) + g(W,Z,X) + g(Z,X)\} + (1-S)g(W,X) + g(X,S)] \Big] \\
&\quad = E_{P_0}\left[\frac{1-S}{\alpha_0} \{h_0(W,X) - \theta_0\} g(O) \right].
\end{aligned}
$$

The scores $Sg(Y \mid Z,W,A,X)$, $Sg(W \mid Z,X)$ and $Sg(Z \mid X)$ were added in the second to last step, which is allowed because the factor $(1-S)$ renders their products zero. Collecting the two results above, we have that

$$
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_\varepsilon\Big|_{\varepsilon=0} = E_{P_0}\left(\left[\frac{S}{\alpha_0} q_0(Z,X)\{\tilde{Y}_0 - h_0(W,X)\} + \frac{1-S}{\alpha_0}\{h_0(W,X) - \theta_0\} \right] g(O) \right),
$$

which shows that $\phi_0(o)$, the factor next to $g(o)$, is an influence function of the parameter θ_0 at $P_0 \in \mathcal{P}$.

Next, we will show the efficient influence function of θ_0. Define the function

$$\varphi_0(o) = \phi_0(o) - \frac{s}{\alpha_0} q_0(z,x) E_{P_0}\left[\frac{Y}{\{e_0(A \mid X)\}^2} \;\middle|\; Z = z, W = w, X = x, S = 1 \right] \{a - e_0(1 \mid x)\}.$$

It is an influence function of θ_0, because the term after $\phi_0(o)$ is an element of Λ, and thus $(\mathrm{d}/\mathrm{d}\varepsilon)\theta_\varepsilon|_{\varepsilon=0} = E_{P_0}\{\varphi_0(O)g(O)\}$ for all $g \in \dot{\mathcal{P}}_0$. The function expands as

$$\begin{aligned}
\varphi_0(o) = {} & \frac{s}{\alpha_0} q_0(z,x) \frac{(2a-1)}{e_0(a \mid x)} \{y - E_{P_0}(Y \mid Z = z, W = w, A = a, X = x, S = 1)\} \\
& - \frac{s}{\alpha_0} q_0(z,x) h_0(w,x) + \frac{1-s}{\alpha_0} \{h_0(w,x) - \theta_0\} \\
& + \frac{s}{\alpha_0} q_0(z,x) \frac{(2a-1)}{e_0(a \mid x)} E_{P_0}(Y \mid Z = z, W = w, A = a, X = x, S = 1) \\
& - \frac{s}{\alpha_0} q_0(z,x) \sum_{a' \in \{0,1\}} \frac{E_{P_0}(Y \mid Z = z, W = w, A = a', X = x, S = 1)}{e_0(a' \mid x)} \{a - e_0(1 \mid x)\} \\
= {} & \frac{s}{\alpha_0} q_0(z,x) \frac{(2a-1)}{e_0(a \mid x)} \{y - E_{P_0}(Y \mid Z = z, W = w, A = a, X = x, S = 1)\} \\
& - \frac{s}{\alpha_0} q_0(z,x) h_0(w,x) + \frac{1-s}{\alpha_0} \{h_0(w,x) - \theta_0\} \\
& + \frac{s}{\alpha_0} q_0(z,x) E_{P_0}(\tilde{Y}_0 \mid Z = z, W = w, X = x, S = 1).
\end{aligned}$$

To conclude the proof, we check that the function $\varphi_0(o)$ is indeed an element of $\dot{\mathcal{P}}_0$. Consider the decomposition that

$$\varphi_0(o) = s g^*(y,z,w,a,x) + s g^*(w,z,x) + (1-s) g^*(w,x) + g^*(x,s),$$

where

$$\begin{aligned}
g^*(y,z,w,a,x) &= \frac{1}{\alpha_0} q_0(z,x) \frac{(2a-1)}{e_0(a \mid x)} \{y - E_{P_0}(Y \mid Z = z, W = w, A = a, X = x, S = 1)\}, \\
g^*(w,z,x) &= \frac{1}{\alpha_0} q_0(z,x) \{E_{P_0}(\tilde{Y}_0 \mid Z = z, W = w, X = x, S = 1) - h_0(w,x)\}, \\
g^*(w,x) &= \frac{1}{\alpha_0} [h_0(w,x) - E_{P_0}\{h_0(w,x) \mid X = x, S = 0\}], \\
g^*(x,s) &= \frac{1-s}{\alpha_0} [E_{P_0}\{h_0(W,X) \mid X = x, S = 0\} - \theta_0].
\end{aligned}$$

We need to check that $E_{P_0}\{g^*(Y,Z,W,A,X) \mid Z,W,A,X,S = 1\} = 0$, $E_{P_0}\{g^*(W,Z,X) \mid Z,X,S = 1\} = 0$, $E_{P_0}\{g^*(W,X) \mid X,S = 0\} = 0$, as well as $E_{P_0}\{g^*(X,S)\} = 0$, all of which hold true by the definition of h_0 and θ_0. Therefore, the influence function $\varphi_0(o) \in \dot{\mathcal{P}}_0$ is the efficient influence function of θ_0 at $P_0 \in \mathcal{P}$.

S6.5. *Proof of Theorem 1*

Let

$$\begin{aligned}
\ell_0(o) &= \frac{s}{\alpha_0} q_0 \{\tilde{y}_0 - h_0\} + \frac{1-s}{\alpha_0} h_0, \\
\hat{\ell}(o) &= \frac{s}{\hat{\alpha}} \hat{q} \{\tilde{y}_0 - \hat{h}\} + \frac{1-s}{\hat{\alpha}} \hat{h}.
\end{aligned}$$

Then ℓ_0, $\hat{\ell}$ belong to the P_0-Donsker class $\mathcal{G}_0$, which is also P_0-Glivenko-Cantelli.

We first show consistency. Consider the difference

$$\hat{\theta} - \theta_0 = (P_n - P_0)\hat{\ell} + \left(P_0\hat{\ell} - \frac{\alpha_0}{\hat{\alpha}}\theta_0\right) - \frac{\hat{\alpha} - \alpha_0}{\hat{\alpha}}\theta_0. \tag{S4}$$

The absolute value of the first term of (S4) is bounded by $\sup_{g\in\mathcal{G}_0} |(P_n - P_0)g|$, which converges in probability to zero by the uniform law of large numbers applied to the P_0-Glivenko-Cantelli class $\mathcal{G}_0$. The absolute value of the second term of (S4) is

$$\begin{aligned}
\left|P_0\hat{\ell} - \frac{\alpha_0}{\hat{\alpha}}\theta_0\right| &= \left|\frac{1}{\hat{\alpha}}P_0\{S\hat{q}(\tilde{Y}_0 - \hat{h}) + (1-S)\hat{h}\} - \frac{\alpha_0}{\hat{\alpha}}\theta_0\right| \\
&= \left|\frac{1}{\hat{\alpha}}P_0\{S\hat{q}(h_0 - \hat{h}) + Sq_0\hat{h}\} - \frac{\alpha_0}{\hat{\alpha}}\theta_0\right| \\
&= \left|\frac{1}{\hat{\alpha}}P_0\{S(\hat{q} - q_0)(h_0 - \hat{h}) + Sq_0h_0\} - \frac{\alpha_0}{\hat{\alpha}}\theta_0\right| \\
&= \left|\frac{1}{\hat{\alpha}}P_0\{S(\hat{q} - q_0)(h_0 - \hat{h})\}\right| \\
&\leq M\|\hat{q} - q_0\|_{P_0^1}\|\hat{h} - h_0\|_{P_0^1} = o_{P_0}(1).
\end{aligned} \tag{S5}$$

The fourth step above is due to the observation

$$\begin{aligned}
E_{P_0}\{Sq_0(Z,X)h_0(W,X)\} &= E_{P_0}[Sq_0(Z,X)E_{P_0}\{h_0(W,X) \mid Z, X, S=1\}] \\
&= E_{P_0}\{Sq_0(Z,X)E_{P_0}(\tilde{Y}_0 \mid Z, X, S=1)\} = E_{P_0}\{Sq_0(Z,X)\tilde{Y}_0\} = \alpha_0\theta_0.
\end{aligned}$$

The absolute value of the third term of (S4) converges in probability to zero by the trivial consistency of $\hat{\alpha}$ and Slutsky's theorem. The triangle inequality shows $\hat{\theta} \xrightarrow{\text{P}} \theta_0$.

We now show asymptotic linearity. Working under the additional assumption $\|\hat{q} - q_0\|_{P_0^1}\|\hat{h} - h_0\|_{P_0^1} = o_{P_0}(n^{-1/2})$, we further express the difference (S4) as

$$\begin{aligned}
\hat{\theta} - \theta_0 &= (P_n - P_0)(\hat{\ell} - \ell_0) + P_n\ell_0 - \theta_0 + \left(P_0\hat{\ell} - \frac{\alpha_0}{\hat{\alpha}}\theta_0\right) - \frac{\hat{\alpha} - \alpha_0}{\hat{\alpha}}\theta_0 \\
&= P_n\left(\phi_0 + \frac{1-S}{\alpha_0}\theta_0\right) + (P_n - P_0)(\hat{\ell} - \ell_0) - \frac{2\hat{\alpha} - \alpha_0}{\hat{\alpha}}\theta_0 + \left(P_0\hat{\ell} - \frac{\alpha_0}{\hat{\alpha}}\theta_0\right) \\
&= P_n\phi_0 + (P_n - P_0)(\hat{\ell} - \ell_0) + \frac{(\hat{\alpha} - \alpha_0)^2}{\alpha_0\hat{\alpha}}\theta_0 + o_{P_0}(n^{-1/2}).
\end{aligned} \tag{S6}$$

For the last equality, we use the bound from (S5) but with $\|\hat{q} - q_0\|_{P_0^1}\|\hat{h} - h_0\|_{P_0^1} = o_{P_0}(n^{-1/2})$. The second term of (S6) is an empirical process term of order $o_{P_0}(n^{-1/2})$ if $\|\hat{\ell} - \ell_0\|_{P_0} = o_{P_0}(1)$, since $\mathcal{G}_0$ is P_0-Donsker. Applying the central limit theorem to $\hat{\alpha}$ and then Slutsky's theorem, the third term of (S6) is $O_{P_0}(n^{-1}) = o_{P_0}(n^{-1/2})$. To conclude the proof, we show that $\|\hat{\ell} - \ell_0\|_{P_0}$ indeed converges in probability to zero under $\|\hat{h} - h_0\|_{P_0^1} = o_{P_0^1}(1)$, $\|\hat{q} - q_0\|_{P_0^1} = o_{P_0^1}(1)$ and the boundedness conditions. We have $\|\hat{q}\|_{P_0^1} = O_{P_0}(1)$ because it is bounded, and $\|\hat{h}\|_{P_0^1} \leq \|\hat{h} - h_0\|_{P_0^1} + \|h_0\|_{P_0^1} = o_{P_0}(1) + O_{P_0}(1) = O_{P_0}(1)$. Furthermore, by the boundedness of $\hat{q}$, the terms $\|\hat{q}(\hat{h} - h_0)\|_{P_0^1}$ and $\|\hat{q}h_0\|_{P_0^1}$ are also bounded in probability. The $L_2(P_0)$-norm of the plug-in function $\hat{\ell}$ is

$$\|\hat{\ell}\|_{P_0} \leq M\left\{\|S\hat{q}\tilde{Y}_0\|_{P_0} + \|S\hat{q}(\hat{h} - h_0)\|_{P_0} + \|S\hat{q}h_0\|_{P_0} + \|(1-S)\hat{h}\|_{P_0}\right\}$$

$$\leq M^3\{E_{P_0}(SY^2)\}^{1/2} + M\left[E_{P_0}\left\{S\frac{P_0(S=0 \mid W,X)}{P_0(S=1 \mid W,X)}\hat{h}^2\right\}\right]^{1/2} + O_{P_0}(1)$$
$$\leq M^{7/2} + M^{3/2}\|\hat{h}\|_{P_0^1} + O_{P_0}(1),$$

which is indeed bounded in probability. The $L_2(P_0)$-distance between the plugin and the true function is

$$\|\hat{\ell} - \ell_0\|_{P_0}$$
$$\leq M\{\|S(\hat{q}-q_0)\tilde{Y}_0\|_P + \|S(\hat{q}\hat{h} - q_0h_0)\|_{P_0} + \|(1-S)(\hat{h}-h_0)\|_{P_0} + |\hat{\alpha}-\alpha_0|\|\hat{\ell}\|_{P_0}\},$$

and by similar arguments above, we bound the distance by

$$\leq M^2\|\hat{q}-q_0\|_{P_0^1} + M\|(\hat{q}-q_0)h_0\|_{P_0^1} + M\|\hat{q}(\hat{h}-h_0)\|_{P_0^1}$$
$$+ M^{1/2}\|\hat{h}-h_0\|_{P_0^1} + M|\hat{\alpha}-\alpha_0|O_{P_0}(1)$$
$$\leq 2M^2\|\hat{q}-q_0\|_{P_0^1} + (M^2+M^{1/2})\|\hat{h}-h_0\|_{P_0^1} + o_{P_0}(1) = o_{P_0}(1).$$

This shows $\hat{\theta} - \theta_0 = P_n\phi_0 + o_{P_0}(n^{-1/2})$.

S7. Handling missing outcome

S7.1. *Identifiability*

In our discussion thus far, we have ignored a common issue in many RCTs: study participants are typically followed over a period of time, during which some may drop out before the end of study, and their outcomes are not recorded. When the dropout mechanism is not missing completely at random, applying the proximal indirect comparison estimator from previous sections to the nonmissing population may not identify the full-compliance ATE in the target RCT θ due to potential selection bias. In this section, we propose an estimator which correctly identifies the target parameter θ under a missing-at-random (MAR) dropout pattern.

The binary missingness indicator Δ takes the value 0 when a study participant's outcome information is missing. Let the conditional probability of no dropout from the source trial be $\pi(Z,W,A,X) = \text{pr}(\Delta = 1 \mid Z,W,A,X,S=1)$. We assume that the missingness in the source trial is noninformative of the outcome, conditioning on all other observed variables, which is formalized below.

Assumption S5 (Missing at random).
(i) $\Delta \perp\!\!\!\perp Y \mid (Z,W,A,X,S=1)$;
(ii) $\pi(Z,W,A,X) > 0$ whenever $\text{pr}(S=1 \mid Z,W,A,X) > 0$.

In particular, Assumption S5 requires that the unobserved effect modifiers U do not directly affect the missing pattern. They are allowed to have an indirect effect through the proxies and the baseline covariates, upon controlling for which the missingness is ignorable. If the outcome is MAR, one can devise augmented estimators from the influence functions of the target parameter θ defined on the data without missingness (Tsiatis, 2006). If missing outcomes are also present in the target RCT, the identifiability of the ATE θ comparing treatments $A=0$ and $A=-1$ relying only on randomization is lost.

The observed data model subject to missingness $\mathcal{P}^{\text{C}}$ is the collection of distributions over $O^{\text{C}} = (S, S\Delta, SA, X, W, SZ, S\Delta Y)$ such that $P^{\text{C}}(O^{\text{C}}, S\Delta = 1) = \pi(Z,W,A,X)P(O, S=1)$ and $P^{\text{C}}(O^{\text{C}}, S=0) = P(O,S=0)$ for all $P \in \mathcal{P}$. In this sense, we can write every $P^{\text{C}} \in \mathcal{P}^{\text{C}}$ as a function $P^{\text{C}}(P)$. The definition of the sets of bridge functions $\mathcal{H}$ and $\mathcal{Q}$ is valid for the model $\mathcal{P}^{\text{C}}$

under MAR, which immediately makes the target parameter identifiable. Note that conditional mean of the outcome among subjects without missingness in the source trial $E(Y \mid \Delta = 1, Z = z, W = w, A = a, X = x, S = 1)$ is the same as the conditional mean $\mu(z, w, a, x) = E(Y \mid Z = z, W = w, A = a, X = x, S = 1)$ where the outcome is always observed.

S7.2. *Estimation*

Proposition S3. *Suppose Assumption S5 holds. For $P_0 \in \mathcal{P}$ under Assumption 4, the efficient influence function of the target parameter θ_0 at $P_0^{\mathrm{C}} \in \mathcal{P}^{\mathrm{C}}$ is*

$$\begin{aligned}\varphi_0^{\mathrm{C}}(o) = \frac{s}{\alpha_0} q_0(z,x) \frac{\delta(2a-1)}{\pi_0(z,w,a,x) e_0(a \mid x)} \{y - \mu_0(z,w,a,x)\} \\ + \frac{s}{\alpha_0} q_0(z,x) \{\mu_0(z,w,1,x) - \mu_0(z,w,0,x) - h_0(w,x)\} + \frac{1-s}{\alpha_0} \{h_0(w,x) - \theta_0\}.\end{aligned}$$

Proof. Recall that in the proof of Proposition 3, we have derived the tangent space $\dot{\mathcal{P}}_0$ and it orthogonal complement Λ. Therefore, the translation $\varphi(o) + \Lambda$ is the space of all influence functions of the parameter θ_0 at P_0 under the model $\mathcal{P}$. More explicitly, the influence functions share the form

$$\tilde{\phi}_0(o; c) = \varphi_0(o) + sc(z,w,x)\{a - e_0(1 \mid x)\},$$

where $c \in L_2(P_0^1)$ is arbitrary. Following Example 25.43 in van der Vaart (1998), all influence functions of θ_0 at P^{C} under the model $\mathcal{P}^{\mathrm{C}}$ can be characterized as

$$\begin{aligned}\tilde{\phi}_0^{\mathrm{C}}(o; c, b) &= \left\{ \frac{s\delta}{\pi_0(z,w,a,x)} + (1-s) \right\} \tilde{\phi}_0(o; c) + sb(z,w,a,x)\{\delta - \pi_0(z,w,a,x)\} \\ &= \left\{ \frac{s\delta}{\pi_0(z,w,a,x)} + (1-s) \right\} \left[\varphi_0(o) + sc(z,w,x)\{a - e_0(1 \mid x)\} \right] \\ &\quad + sb(z,w,a,x)\{\delta - \pi_0(z,w,a,x)\} \\ &= \frac{s}{\alpha_0} \frac{\delta}{\pi_0(z,w,a,x)} q_0(z,x) \frac{2a-1}{e_0(a \mid x)} \{y - \mu_0(z,w,a,x)\} \\ &\quad + \frac{s}{\alpha_0} \frac{\delta}{\pi_0(z,w,a,x)} q_0(z,x) \{\mu_0(z,w,1,x) - \mu_0(z,w,0,x) - h_0(w,x)\} \\ &\quad + \frac{1-s}{\alpha_0} \{h_0(w,x) - \theta_0\} + \frac{s\delta}{\pi_0(z,w,a,x)} c(z,w,x)\{a - e_0(1 \mid x)\} \\ &\quad + sb(z,w,a,x)\{\delta - \pi_0(z,w,a,x)\}.\end{aligned}$$

To find the efficient influence function, we first optimize over b. This is equivalent to calculating the projection of

$$\left\{ \frac{s\delta}{\pi_0(z,w,a,x)} + (1-s) \right\} \tilde{\phi}_0(o; c)$$

onto $(\Lambda^{\mathrm{C}})^{\perp}$, where

$$\Lambda^{\mathrm{C}} = \{sb(z,w,a,x)\{\delta - \pi_0(z,w,a,x)\} : b(z,w,a,x) \in L_2(P_0^1)\}.$$

Suppose the projection has the form $s\{\delta - \pi_0(z,w,a,x)\}b^*(z,w,a,x)$. Then the function b^* satisfies the equation

$$\begin{aligned}E_{P_0}\Bigg(\Bigg[\frac{\Delta \tilde{\phi}_0(O; c)}{\pi_0(Z,W,A,X)} - \{\Delta - \pi_0(Z,W,A,X)\} b^*(Z,W,A,X) \Bigg] \\ \{\Delta - \pi_0(Z,W,A,X)\} \,\Big|\, Z,W,A,X,S=1 \Bigg) = 0.\end{aligned}$$

The solution is

$$b^*(z,w,a,x) = \frac{q_0(z,x)}{\alpha_0\pi_0(z,w,a,x)}\{\mu_0(z,w,1,x) - \mu_0(z,w,0,x) - h_0(w,x)\} + \frac{c(z,w,x)}{\pi_0(z,w,a,x)}\{a - e_0(1 \mid x)\},$$

which gives the projection

$$\begin{aligned}\tilde{\phi}_0^{\mathrm{C}}(o;c,-b^*) = {} & \frac{s}{\alpha_0}q_0(z,x)\frac{2a-1}{e_0(a \mid x)}\frac{\delta}{\pi_0(z,w,a,x)}\{y - \mu_0(z,w,a,x)\} \\ & + \frac{s}{\alpha_0}q_0(z,x)\{\mu_0(z,w,1,x) - \mu_0(z,w,0,x) - h_0(w,x)\} \\ & + \frac{1-s}{\alpha_0}\{h_0(w,x) - \theta_0\} \\ & + s\{a - e_0(1 \mid x)\}c(z,w,x). \end{aligned} \tag{S7}$$

We now optimize over c. Observe that the trailing term (S7) is orthogonal to all other terms of $\tilde{\phi}_0^{\mathrm{C}}(o;c,-b^*)$. The optimal solution is $c^* = 0$. From this we conclude that the efficient influence function $\varphi_0^{\mathrm{C}}(o) = \tilde{\phi}_0^{\mathrm{C}}(o;c^*,-b^*)$ is as stated in Proposition S3. □

In §4, the estimation of the outcome bridge with no missing outcome does not involve any nuisance parameter, in that the propensity score is assumed to be known, and $\tilde{Y}_0 = (2A-1)Y/e_0(A \mid X)$ can be treated as the de facto outcome in the analysis. However, in the presence of missingness on the outcome, we resort to two-stage estimation of the outcome bridge function. In the first stage a regression model is fitted for the outcome on the nonmissing participants, so that $\hat{\mu}$ is an estimator for μ_0. Additionally, we fit a binary regression model $\hat{\pi}$ for the probability of nonmissingness π_0. In the second stage, the outcome bridge is estimated by a minimax optimization problem based on

$$\psi_{h',q',\pi',\mu'}(o) = q'(z,x)\left[\frac{2a-1}{e_0(a \mid x)}\frac{\delta}{\pi'(z,w,a,x)}\{y - \mu'(z,w,a,x)\} + \mu'(z,w,1,x) - \mu'(z,w,0,x) - h'(w,x)\right].$$

The intuition is that $\psi_{h',q',\pi',\mu'}$ can be used to construct a doubly-robust estimating equation, in the sense that for any q', the population mean $E_{P^{\mathrm{C}}}(\psi_{h_0,q',\pi',\mu'} \mid S = 1)$ evaluated at the true outcome bridge h_0 is zero if either $\mu' = \mu_0$ or $\pi' = \pi_0$. The nuisance models are plugged into $\psi_{h',q',\pi',\mu'}$, giving the estimated outcome bridge

$$\hat{h} = \arg\inf_{h' \in \mathcal{H}'}\sup_{q' \in \mathcal{Q}'}\left\{\frac{1}{n_1}\sum_{i:S_i=1}\psi_{h',q',\hat{\pi},\hat{\mu}}(O_i)\right\}^2,$$

where $\mathcal{H}'$ is the bridge hypothesis class and $\mathcal{Q}'$ is the critic class (Kallus et al., 2022). The former is the postulated model for the outcome bridge, and the latter is the adversarial class of functions used to construct the worst-case loss. With the nuisance models, we compose an estimator for the target parameter θ_0 motivated by the efficient influence function φ_0^{C} from Proposition S3:

$$\hat{\theta}^{\mathrm{C}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{S_i}{\hat{\alpha}}\psi_{\hat{h},\hat{q},\hat{\pi},\hat{\mu}}(O_i) + \frac{1-S_i}{\hat{\alpha}}\hat{h}(W_i,X_i)\right\}.$$

In Theorem S2, we show that the estimator $\hat{\theta}^{\mathrm{C}}$ is multiply-robust under some regularity conditions and convergence of the nuisance models. We use $\pi_{0,a}$, $\bar{\pi}_a$, $\hat{\pi}_a$, $\mu_{0,a}$, $\bar{\mu}_a$, and $\hat{\mu}_a$ as shorthand notations for the respective functions fixing the corresponding argument at a.

Assumption S6 (Regularity conditions).

(i) The function class

$$\mathcal{G}_0^{\mathrm{C}} = \left\{ g^{\mathrm{C}}(o) = \frac{s}{\alpha'} q' \frac{\delta(2a-1)}{\pi' e} (y - \mu') + \frac{s}{\alpha'} q' (\mu_1' - \mu_0' - h') + \frac{1-s}{\alpha'} h' : \right.$$
$$\left. \alpha' \in [0,1], q' \in L_2(Z, X; P_0^1), \pi', \mu' \in L_2(Z, W, A, X; P_0^1), h' \in L_2(W, X; P_0^1) \right\}$$

is P_0^{C}-Donsker.

(ii) There exists a universal constant $M > 1$ such that $\alpha_0 \geq M^{-1}$, $\hat{\alpha} \geq M^{-1}$, $e_0 \geq M^{-1}$, $\pi_0 \geq M^{-1}$, $\hat{\pi} \geq M^{-1}$, $|\mu_0| \leq M$, $|h_0| \leq M$, $|\hat{q}| \leq M$, $P_0(S = 1 \mid W, X) \geq M^{-1}$, and $E_{P_0^{\mathrm{C}}}(Y^2 \mid \Delta = 1, Z, W, A, X, S = 1) \leq M$.

Theorem S2. *Suppose Assumption S6 holds and that* $\|\hat{\mu}_a - \bar{\mu}_a\|_{P_0^{\mathrm{C}},1} = o_{P_0^{\mathrm{C}}}(1)$, $\|\hat{\pi}_a - \bar{\pi}_a\|_{P_0^{\mathrm{C}},1} = o_{P_0^{\mathrm{C}}}(1)$, $\|\hat{h} - \bar{h}\|_{P_0^{\mathrm{C}},1} = o_{P_0^{\mathrm{C}}}(1)$, $\|\hat{q} - \bar{q}\|_{P_0^{\mathrm{C}},1} = o_{P_0^{\mathrm{C}}}(1)$ *for some nonrandom functions* $\bar{\mu}_a(z,w,x)$, $\bar{\pi}_a(z,w,x)$, $\bar{h}(w,x)$ *and* $\bar{q}(z,x)$ *in* $L_2(P_0^{\mathrm{C},1})$. *Then:*

1. *The estimator* $\hat{\theta}^{\mathrm{C}}$ *is consistent for* θ_0, *if either (a)* $\bar{h} = h_0$ *and* $\bar{\mu}_a = \mu_{0,a}$, *(b)* $\bar{h} = h_0$ *and* $\bar{\pi}_a = \pi_{0,a}$, *(c)* $\bar{q} = q_0$ *and* $\bar{\pi}_a = \pi_{0,a}$, *or (d)* $\bar{q} = q_0$ *and* $\bar{\mu}_a = \mu_{0,a}$.
2. *The estimator* $\hat{\theta}^{\mathrm{C}}$ *is asymptotically linear with influence function* φ_0^{C}, *if (a)* $\bar{h} = h_0$, $\bar{q} = q_0$, $\bar{\pi}_a = \pi_{0,a}$, $\bar{\mu}_a = \mu_{0,a}$, *and (b)* $\|\hat{q} - q_0\|_{P_0^{\mathrm{C}},1} \|\hat{h} - h_0\|_{P_0^{\mathrm{C}},1} + \sum_{a \in \{0,1\}} \|\hat{\pi}_a - \pi_{0,a}\|_{P_0^{\mathrm{C}},1} \|\hat{\mu}_a - \mu_{0,a}\|_{P_0^{\mathrm{C}},1} = o_{P_0^{\mathrm{C}}}(n^{-1/2})$.

Proof. Define

$$\ell_0^{\mathrm{C}}(o) = \frac{s}{\alpha_0} q_0 \frac{\delta(2a-1)}{\pi_0 e_0} (y - \mu_0) + \frac{s}{\alpha_0} q_0 (\mu_{0,1} - \mu_{0,0} - h_0) + \frac{1-s}{\alpha_0} h_0,$$
$$\hat{\ell}^{\mathrm{C}}(o) = \frac{s}{\hat{\alpha}} \hat{q} \frac{\delta(2a-1)}{\hat{\pi} e_0} (y - \hat{\mu}) + \frac{s}{\hat{\alpha}} \hat{q} (\hat{\mu}_1 - \hat{\mu}_0 - \hat{h}) + \frac{1-s}{\hat{\alpha}} \hat{h}.$$

Both ℓ_0^{C} and $\hat{\ell}^{\mathrm{C}}$ belong to the P^{C}-Donsker class $\mathcal{G}_0^{\mathrm{C}}$, which is also P^{C}-Glivenko-Cantelli.

We first show the consistency of $\hat{\theta}^{\mathrm{C}}$. Consider the difference

$$\hat{\theta}^{\mathrm{C}} - \theta_0 = (P_n^{\mathrm{C}} - P_0^{\mathrm{C}}) \hat{\ell}^{\mathrm{C}} + \left(P^{\mathrm{C}} \hat{\ell}^{\mathrm{C}} - \frac{\alpha_0}{\hat{\alpha}} \theta_0 \right) - \frac{\hat{\alpha} - \alpha_0}{\hat{\alpha}} \theta_0. \tag{S8}$$

The absolute value of the first term of (S8) is bounded by $\sup_{g^{\mathrm{C}} \in \mathcal{G}_0^{\mathrm{C}}} |(P_n^{\mathrm{C}} - P_0^{\mathrm{C}}) g^{\mathrm{C}}| \xrightarrow{\mathrm{P}} 0$, since $\mathcal{G}_0^{\mathrm{C}}$ is P_0^{C}-Glivenko-Cantelli. The absolute value of the second term of (S8) is

$$\begin{aligned}
&\left| P_0^{\mathrm{C}} \hat{\ell}^{\mathrm{C}} - \frac{\alpha_0}{\hat{\alpha}} \theta_0 \right| \\
&\quad = \left| P_0^{\mathrm{C}} \left[\frac{S}{\hat{\alpha}} \hat{q} \frac{2A-1}{e_0} \frac{\Delta}{\hat{\pi}} (Y - \hat{\mu}) + \frac{S}{\hat{\alpha}} \hat{q} (\hat{\mu}_1 - \hat{\mu}_0 - \hat{h}) + \frac{1-S}{\hat{\alpha}} \hat{h} \right] - \frac{\alpha_0}{\hat{\alpha}} \theta_0 \right| \\
&\quad = \left| \frac{1}{\hat{\alpha}} P^{\mathrm{C}} \left[S \hat{q} \left\{ \frac{\pi_{0,1}}{\hat{\pi}_1} (\mu_{0,1} - \hat{\mu}_1) - \frac{\pi_{0,0}}{\hat{\pi}_0} (\mu_{0,0} - \hat{\mu}_0) \right\} + S \hat{q} (\hat{\mu}_1 - \hat{\mu}_0 - \hat{h}) + S q_0 \hat{h} \right] - \frac{\alpha_0}{\hat{\alpha}} \theta_0 \right| \\
&\quad = \left| \frac{1}{\hat{\alpha}} P^{\mathrm{C}} \left\{ S \sum_{a \in \{0,1\}} (-1)^{1-a} \hat{q} \frac{\pi_{0,a} - \hat{\pi}_a}{\hat{\pi}_a} (\mu_{0,a} - \hat{\mu}_a) + S \hat{q} (\mu_{0,1} - \hat{\mu}_1 - \mu_{0,0} + \hat{\mu}_0) \right. \right. \\
&\qquad \left. \left. + S \hat{q} (\hat{\mu}_1 - \hat{\mu}_0 - \hat{h}) + S q_0 \hat{h} \right\} - \frac{\alpha_0}{\hat{\alpha}} \theta_0 \right|
\end{aligned}$$

$$= \left| \frac{1}{\hat{\alpha}} P^{\mathrm{C}} \Big\{ S \sum_{a \in \{0,1\}} (-1)^{1-a} \hat{q} \frac{\pi_{0,a} - \hat{\pi}_a}{\hat{\pi}_a} (\mu_{0,a} - \hat{\mu}_a) \right.$$

$$\left. - S(\hat{q} - q_0)(\hat{h} - h_0) + S h_0 q_0 \Big\} - \frac{\alpha_0}{\hat{\alpha}} \theta_0 \right|$$

$$\leq M^3 \sum_{a \in \{0,1\}} \|\hat{\pi}_a - \pi_{0,a}\|_{P_0^{\mathrm{C}},1} \|\hat{\mu}_a - \mu_{0,a}\|_{P_0^{\mathrm{C}},1} + M \|\hat{h} - h_0\|_{P_0^{\mathrm{C}},1} \|\hat{q} - q_0\|_{P_0^{\mathrm{C}},1} \xrightarrow{\mathrm{P}} 0.$$

The absolute value of the third term of (S8) is trivially $o_{P_0^{\mathrm{C}}}(1)$ due to the consistency of $\hat{\alpha}$ and Slutsky's theorem. Collecting these three terms, the triangular inequality shows $|\hat{\theta}^{\mathrm{C}} - \theta_0| = o_{P^{\mathrm{C}}}(1)$. This shows the first part of the theorem.

We now show the asymptotic linearity of $\hat{\theta}^{\mathrm{C}}$. Working under the additional assumption

$$\sum_{a \in \{0,1\}} \|\hat{\pi}_a - \pi_{0,a}\|_{P_0^{\mathrm{C}},1} \|\hat{\mu}_a - \mu_{0,a}\|_{P_0^{\mathrm{C}},1} + \|\hat{h} - h_0\|_{P_0^{\mathrm{C}},1} \|\hat{q} - q_0\|_{P_0^{\mathrm{C}},1} = o_{P_0^{\mathrm{C}}}(n^{-1/2}),$$

we further express the difference as

$$\hat{\theta}^{\mathrm{C}} - \theta_0 = (P_n^{\mathrm{C}} - P_0^{\mathrm{C}})(\hat{\ell}^{\mathrm{C}} - \ell_0^{\mathrm{C}}) + P_n^{\mathrm{C}} \ell_0^{\mathrm{C}} - \theta_0 + \left(P^{\mathrm{C}} \hat{\ell}^{\mathrm{C}} - \frac{\alpha_0}{\hat{\alpha}} \theta_0 \right) - \frac{\hat{\alpha} - \alpha_0}{\hat{\alpha}} \theta_0$$

$$= P_n^{\mathrm{C}} \varphi_0^{\mathrm{C}} + (P_n^{\mathrm{C}} - P_0^{\mathrm{C}})(\hat{\ell}^{\mathrm{C}} - \ell_0^{\mathrm{C}}) + \frac{(\hat{\alpha} - \alpha_0)^2}{\alpha_0 \hat{\alpha}} \theta_0 + o_{P_0^{\mathrm{C}}}(n^{-1/2}).$$

The second term is an empirical process term of order $o_{P_0^{\mathrm{C}}}(n^{-1/2})$ if $\|\hat{\ell}^{\mathrm{C}} - \ell_0^{\mathrm{C}}\|_{P_0^{\mathrm{C}}} = o_{P_0^{\mathrm{C}}}(1)$, since $\mathcal{G}_0^{\mathrm{C}}$ is P_0^{C}-Donsker. By an application of the central limit theorem to $\hat{\alpha}$ and Slutsky's theorem, the third term is of the order $O_{P_0^{\mathrm{C}}}(n^{-1}) = o_{P_0^{\mathrm{C}}}(n^{-1/2})$. To conclude the proof of the second part of the theorem, we show that $\|\hat{\ell}^{\mathrm{C}} - \ell_0^{\mathrm{C}}\|_{P_0^{\mathrm{C}}}$ indeed converges in probability to zero. The $L_2(P_0^{\mathrm{C}})$-norm of the plugin function $\hat{\ell}^{\mathrm{C}}$ is

$$\|\hat{\ell}^{\mathrm{C}}\|_{P_0^{\mathrm{C}}} \leq M \Bigg\{ \left\| S\hat{q} \frac{(2A-1)}{e_0(A \mid X)} \frac{\Delta}{\hat{\pi}} Y \right\|_{P_0^{\mathrm{C}}} + \left\| S\hat{q} \frac{(2A-1)}{e_0(A \mid X)} \frac{\Delta}{\hat{\pi}} \hat{\mu} \right\|_{P_0^{\mathrm{C}}}$$

$$+ \|S\hat{q}(\hat{\mu}_1 - \hat{\mu}_0 - \hat{h})\|_{P_0^{\mathrm{C}}} + \|(1-S)\hat{h}\|_{P_0^{\mathrm{C}}} \Bigg\}$$

$$\leq M^4 \{E_{P^{\mathrm{C}}}(Y^2 \mid \Delta = 1, S = 1)\}^{1/2} + M^{7/2} \sum_{a \in \{0,1\}} \|\hat{\mu}_a\|_{P_0^{\mathrm{C}},1}$$

$$+ \left\{ M^2 \sum_{a \in \{0,1\}} \|\hat{\mu}_a\|_{P_0^{\mathrm{C}},1} + M^2 \|\hat{h}\|_{P_0^{\mathrm{C}},1} \right\} + M^{3/2} \|\hat{h}\|_{P_0^{\mathrm{C}},1}$$

$$= O_{P_0^{\mathrm{C}}}(1).$$

This is because the norms of the nuisance estimators $\|\hat{q}\|_{P_0^{\mathrm{C}},1} \leq M = O_{P_0^{\mathrm{C}}}(1)$, $\|\hat{h}\|_{P_0^{\mathrm{C}},1} \leq \|\hat{h} - \bar{h}\|_{P_0^{\mathrm{C}},1} + \|\bar{h}\|_{P_0^{\mathrm{C}},1} = O_{P_0^{\mathrm{C}}}(1)$ and $\|\hat{\mu}_a\|_{P_0^{\mathrm{C}},1} \leq \|\hat{\mu}_a - \bar{\mu}_a\|_{P_0^{\mathrm{C}},1} + \|\bar{\mu}_a\|_{P_0^{\mathrm{C}},1} = O_{P_0^{\mathrm{C}}}(1)$ are bounded by probability. The $L_2(P_0^{\mathrm{C}})$-distance between the plugin and the true function is

$$\|\hat{\ell}^{\mathrm{C}} - \ell_0^{\mathrm{C}}\|_{P_0^{\mathrm{C}}}$$

$$\leq \frac{1}{\alpha_0} \left\| S(\hat{q} - q_0) \frac{2A-1}{e_0} \frac{\Delta}{\pi_0} Y \right\|_{P_0^{\mathrm{C}}} + \frac{1}{\alpha_0} \left\| S\hat{q} \frac{2A-1}{e_0} \frac{\Delta(\pi_0 - \hat{\pi})}{\hat{\pi}\pi_0} Y \right\|_{P_0^{\mathrm{C}}}$$

$$+ \frac{1}{\alpha_0} \left\| S(\hat{q} - q_0) \frac{2A-1}{e} \frac{\Delta}{\pi_0} \mu_0 \right\|_{P_0^{\mathrm{C}}} + \frac{1}{\alpha_0} \left\| S\hat{q} \frac{2A-1}{e_0} \frac{\Delta(\pi_0 - \hat{\pi})}{\hat{\pi}\pi_0} \mu \right\|_{P_0^{\mathrm{C}}}$$

$$+\frac{1}{\alpha_0}\left\|S\hat{q}\frac{2A-1}{e_0}\frac{\Delta}{\hat{\pi}}(\hat{\mu}-\mu_0)\right\|_{P_0^{\mathrm{C}}}+\frac{1}{\alpha_0}\|S\{\hat{q}(\hat{\mu}_1-\hat{\mu}_0-\hat{h})-q_0(\mu_{0,1}-\mu_{0,0}-h_0)\}\|_{P_0^{\mathrm{C}}}$$
$$+\frac{1}{\alpha_0}\|(1-S)(\hat{h}-h_0)\|_{P_0^{\mathrm{C}}}+\frac{|\hat{\alpha}-\alpha_0|}{\alpha_0}\|\hat{\ell}^{\mathrm{C}}\|_{P_0^{\mathrm{C}}},$$

and by similar arguments above, we bound the distance by

$$\begin{aligned}
&\leq M^3[P_0^{\mathrm{C}}\{S(\hat{q}-q_0)^2E_{P_0^{\mathrm{C}}}(Y^2\mid\Delta=1,Z,X,S=1)\}]^{1/2}\\
&\quad+M^4\sum_{a\in\{0,1\}}[P_0^{\mathrm{C}}\{S(\hat{\pi}a-\pi_{0,a})^2E_{P_0^{\mathrm{C}}}(Y^2\mid\Delta=1,Z,W,A=a,X,S=1)\}]^{1/2}\\
&\quad+M^3\|\hat{q}-q_0\|_{P_0^{\mathrm{C}},q}+M^5\sum_{a\in\{0,1\}}\|\hat{\pi}a-\pi_{0,a}\|_{P_0^{\mathrm{C}},1}+M^{7/2}\sum_{a\in\{0,1\}}\|\hat{\mu}a-\mu_{0,a}\|_{P_0^{\mathrm{C}},1}\\
&\quad+\{3M^2\|\hat{q}-q_0\|_{P_0^{\mathrm{C}},1}+M^2\|\hat{\mu}_1-\mu_{0,1}\|_{P_0^{\mathrm{C}},1}+M^2\|\hat{\mu}_0-\mu_{0,0}\|_{P_0^{\mathrm{C}},1}+M^2\|\hat{h}-h_0\|_{P_0^{\mathrm{C}},1}\}\\
&\quad+M^{3/2}\|\hat{h}-h_0\|_{P_0^{\mathrm{C}},1}+M|\hat{\alpha}-\alpha_0|O_{P_0^{\mathrm{C}}}(1)\\
&=o_{P_0^{\mathrm{C}}}(1).
\end{aligned}$$

We conclude that $\hat{\theta}^{\mathrm{C}}-\theta_0=P_n^{\mathrm{C}}\varphi_0^{\mathrm{C}}+o_{P_0^{\mathrm{C}}}(n^{-1/2})$. □

S7.3. *Simulation*

In the simulation study for the indirect comparison estimator in the presence of missing outcomes, we generated the full data $(S\Delta,U,S,X,SA,Y,W,SZ)$ including an missing indicator Δ. We sampled from the distribution of (U,S,X,SA,Y,W,SZ) specified in §5.1 and drew Δ from the distribution

$$\Delta\mid(Z,W,A,X,S=1)\sim\text{Bernoulli}\{\text{expit}(0.1Z^{\mathrm{T}}1+0.1W^{\mathrm{T}}1+0.7A+0.3X^{\mathrm{T}}1)\}.$$

We only investigated the multiply robust estimator $\hat{\theta}^{\mathrm{C}}$. On the source RCT sample, we fitted the adherence probability model $\hat{\pi}(z,w,a,x)$ using a logistic regression linear in all covariates and the mean outcome model $\hat{\mu}(z,w,a,x)$ on the subsample where $\Delta=1$ using an ordinary linear regression with interaction between A and $(Z,W,X)^{\mathrm{T}}$. The nuisance estimator of the outcome difference bridge was subsequently obtained as

$$\hat{\eta}=\arg\min_{\eta'}\left\|\frac{1}{n_1}\sum_{i:S_i=1}\psi_{h_{\eta'},b,\hat{\pi},\hat{\mu}}(O_i)\right\|^2,$$

while the estimation of the participation odds bridge remained the same as in the setup without nonadherence. To demonstrate the robustness against model misspecifications, we considered the configurations where none of h, q, π, and μ was misspecified (experiment 14), where q and π were misspecified (experiment 15), where q and μ were misspecified (experiment 16), where h and π were misspecified (experiment 17), where h and μ were misspecified (experiment 18) and where all of h, q, π, and μ were misspecified (experiment 19). The misspecified models were fitted by replacing W and Z with $|W|^{1/2}$ and $|Z|^{1/2}$ wherever appropriate. However, the true model μ_0 for μ does not have an easy closed-form expression. Therefore, all the posited models for μ could have been misspecified, whether intentionally or not. The results are displayed in Table S4. The estimator $\hat{\theta}^{\mathrm{C}}$ retained small empirical biases under model misspecifications as expected, except when all nuisance models were misspecified. The influence-function-based standard error in experiments 15–16 where the q model was misspecified led to anticonservative confidence intervals.

Table S4. *Simulation results of experiments 14–19.*

n	Experiment	Mean	Bias	RMSE	SE	Coverage
1000	14	−2.64	1.57	1.19	1.15	93.9
	15	−2.64	0.33	1.24	1.12	91.8
	16	−2.64	3.08	3.08	2.68	90.7
	17	−2.64	5.97	1.24	1.27	95.7
	18	−2.63	11.51	3.30	3.16	94.0
	19	−2.40	248.88	3.73	2.72	82.3
2000	14	−2.65	−0.36	0.84	0.81	94.1
	15	−2.65	−0.24	0.84	0.77	93.1
	16	−2.65	−5.38	2.08	1.85	91.7
	17	−2.64	2.67	0.86	0.88	95.8
	18	−2.65	−3.09	2.23	2.22	95.1
	19	−2.39	250.89	3.15	1.85	72.8

Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-1}; SE: average of standard error estimates, 10^{-1}; Coverage: 95% confidence interval coverage, %.

References

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

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2024). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernán, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39(14):1999–2014.

Davies, M., Færch, L., Jeppesen, O. K., Pakseresht, A., Pedersen, S. D., Perreault, L., Rosenstock, J., Shimomura, I., Viljoen, A., Wadden, T. A., and Lingvay, I. (2021). Semaglutide 2.4 mg once a week in adults with overweight or obesity, and type 2 diabetes (STEP 2): A randomised, double-blind, double-dummy, placebo-controlled, phase 3 trial. *The Lancet*, 397(10278):971–984.

Davies, M. J., Bergenstal, R., Bode, B., Kushner, R. F., Lewin, A., Skjøth, T. V., Andreasen, A. H., Jensen, C. B., DeFronzo, R. A., and for the NN8022-1922 Study Group (2015). Efficacy of liraglutide for weight loss among patients with type 2 diabetes: The SCALE diabetes randomized clinical trial. *JAMA*, 314(7):687–699.

Kallus, N., Mao, X., and Uehara, M. (2022). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv*: 2103.14029v4.

Kesavan, S. (2022). *Nonlinear functional analysis: A first course*, volume 28 of *Texts and Readings in Mathematics*. Springer.

Rubino, D. M., Greenway, F. L., Khalid, U., O'Neil, P. M., Rosenstock, J., Sørrig, R., Wadden, T. A., Wizert, A., Garvey, W. T., and STEP 8 Investigators (2022). Effect of weekly subcutaneous semaglutide vs daily liraglutide on body weight in adults with overweight or obesity without diabetes: The STEP 8 randomized clinical trial. *JAMA*, 327(2):138–150.

Stewart, L. A., Clarke, M., Rovers, M., Riley, R. D., Simmonds, M., Stewart, G., Tierney, J. F., and for the PRISMA-IPD Development Group (2015). Preferred Reporting Items for a Systematic Review and Meta-analysis of Individual Participant Data: The PRISMA-IPD Statement. *JAMA*, 313(16):1657–1665.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. (2024). An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer Series in Statistics. Springer.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Manuscript II

Efficient estimation of target population average treatment effect from multi-source data

Zehao Su
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
zehao.su@sund.ku.dk

Helene C. Rytgaard
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
hely@sund.ku.dk

Henrik Ravn
Novo Nordisk A/S,
Vandtårnsvej 108, DK-2860 Søborg, Denmark
hnrv@novonordisk.com

Frank Eriksson
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
eriksson@sund.ku.dk

Abstract

We consider estimation of the average treatment effect (ATE) when outcome information is unavailable from the target population. Instead, we observe the outcome in multiple source populations and wish to combine the treatment effects therein to make inference on the target population ATE. Statistical analyses based on transportability methods typically standardize over subject characteristics to account for differences between these populations. In contrast to existing works that assume transportability on the conditional distribution of potential outcomes or conditional treatment-specific means, we work under a weaker form of effect transportability. In particular, we assume transportability of conditional average treatment effects across multiple populations, which may hold with fewer standardization variables. Under this assumption, we derive the semiparametric efficiency bound of the target population ATE. We characterize a class of doubly robust and asymptotically linear estimators that vary in the weights assigned to observations from different data sources. Specifically, we propose an efficient estimator whose asymptotic variance cannot be improved upon unless stronger transportability assumptions hold. For a low-dimensional summary of effect heterogeneity in the target population, we suggest estimating the projected conditional ATE. We illustrate the use of the proposed estimators on a multi-center weight management clinical trial for semaglutide, a glucagon-like peptide-1 receptor agonist, on patients with obesity. Using outcome information from other regions, we estimate the weight loss effect of semaglutide in the United States subgroup.

Keywords: Effect modification; Transportability; Meta-analysis; Semiparametric efficiency bound.

1. Introduction

One of the major tasks in health technology assessment (HTA) is to synthesize new evidence of interventions on a target population from a pool of existing data sources. Usually there are non-negligible differences between the target population and the population from which a data source is sampled. Given an intervention effect of interest, researchers compile a set of baseline covariates that summarize the interpopulation differences. These covariates are referred to as relevant variables for transportability. Loosely speaking, when individual patient data (IPD) are available, the synthesized effect on the target population can often be obtained by standardizing the effect from the data source through adjustment of the relevant variables or by weighting the samples from the data source to account for the distributional differences of the relevant variables.

Two practical questions are important when discussing evidence synthesis for the effect of an intervention. The first is the scope of the relevant variables. The basis of evidence synthesis methods is transportability of intervention effects across study populations. For transportability on summary measures involving a single intervention, the relevant variables requried are called prognostic variables in the meta-analysis literature. An instance of such summary measures is the conditional mean outcome of a treatment arm in a clinical trial. For transportability on effect measures, the required relevant variables are called effect modifiers (VanderWeele and Robins, 2007). They are a subset of prognostic variables, making them suitable for evidence synthesis as the total number of collected variables in all populations can be rather limited. Moreover, transportability on a single-intervention summary measure is unlikely to hold due to unmeasured factors that impact the outcome level, such as the overall quality of healthcare.

The second question concerns the overlap between populations. For example, if the data source is a clinical trial, the target population may include a group of subjects which do not fulfill the inclusion criteria of the trial. When there is a lack of population overlap, the synthesized evidence is subject to bias because it relies on extrapolating intervention effects. The external validity may be restored using a trimmed target population relative to the source population to avoid extrapolation (Chen et al., 2023). However, the new, artificially defined target population can be hard to translate into any sensible cohort in reality. When there are multiple source trials available that examine the interventions of interest, the overlap may be achieved by treating the source trials as one collective, large data source. While individual source trials may exhibit overlap violations with respect to the target population, the risk of unintended extrapolation is reduced when multiple trials are jointly taken into account.

Recently there has been a large mass of research works on transportability under the causal inference framework and meta-analysis methods focusing on causal estimands. Most methods on transportability and generalizability work under the premise of a single data source (Lee et al., 2022; Dahabreh et al., 2020; Josey et al., 2021; Li et al., 2023a). Some of these methods can be extended to handle multiple data sources, but would instead rely on unnecessarily restrictive transportability assumptions for the in-

tervention effect of interest. In the context of incorporating external controls in clinical trials, Li et al. (2023b) derived the efficiency bound for the fused intervention effect assuming transportability on the conditional mean under placebo. There is also a general framework for data fusion with multi-source longitudinal data (Li and Luedtke, 2023). It hinges on pairwise full overlap across fusible data sources as well as distribution-level transportability, so it does not answer the two questions that commonly arise in HTA and causal meta-analysis. Wang et al. (2024b) considered subgroup causal effect estimation under partial overlap across the multi-source data. Their proposal works under transportability of the conditional distribution of the potential outcome, which we avoid in this work.

Dahabreh et al. (2023) showed identifiability of the target population average treatment effect (ATE) under a weaker overlap condition by utilizing multiple source trials. The identifiability relies on pairwise transportability of the conditional average treatment effect (CATE) between the target population and a source trial, but only on the subset of relevant variables that are common to both study populations. Nonetheless, neither the semiparametric efficiency bound nor estimation procedures of the identified parameter has been studied. It is well-known that efficient estimators of a parameter can sometimes be constructed from the efficient influence function. Through the characterization of the influence functions, if there is more than one, we can gain insights into the estimation problem, especially when the identifiability conditions induce restrictions on the model. We present the efficient influence function of the target population ATE and propose asymptotically linear estimators, including one that can attain the semiparametric efficiency bound. Notably, we retain the weaker identifiability assumptions as in Dahabreh et al. (2023). Under the same assumptions, we study a simple form of effect heterogeneity on the target population, specifically the projection of the CATE onto a prespecified basis expansion.

2. Identifiability under CATE transportability

The data sources are a collection of m randomized clinical trials labeled by a discrete variable $D \in [m]$, where $[m]$ is a shorthand for the index set $\{1, \ldots, m\}$. To any participant in a source trial, a binary treatment $A \in \{0, 1\}$ is administered randomly, and the outcome Y is recorded at the end of the study. We assume that the outcome is real valued. Additionally in all source trials, a common set of baseline covariates X is measured, which contains the relevant variables for transportability. The total number of subjects from the source trials is denoted by $n_0 = \sum_{d \in [m]} n_{0d}$. Within the source trial $D = d$, the observed data is n_{0d} independent and identically distributed (i.i.d.) copies of the tuple (Y, A, X). From the target population, an i.i.d. sample of the covariates X of size n_1 is collected. The total sample size combining the source and the target population is thus $n = n_0 + n_1$. We further introduce a binary indicator G for whether a data point belongs to any source trial ($G = 0$) or the target population ($G = 1$).

The sample sizes of the source trials n_{0d} often cannot be controlled by researchers working on HTA, as they reflect the practical data collection decisions in the respective clinical trials. A common example is the minimum sample size determined from power calculations. Although the actual sampling of data is specific to the study population, it is helpful to view the complete sample $\mathcal{O} = \big\{\{(1 - G_i)Y_i, (1 - G_i)A_i, X_i, (1 - G_i)D_i, G_i\} : i = 1, \ldots, n\big\}$, as an i.i.d. sample from some joint distribution over the

observed data $O = \{(1-G)Y, (1-G)A, X, (1-G)D, G\}$. We make the following assumption on the sample sizes throughout the paper.

Assumption 1 (Sampling proportions). There exist fixed values $\alpha_{0d} \in (0,1)$ such that the proportions $n_{0d}/n \to \alpha_{0d}$ for $d \in [m]$ and $n_1/n \to \alpha$ when $n \to \infty$.

Then we have $\text{pr}(D = d) = \alpha_{0d}$ and $\text{pr}(G = 1) = \alpha$, while the marginal probabilities of D and G do not reflect the relative sizes of the underlying study population nor the sampling mechanism adopted. We can hypothesize the existence of an artificial superpopulation from which every data point is drawn. Random sampling in the superpopulation is asymptotically equivalent to actual biased sampling.

Let $Y(a)$ denote the potential outcome of Y under the static intervention of $a \in \{0,1\}$. We define parameter θ as the target population ATE $E\{Y(1) - Y(0) \mid G = 1\}$. Define the support of baseline covariates in the target population as $\mathcal{X}_1 = \{x : \text{pr}(G = 1 \mid X = x) > 0\}$. Let $\mathcal{D}_x = \{d : \text{pr}(D = d \mid X = x, G = 0) > 0\}$. This is the set of indices of the source trials whose supports of the baseline covariates cover the value $x \in \mathcal{X}_1$. When we condition on $D = d$, it is implicitly understood that we also condition on $G = 0$.

Assumption 2 (Identifiability).

(i) (Overlap) $\mathcal{D}_x \neq \emptyset$ for $x \in \mathcal{X}_1$;
(ii) (Positivity) $\text{pr}(A = a \mid X = x, D = d) > 0$ for $a \in \{0,1\}$, $x \in \mathcal{X}_1$, and $d \in \mathcal{D}_x$;
(iii) (Mean exchangeability) $E\{Y(a) \mid X = x, D = d\} = E\{Y(a) \mid A = a', X = x, D = d\}$ for $a, a' \in \{0,1\}$, $x \in \mathcal{X}_1$, and $d \in \mathcal{D}_x$;
(iv) (Consistency) $Y_i(a) = Y_i$ if $A_i = a$ for $a \in \{0,1\}$.
(v) (Transportability) $E\{Y(1) - Y(0) \mid X = x, G = 1\} = E\{Y(1) - Y(0) \mid X = x, D = d\}$ for $x \in \mathcal{X}_1$ and $d \in \mathcal{D}_x$.

Assumptions 2(i)–(ii) and (v) here correspond to Assumptions A5†, A3 and A4‡ in Dahabreh et al. (2023). Assumption 2(iii) is a weaker version of Assumption A2 in Dahabreh et al. (2023), which is sufficient for identification of the ATE. Though common in the causal inference literature, Assumption 2(iv) should be carefully examined in the source trials, particularly in the context of HTA. For instance, the presence of multiple versions of placebo can pose challenges to consistency. However, if the differences among them are negligible with regards to their acting mechanisms on the outcome, these placebos can be treated as a single entity. Transportability of the conditional effect measure only requires the transportability assumption to hold conditioning on shifted effect modifiers instead of all prognostic variables (Colnet et al., 2024). It should be noted that the scope of effect modifiers is largely dependent on the chosen effect measure. Furthermore, Assumption 2(v) is conducive to identifiability of the ATE, but can be futile for identifiability of other marginal causal effects. Therefore, it is highly advisable to evaluate the validity of such assumptions according to the concrete intervention effect.

Before presenting identifiability of the target parameter with the observed data, we introduce the following notations on the observed data distribution. Define the selection score of being in the target population as $\pi(x) = \text{pr}(G = 1 \mid X = x)$, the selection score of being in the source trial $D = d$ when in the source population as $\zeta(d \mid x) = \text{pr}(D = d \mid X = x, G = 0)$. In each source trial $D = d$, the propensity score of receiving treatment $A = a$ is $e(a \mid x, d) = \text{pr}(A = a \mid X = x, D = d)$, and the conditional

outcome mean under treatment $A = a$ is $\mu(a, x, d) = E(Y \mid A = a, X = x, D = d)$. By Assumption 2, when we fix the baseline characteristics, the difference between the conditional outcome means under the two interventions remains constant for any trial in the source trial pool; that is, for any $x \in \mathcal{X}_1$ and $d, d' \in \mathcal{D}_x$, we have

$$\mu(1, x, d) - \mu(0, x, d) = \mu(1, x, d') - \mu(0, x, d') = \delta(x). \tag{1}$$

Note that the difference $\delta(x)$ does not vary across the source trials for any $x \in \mathcal{X}_1$ and is a function of the baseline covariates only. For simplicity, we also call this function the CATE.

Define the model $\mathcal{P}$ as the collection of probability measures on O that respect the conditional mean difference restriction (1).

Lemma 1 (Identifiability). *Suppose Assumption 2 holds. The target parameter is identifiable in the observed data distribution $P \in \mathcal{P}$ as*

$$\theta = E\{\delta(X) \mid G = 1\}. \tag{2}$$

The g-formula representation of the target parameter on the observed data does not depend directly on the membership to any source trial D, while the difference function $\delta(x)$ implicitly involves this information, as it is defined strictly on the subset of the source trials $\mathcal{D}_x$ for a given level of baseline covariates x.

3. Efficient estimation of the transported ATE

3.1. *Semiparametric efficiency bound under CATE transportability*

For the well-defined observed data target parameter θ, it is natural to study its semiparametric efficiency bound as well as estimators that can achieve this bound. We first motivate the efficient influence function of θ through a class of candidate estimators.

The representation of θ as (2) is not unique. For all $h(x, d)$ such that $E\{h(X, D) \mid X = x, G = 0\} = 1$, the target parameter is also identifiable as

$$\theta = \frac{1}{\alpha} E\left\{ \frac{(1 - G)\pi(X)}{1 - \pi(X)} \frac{2A - 1}{e(A \mid X, D)} h(X, D) Y \right\}. \tag{3}$$

The function h in the inverse probability weighting representation in (3) can be thought of as assigning trial-specific weights for subjects in the pooled source trial population. In the special case when $h(x, d) = 1$, the outcomes from all source trials are weighted the same besides the inverse propensity score of their corresponding treatment arm. Suppose we have the knowledge of the true nuisance parameters, then the representation (3) suggests a familiar class of reweighting estimators

$$\frac{1}{n_0} \sum_{i: G_i = 0} \frac{\pi(X_i)}{1 - \pi(X_i)} \frac{2A_i - 1}{e(A_i \mid X_i, D_i)} h(X_i, D_i) Y_i.$$

To minimize the variance of estimators in this class, the outcomes from source trials with a higher conditional variance should be downweighed. Let the conditional variance

of the outcome under treatment arm $A = a$ in source trial $D = d$ be $V(a, x, d) = \mathrm{var}(Y \mid A = a, X = x, D = d)$ for $x \in \mathcal{X}_1$ and $d \in \mathcal{D}_x$. Now consider the weighting function

$$w(x, d) = \left\{ \frac{V(1, x, d)}{e(1 \mid x, d)} + \frac{V(0, x, d)}{e(0 \mid x, d)} \right\}^{-1}. \tag{4}$$

It can be rescaled so that the function $w(x, d) \big/ \sum_{d' \in [m]} w(x, d') \zeta(d' \mid x)$ can take the place of $h(x, d)$ in (3). This weighting function is optimal in the sense of Lemma 2 below.

In the sequel, we assume the data follows some true distribution $P_0 \in \mathcal{P}$, since the semiparametric efficiency bound is derived under local regularity conditions at P_0. Quantities defined on P_0 receive the subscript 0.

Assumption 3 (Regularity condition). There exists a universal constant $C > 1$ such that $e_0(a \mid x, d) \geqslant C^{-1}, V_0(a, x, d) \leqslant C, C^{-1} \leqslant w_0(x, d) \leqslant C$, and $|y - \mu_0(a, x, d)| \leqslant C$.

Lemma 2 (Efficient influence function). *Suppose Assumption 3 holds. The efficient influence function of θ_0 at the distribution P_0 under the nonparametric model $\mathcal{P}$ is*

$$\varphi_{w_0}(o) = \frac{1 - g}{\alpha_0} \frac{\pi_0(x)}{1 - \pi_0(x)} \frac{w_0(x, d)}{\sum_{d' \in [m]} \zeta_0(d' \mid x) w_0(x, d')} \frac{2a - 1}{e_0(a \mid x, d)} \{y - \mu_0(a, x, d)\} + \frac{g}{\alpha_0} \{\delta_0(x) - \theta_0\}. \tag{5}$$

To understand the lemma, consider the ATE $E_0\{\mu_0(1, X, d) - \mu_0(0, X, d) \mid D = d\}$ in the population $D = d$ defined on the model of probability distributions over $I(D = d)O$. The semiparametric efficiency bound of this parameter (Hirano et al., 2003) is

$$E_0\{w_0^{-1}(X, d) \mid D = d\} + \mathrm{var}_0\{\mu_0(1, X, d) - \mu_0(0, X, d) \mid D = d\},$$

which consists of an expectation part and a variance part. The semiparametric efficiency bound Ω_0 of θ_0 has a similar structure. By some algebra, we have

$$\Omega_0 = E_0 \varphi_{w_0}^2(O) = \frac{1 - \alpha_0}{\alpha_0^2} E_0 \left[\left\{ \frac{\pi_0(X)}{1 - \pi_0(X)} \right\}^2 \left\{ \sum_{d \in [m]} \zeta_0(d \mid X) w_0(X, d) \right\}^{-1} \,\middle|\, G = 0 \right] + \frac{1}{\alpha_0} \mathrm{var}_0\{\delta_0(X) \mid G = 1\}.$$

Each data source contributes to Ω_0 with $w_0(x, d)$, which are first weighted by the selection score $\zeta_0(d \mid x)$ and then inverted to give the inside of the expectation part of the semiparametric efficiency bound, up to the odds of selection score $\pi_0(x)$ and constants.

The following result is a direct consequence of Lemma 2 and its proof.

Corollary 1 (Influence functions). *Suppose Assumption 3 holds. The linear subspace*

$$\Lambda_0 = \left\{ \frac{1 - g}{\alpha_0} h(x, d) \frac{2a - 1}{e_0(a \mid x, d)} \{y - \mu_0(a, x, d)\} : E_0\{h(X, D) \mid X, G = 0\} = 0 \right\}$$

is the orthogonal complement of the tangent space of the model $\mathcal{P}$ at P_0. Consequently, for any weight function $\tilde{w}(x, d)$ such that $\sum_{d' \in [m]} \zeta_0(d' \mid x) \tilde{w}(x, d') \neq 0$ for $x \in \mathcal{X}_1$, $\varphi_{\tilde{w}}(o)$ is an influence function of θ_0.

Corollary 1 indicates that if we wish to construct estimators based on influence functions, the weights used for the recalibration of observations from different source trials need not be the optimal weight w_0 in Lemma 2.

3.2. *Efficient estimation of the target parameter*

The efficient influence function (5) motivates an estimating equation for the target parameter θ_0. In the following, we describe an estimation procedure with crossfitted nuisance parameters (Zheng and van der Laan, 2011; Chernozhukov et al., 2018) and analyze the resulting crossfitted estimator of θ_0. Consider a random partition of the data into K splits with index sets $\mathcal{I}_k$ such that $\cup_{k\in[K]}\mathcal{I}_k = [n]$. Without loss of generality, assume every index set has cardinality n/K. Let $\mathcal{O}_{-k} = \{O_i : i \notin \mathcal{I}_k\}$ denote the observations not belonging to split k. On each $\mathcal{O}_{-k}$, the proportion of samples from the target population $\hat{\alpha}_k$ is a natural estimator of α_0. We have outcome regression models $\hat{\mu}_k(a, x, d)$ for the conditional outcome mean $\mu_0(a, x, d)$ of the intervention $A = a$ in source trial $D = d$, and a selection score model $\hat{\zeta}_k(d \,|\, x)$ among the source trials for the probability $\zeta_0(d \,|\, x)$. Within each source trial the propensity scores for treatment assignment $\hat{e}_k(a \,|\, x, d)$ model the probability $e_0(a \,|\, x, d)$. The selection score model into the target population $\hat{\pi}_k(x)$ approximates the probability $\pi_0(x)$. Additionally, we can choose possibly random weights $\check{w}_k(x, d)$ and rescale them as $\hat{w}_k(x, d) = \check{w}_k(x, d)/\big\{\sum_{d'\in[m]} \check{w}_k^2(x, d')\big\}^{1/2}$ to ensure that the weight vector $\{\hat{w}_k(x, 1), \ldots, \hat{w}_k(x, m)\}^{\mathrm{T}}$ is normalized for every x.

We do not require the sample version of (1). That is, we allow the difference between the fitted conditional means $\hat{\mu}_k(1, x, d) - \hat{\mu}_k(0, x, d)$ to vary across d. Rather, we estimate the difference function by

$$\hat{\delta}_k(x) = \sum_{d\in[m]} \frac{\hat{w}_k(x, d)\hat{\zeta}_k(d \,|\, x)\{\hat{\mu}_k(1, x, d) - \hat{\mu}_k(0, x, d)\}}{\sum_{d'\in[m]} \hat{w}_k(x, d')\hat{\zeta}_k(d' \,|\, x)}.$$

In light of the structure of influence functions described in Corollary 1, we propose the estimator

$$\hat{\theta} = \frac{1}{n} \sum_{k\in[K]} \sum_{i\in\mathcal{I}_k} \ell_{\hat{\eta}_k}(O_i),$$

where

$$\ell_{\hat{\eta}_k}(o) = \frac{1-g}{\hat{\alpha}_k} \frac{\hat{\pi}_k(x)}{1-\hat{\pi}_k(x)} \frac{\hat{w}_k(x, d)}{\sum_{d'\in[m]} \hat{\zeta}_k(d' \,|\, x)\hat{w}_k(x, d')} \frac{2a-1}{\hat{e}_k(a \,|\, x, d)}\{y - \hat{\mu}_k(a, x, d)\} + \frac{g}{\hat{\alpha}_k}\hat{\delta}_k(x)$$

is indexed by the set of crossfitted nuisance parameters $\hat{\eta}_k = \{\hat{\alpha}_k, \hat{\pi}_k, \hat{\zeta}_k, \hat{w}_k, \hat{e}_k, \hat{\mu}_k, \hat{\delta}_k\}$.

Assumption 4 (Regularity conditions).

(i) The probability limits with respect to the $L_2(P_0)$-norm are well-defined for the nuisance parameter estimates such that

$$\|(\hat{\pi}_k - \bar{\pi})(X)\|_{P_0} = o_{P_0}(1), \|(\hat{\zeta}_k - \bar{\zeta})(d \,|\, X)\|_{P_0} = o_{P_0}(1),$$
$$\|(\hat{w}_k - \bar{w})(X, d)I\{\zeta_0(d \,|\, X) > 0\}\|_{P_0} = o_{P_0}(1),$$

$$\|(\hat{e}_k - \bar{e})(a \mid X, d)\}I\{\zeta_0(d \mid X) > 0\}\|_{P_0} = o_{P_0}(1),$$
$$\|\{\hat{\mu}_k(a, X, d) - \bar{\mu}(a, X, d)\}I\{\zeta_0(d \mid X) > 0\}\|_{P_0} = o_{P_0}(1);$$

(ii) There exists a universal constant $C > 1$ such that

$$\alpha_0 \geqslant C^{-1}, \hat{\alpha} \geqslant C^{-1}, \hat{\pi}_k(x) \leqslant 1 - C^{-1}, e_0(a \mid x, d) \geqslant C^{-1}, \hat{e}_k(a \mid x, d) \geqslant C^{-1},$$
$$|\hat{\mu}_k(a, x, d)| \leqslant C, V_0(a, x, d) \leqslant C, |\hat{w}_k(x, d)| \leqslant C, |\bar{w}(x, d)| \leqslant C,$$
$$\left|\sum_{d' \in [m]} \hat{\zeta}_k(d' \mid x)\hat{w}_k(x, d')\right| \geqslant C^{-1}, \left|\sum_{d' \in [m]} \zeta_0(d' \mid x)\bar{w}(x, d')\right| \geqslant C^{-1}.$$

Assumption 5a (Correct specifications). Either $\bar{\mu} = \mu_0$ or $\bar{e} = e_0$, $\bar{\zeta} = \zeta_0$, and $\bar{\pi} = \pi_0$.

Assumption 5b (Rate conditions).

(i) $\bar{\mu} = \mu_0$, $\bar{e} = e_0$, $\bar{\zeta} = \zeta_0$, and $\bar{\pi} = \pi_0$;

(ii) $\|(\hat{\mu}_k - \bar{\mu})(a, X, d)I\{\zeta_0(d \mid X) > 0\}\|_{P_0}\big\{\|(\hat{e}_k - \bar{e})(a \mid X, d)I\{\zeta_0(d \mid X) > 0\}\|_{P_0} + \|(\hat{\zeta}_k - \bar{\zeta})(d \mid X)\|_{P_0} + \|(\hat{\pi}_k - \bar{\pi})(X)\|_{P_0}\big\} = o_{P_0}(n^{-1/2})$.

Theorem 1 (Asymptotic behavior). *Suppose Assumption 4 holds. Then:*

1. $\hat{\theta} \xrightarrow{\text{p}} \theta_0$ *under Assumption 5a.*
2. $\hat{\theta} - \theta_0 = n^{-1}\sum_{i=1}^{n} \varphi_{\bar{w}}(O_i) + o_{P_0}(n^{-1/2})$ *under Assumption 5b. Furthermore,* $\hat{\theta}$ *achieves the local semiparametric efficiency bound* Ω_0 *if there exists a function* $c(x) \neq 0$ *such that*

$$\bar{w}(x, d) = c(x)w_0(x, d)$$

and Assumption 3 holds.

The influence-function-based estimator $\hat{\theta}$ is doubly robust in the sense that it is consistent when either the outcome regression model is correctly specified or when the selection score model for the target population, the selection score model among source trials, and the propensity score model for the treatment assignment are all correctly specified. In practice, the propensity score model e_0 is usually known or estimable with parametric rates in the respective source trials. Therefore, the product term $\|\hat{e}_k - e_0\|_{P_0}\|\hat{\mu}_k - \mu_0\|_{P_0}$ is often negligible asymptotically. However, the models $\{\pi_0, \zeta_0\}$ are generally more complicated because they inherit the complexity of the sampling procedure, as well as the pragmatic choice of source trials to include in the analysis. Therefore, it can be important that the CATE estimator $\hat{\delta}_k$ and hence $\hat{\mu}_k$ are correctly specified for the consistency of $\hat{\theta}$. The asymptotic linearity of the target parameter estimate further requires that all nuisance estimators converge to the truth at a reasonable, possibly subparametric rate, such as $o_{P_0}(n^{-1/4})$. This rate is achieved by flexible curve-fitting algorithms, such as the highly adaptive lasso estimator (Benkeser and van der Laan, 2016) and the sieve neural network (Chen and White, 1999).

An estimator of the asymptotic variance of $\hat{\theta}$ is given by the crossfitted squared empirical L_2-norm of the influence function

$$\hat{\Omega} = \frac{1}{n}\sum_{k \in [K]}\sum_{i \in \mathcal{I}_k}\left\{\ell_{\hat{\eta}_k}(O_i) - \frac{G_i}{\hat{\alpha}_k}\hat{\theta}\right\}^2.$$

We show its consistency in the Supplementary Material §S2. Therefore, an asymptotic

$(1-\gamma)$-confidence interval of θ_0 is given by $\hat{\theta} \pm n^{-1/2}\Phi^{-1}(1-\gamma/2)\hat{\Omega}^{1/2}$, where Φ is the distribution function of a standard normal random variable.

3.3. *Choice of weight functions*

The choice of weight function $\hat{w}_k(x,d)$ does not affect the asymptotic linearity of the estimator beyond Assumption 4. However, only the particular specification of the weight function in (4) yields the efficient estimator, up to some scaling function $c(x)$. We can pick non-source-specific weights such that $\hat{w}_k(x)$ is constant in d for convenience. For these weights, the asymptotic linearity of $\hat{\theta}$ can be established without consistency of $\hat{\zeta}_k(d \mid x)$. However, precision may be gained from estimation of the weights. Suppose there exist two regular asymptotically linear estimators for θ_0, one with the efficient influence function φ_{w_0}, and the other with influence function $\varphi_{\tilde{w}}$ using a weight $\tilde{w}(x)$ that does not vary with d. The difference of their asymptotic variances, shown in Supplementary Material §S3.1, is

$$E_0(\varphi_{w_0}^2 - \varphi_{\tilde{w}}^2) = E_0\left(\frac{\pi_0(X)}{\alpha_0\{1-\pi_0(X)\}}\left[\left\{\sum_{d'\in[m]} w_0(X,d')\zeta_0(d' \mid X)\right\}^{-1} - \left\{\sum_{d'\in[m]} w_0^{-1}(X,d')\zeta_0(d' \mid X)\right\}\right] \,\middle|\, G=1\right) \leqslant 0,$$

where the bound follows from the Cauchy-Schwarz inequality.

A straightforward estimator of the optimal weight function can be constructed by plugging in the nuisance estimators $\hat{e}_k(a \mid x,d)$ and $\hat{V}_k(a,x,d)$. However, the optimal weight function involves inverting the fitted conditional variances, which can introduce numerical instability to the estimator $\hat{\theta}$. Therefore, nonparametric conditional variance estimators, such as the local kernel linear regression (Fan and Yao, 1998), may deteriorate finite sample performance of the estimator. To circumvent the inconvenience, we re-express the optimal weight function as the inverse of a conditional expectation

$$w_0(x,d) = \left(E_0\left[\left\{\frac{Y-\mu_0(A,X,D)}{e_0(A \mid X,D)}\right\}^2 \,\middle|\, X=x, D=d\right]\right)^{-1}.$$

Following Hines et al. (2024), we notice that w_0 is the minimizer of the weighted regression loss

$$E_0\left(\left\{\frac{Y-\mu_0(A,X,D)}{e_0(A \mid X,D)}\right\}^2\left[\left\{\frac{Y-\mu_0(A,X,D)}{e_0(A \mid X,D)}\right\}^{-2} - w(X,D)\right]^2 \,\middle|\, G=0\right).$$

Given a function class $\mathcal{F}$, we propose to minimize the empirical loss so that

$$\check{w}_{k,\text{optimal}}(x,d) = \arg\min_{f\in\mathcal{F}} \sum_{i:i\in\mathcal{I}_k, G_i=0}\left[-2f(X_i,D_i) + \left\{\frac{Y_i-\hat{\mu}_k(A_i,X_i,D_i)}{\hat{e}_k(A_i \mid X_i,D_i)}\right\}^2 f^2(X_i,D_i)\right]. \quad (6)$$

Under the conditions of Theorem 1, if the estimated weight function is consistent for the optimal weight function, the resulting target population ATE estimator will be efficient.

Notably, efficiency does not assume any convergence rate of the weight function beyond consistency. Nonetheless, if either $\hat{\mu}_k$ or $\hat{e}_k$ is misspecified or if the hypothesis class $\mathcal{F}$ does not contain functions that can be normalized to w_0, there is no guarantee that the asymptotic variance of the resulting estimator will be lower than the estimator using constant weights.

3.4. *Semiparametric efficiency bounds under other transportability assumptions*

Stronger transportability assumptions typically induce models smaller than $\mathcal{P}$. In the following, we present the semiparametric efficiency bounds under these models.

Let the model $\mathcal{P}^\dagger$ contain all probability measures satisfying the conditional mean restriction

$$\mu(a,x,d) = \mu(a,x,d') = \mu(a,x), \tag{7}$$

for $x \in \mathcal{X}_1$, $d, d' \in \mathcal{D}_x$ and $a \in \{0,1\}$. Restriction (7) is implied by the transportability of conditional treatment-specific means $E\{Y(a) \mid X = x, D = d\} = E\{Y(a) \mid X = x, D = d'\}$ and Assumptions 2(ii)–(iv). Define $\zeta(d \mid a,x) = \text{pr}(D = d \mid A = a, X = x)$ and $e(a \mid x) = \text{pr}(A = a \mid X = x, G = 0)$.

Assumption 6 (Regularity condition). There exists a universal constant $C > 1$ such that $e_0(a \mid x) \geqslant C^{-1}$, $V_0(a,x,d) \leqslant C$, and $|y - \mu_0(a,x)| \leqslant C$.

Proposition 1. *Suppose Assumption 6 holds. The efficient influence function of θ_0 at $P_0 \in \mathcal{P}^\dagger$ is*

$$\varphi^\dagger_{w_0^\dagger}(o) = \frac{1-g}{\alpha_0}\frac{\pi_0(x)}{1-\pi_0(x)}\frac{2a-1}{e_0(a \mid x)}\frac{w_0^\dagger(a,x,d)}{\sum_{d' \in [m]} w_0^\dagger(a,x,d')\zeta_0(d' \mid a,x)}\{y - \mu_0(a,x)\} + \frac{g}{\alpha_0}\{\delta_0(x) - \theta_0\},$$

where the weights $w_0^\dagger(a,x,d) = V_0^{-1}(a,x,d)$. The orthocomplement of the tangent space of $\mathcal{P}^\dagger$ at P_0 is

$$\Lambda_0^\dagger = \left\{(1-g)\frac{2a-1}{e_0(a \mid x)}q(a,x,d)\{y - \mu_0(a,x)\} : E_0\{q(A,X,D) \mid A,X,G=0\} = 0\right\}.$$

For any $\tilde{w}(a,x,d)$ such that $\sum_{d' \in [m]} \tilde{w}(a,x,d')\zeta_0(d' \mid a,x) \neq 0$, $\varphi^\dagger_{\tilde{w}^\dagger}$ is an influence function of θ_0.

The weight $w_0^\dagger(a,x,d)$ can be different for observations from different treatment arms in the same source population, in contrast to the weights $w_0(x,d)$ previously seen in Lemma 2, which are only source-specific. This is a generalization of the result from Li et al. (2023b) to accommodate multiple source trials.

Consider the model $\mathcal{P}^\ddagger$ consisting of probability measures on O where the only constraint is $Y \perp\!\!\!\perp D \mid (A,X,G=0)$. This corresponds to the distribution-level transportability $Y(a) \perp\!\!\!\perp D \mid (X, G=0)$ in Wang et al. (2024b). They showed that the efficient influence function of θ_0 under $P_0 \in \mathcal{P}^\ddagger$ is

$$\varphi^\ddagger(o) = \frac{1-g}{\alpha_0}\frac{\pi_0(x)}{1-\pi_0(x)}\frac{2a-1}{e_0(a \mid x)}\{y - \mu_0(a,x)\} + \frac{g}{\alpha_0}\{\delta_0(x) - \theta_0\}.$$

For completeness, we provide the orthocomplement of the tangent space of the model $\mathcal{P}^{\ddagger}$ at P_0, which is

$$\Lambda_0^{\ddagger} = \big\{(1-g)h(y,a,x,d) : E_0\{h(Y,A,X,D) \mid A,X,D\} = 0, \\ E_0\{h(Y,A,X,D) \mid Y,A,X,G=0\} = 0\big\}.$$

The three models discussed so far are nested such that $\mathcal{P}^{\ddagger} \subset \mathcal{P}^{\dagger} \subset \mathcal{P}$. Modulo regularity assumptions, φ_{w_0} from Lemma 2 and and $\varphi_{\tilde{w}}$ from Corollary 1 are valid influence function of θ_0 if P_0 belongs to the smaller models $\mathcal{P}^{\dagger}$ or $\mathcal{P}^{\ddagger}$.

4. Target-population effect heterogeneity estimation

We have assumed CATE transportability to make inference on the target population average treatment effect. In many cases, the CATE function itself may be of interest, such as in meta-analyses with IPD (Rubin, 1992). With the combined sample size from multiple data sources, we may have enough precision to find a larger body of evidence than a single-parameter summary of the intervention effect on the target population, which is often encouraged in HTA. For example, practitioners and policymakers might seek to identify specific subgroups within the target population who could either benefit from or be harmed by the intervention, while the outcome and intervention information is only available from existing clinical trials.

Parametric formulations of causal effect heterogeneity are useful in many clinical settings thanks to their interpretability. In some applications, it may be reasonable to describe effect heterogeneity in a semiparametric model where the CATE function $\delta(x)$ is known up to a Euclidean parameter. In §S3.2 of the Supplementary Material, we derive the efficient score of this parameter. In this section, we describe an alternative finite-dimensional parameterization of effect heterogeneity related to CATE and present an efficient estimation strategy.

For a subset of the baseline covariates $Z \subset X$, consider the basis function

$$\psi(z) = \{\psi_1(z), \ldots, \psi_q(z)\}^{\mathrm{T}}$$

with a fixed dimension $q \geqslant 1$. A natural description of effect heterogeneity is the projected CATE (Semenova and Chernozhukov, 2021; Cui et al., 2023), which is the best approximation of the CATE $\delta_0(x)$ in the linear span of the basis $\psi(z)$. The function $\beta_0^{\mathrm{T}}\psi(z)$, where

$$\beta_0 \in \arg\min_{\beta \in \mathbb{R}^q} E_0[\{\delta_0(X) - \beta^{\mathrm{T}}\psi(Z)\}^2 \mid G=1],$$

is a low-dimensional characterization of the CATE in the nonparametric model in the target population. If Z comprises only categorical variables and the basis $\psi(z)$ is dummy variables, the coefficients in β_0 reduce to subgroup-specific target population ATEs (Wang et al., 2024b). Let $\|v\|$ denote the Euclidean norm of $v \in \mathbb{R}^q$. Under the following assumption, β_0 is uniquely identifiable.

Assumption 7 (Uniqueness). $\|\psi\| \in L_2(P_0)$ and $E_0\{\psi^{\otimes 2}(Z) \mid G=1\}$ is invertible.

Proposition 2. *Suppose Assumptions 3 and 7 hold. The efficient influence function of β_0 at $P_0 \in \mathcal{P}$ is*

$$\varphi^{\beta}_{w_0}(o) = \left[E_0\{\psi^{\otimes 2}(Z) \mid G = 1\}\right]^{-1}\psi(z)$$
$$\left[\frac{1-g}{\alpha_0}\frac{\pi_0(x)}{1-\pi_0(x)}\frac{w_0(x,d)}{\sum_{d'\in[m]}\zeta_0(d' \mid x)w_0(x,d')}\frac{2a-1}{e_0(a \mid x,d)}\{y-\mu_0(a,x,d)\}\right.$$
$$\left. + \frac{g}{\alpha_0}\{\delta_0(x) - \beta_0^{\mathrm{T}}\psi(z)\}\right].$$

The efficient influence function $\varphi^{\beta}_{w_0}$ motivates the crossfitted least-squares estimator

$$\hat{\beta} = \left\{\frac{1}{n_1}\sum_{i:G_i=1}\psi^{\otimes 2}(Z_i)\right\}^{-1}\left\{\frac{1}{n}\sum_{k\in[K]}\sum_{i\in\mathcal{I}_k}\psi(Z_i)\ell_{\hat{\eta}_k}(O_i)\right\}.$$

Denote the support of Z in the target population by $\mathcal{Z}_1$. We have the following result for $\hat{\beta}$.

Theorem 2 (Asymptotic behavior). *Suppose Assumptions 4 and 7 hold and that* $\sup_{z\in\mathcal{Z}_1}\|\psi(z)\| \leqslant C$ *for a universal constant* $C > 0$. *Then* $\hat{\beta} \xrightarrow{\mathrm{p}} \beta_0$ *under Assumption 5a, and* $\hat{\beta} - \beta_0 = n^{-1}\sum_{i=1}^{n}\bar{\varphi}^{\beta}_{\bar{w}}(O_i) + o_{P_0}(n^{-1/2})$ *under Assumption 5b.*

A pointwise asymptotic $(1-\gamma)$-confidence interval of $\beta_0^{\mathrm{T}}\psi(z)$ for any $z \in \mathcal{Z}_1$ can be constructed as

$$\hat{\beta}^{\mathrm{T}}\psi(z) \pm n^{-1/2}\Phi^{-1}(1-\gamma/2)\{\psi^{\mathrm{T}}(z)\hat{\Omega}^{\beta}\psi(z)\}^{1/2},$$

where we use the sandwich estimator

$$\hat{\Omega}^{\beta} = \left\{\frac{1}{n_1}\sum_{i:G_i=1}\psi^{\otimes 2}(Z_i)\right\}^{-1}\left[\frac{1}{n}\sum_{k\in[K]}\sum_{i\in\mathcal{I}_k}\psi^{\otimes 2}(Z_i)\left\{\ell_{\hat{\eta}_k}(O_i) - \frac{G_i}{\hat{\alpha}}\hat{\beta}^{\mathrm{T}}\psi(Z_i)\right\}^2\right]$$
$$\left\{\frac{1}{n_1}\sum_{i:G_i=1}\psi^{\otimes 2}(Z_i)\right\}^{-1}$$

of the asymptotic variance $\bar{\Omega}^{\beta} = E_0\{(\varphi^{\beta}_{\bar{w}})^{\otimes 2}\}$ of $\hat{\beta}$. We show its consistency in the Supplementary Material §S2.

Under stronger regularity assumptions, we establish uniform inference of the projected CATE over $\mathcal{Z}_1$. Let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the mininum and maximum eigenvalues of a matrix M.

Assumption 8 (Bounded basis). There exist universal constants $C \geqslant 1$ such that $C^{-1} \leqslant \inf_{z\in\mathcal{Z}_1}\|\psi(z)\| \leqslant \sup_{z\in\mathcal{Z}_1}\|\psi(z)\| \leqslant C$ and $C^{-1} \leqslant \lambda_{\min}(\bar{\Omega}^{\beta}) \leqslant \lambda_{\max}(\bar{\Omega}^{\beta}) \leqslant C$.

Corollary 2 (Weak convergence). *Suppose Assumptions 4, 5b, 7, and 8 hold. Then*

$$\frac{n^{1/2}\psi^{\mathrm{T}}(z)(\hat{\beta}-\beta_0)}{\{\psi^{\mathrm{T}}(z)\hat{\Omega}^{\beta}\psi(z)\}^{1/2}} \rightsquigarrow \mathbb{T}(z) \qquad \text{in } \ell^{\infty}(\mathcal{Z}_1),$$

where $\ell^{\infty}(\mathcal{Z}_1)$ *is the space of bounded functions over* $\mathcal{Z}_1$, *and* $\mathbb{T}(z)$ *is a mean-zero Gaussian process over* $\mathcal{Z}_1$ *with covariance function*

$$\mathrm{cov}_0\{\mathbb{T}(z), \mathbb{T}(z')\} = \frac{\psi^{\mathrm{T}}(z)\bar{\Omega}^{\beta}\psi(z')}{\{\psi^{\mathrm{T}}(z)\bar{\Omega}_{\beta}\psi(z)\}^{1/2}\{\psi^{\mathrm{T}}(z')\bar{\Omega}^{\beta}\psi(z')\}^{1/2}}.$$

A uniform asymptotic $(1-\gamma)$-confidence interval of $\beta_0^{\mathrm{T}}\psi(z)$ is

$$\hat{\beta}^{\mathrm{T}}\psi(z) \pm n^{-1/2} c_T(1-\gamma)\{\psi^{\mathrm{T}}(z)\hat{\Omega}^{\beta}\psi(z)\}^{1/2},$$

where c_T is the quantile function of $T = \sup_{z\in\mathcal{Z}_1} |\mathbb{T}(z)|$, the supremum of the Gaussian process over $\mathcal{Z}_1$. The distribution of T depends on unknown nuisance parameters. In practice, it can be approximated via Monte-Carlo methods, such as t-bootstrap (Belloni et al., 2015) and multiplier bootstrap (Belloni et al., 2018).

5. Simulated data example

We consider $m = 3$ source trials in the simulation study. The baseline covariates $X = (X_1, X_2, X_3)^{\mathrm{T}}$ include three continuous measurements. Let $\tilde{X} = (1, X^{\mathrm{T}})^{\mathrm{T}}$. We generate the observed data $O = \{(1-G)Y, (1-G)A, X, (1-G)D, G\}$ sequentially as follows:

$$\begin{aligned}
X &\sim 2\Phi[\mathrm{Normal}\{(0,0,0)^{\mathrm{T}}, \Sigma\}] - 1,\\
G \mid X &\sim \mathrm{Bernoulli}\{\mathrm{expit}(\xi_g^{\mathrm{T}}\tilde{X})\},\\
D \mid (X, G=0) &\sim \mathrm{Multinomial}\{\mathrm{softmax}(1, \xi_{d1}^{\mathrm{T}}\tilde{X}, \xi_{d2}^{\mathrm{T}}\tilde{X})\},\\
A \mid (X, D) &\sim \mathrm{Bernoulli}(e_D),\\
Y \mid (A, X, D) &\sim \mathrm{Normal}\{A\delta(X) + \mu(0, X, D), \sigma^2(X, D)\}.
\end{aligned}$$

where

$$\begin{aligned}
\Sigma &= \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix},\\
\xi_g &= \{-\log 3, \log(1.5), \log(1.5), \log(1.5)\}^{\mathrm{T}},\\
\xi_{d1} &= \{\log(1.5), \log(1.5), \log(1.5), \log(1.5)\}^{\mathrm{T}},\\
\xi_{d2} &= \{-\log(0.75), \log(0.75), \log(0.75), \log(0.75)\}^{\mathrm{T}},\\
e_d &= I(d=1)0.5 + I(d=2)0.4 + I(d=3)0.6,\\
\delta(x) &= 1 + 0.5x_1 - 0.2x_2 + 0.4x_3 + \exp(0.3x_1) + \sin(0.25x_2) + \cos(0.5x_3),\\
\mu(0, x, d) &= 0.25d + 0.7x_1 - 0.1x_2 - 0.3x_3 + (d-2)(-0.2x_1 + 0.2x_2 - 0.1x_3),
\end{aligned}$$

and $\sigma^2(x, d)$ is to be specified later.

For simplicity, we work under the simplifying condition that the treatment assignment probability is known in each source trial, which has no impact on the local semiparametric efficiency bound of the parameter θ. We considered four estimators with different weighting functions that share the form

$$\hat{\theta}_\bullet = \frac{1}{n}\sum_{k=1}^{5}\sum_{i\in\mathcal{I}_k} \ell_{\hat{\eta}_{k,\bullet}}(O_i),$$

where $\bullet \in \{\mathrm{oracle}, \mathrm{overlap}, \mathrm{constant}, \mathrm{optimal}\}$, the set of nuisance parameter estimates $\hat{\eta}_{k,\bullet}$ includes $\{\hat{w}_{k,\bullet}, \hat{\delta}_{k,\bullet}\}$, and the unnormalized weights are

$$\check{w}_{k,\mathrm{oracle}}(x, d) = \sigma^{-2}(x, d)e_0(1 \mid x, d)e_0(0 \mid x, d),$$

$$\check{w}_{k,\text{overlap}}(x,d) = e_0(1 \mid x,d)e_0(0 \mid x,d),$$
$$\check{w}_{k,\text{constant}}(x,d) = 1,$$

and finally $\check{w}_{k,\text{optimal}}$ corresponds to the strategy where the optimal weight is learned through empirical risk minimization (6). The weight function $\check{w}_{k,\text{oracle}}$ is the oracle optimal weight function and is used as the benchmark. The weight $\check{w}_{k,\text{overlap}}$ assumes homoscedasticity of the outcome across source trials, and $\check{w}_{k,\text{constant}}$ further ignores the difference between propensity scores.

To evaluate the performance of the estimators under different weight functions, we simulated data under three specifications of conditional variances of the outcome:

$$\sigma_1^2(x,d) = 1 + |d-2|, \qquad \text{(setting 1)}$$
$$\sigma_2^2(x,d) = 2[1 - 0.5|d-2| + \{0.2I(d=2) + 0.1I(d=3)\}(0.5x^{\mathrm{T}}1 + 1.5)]^{-1}, \qquad \text{(setting 2)}$$
$$\sigma_3^2(x,d) = 1, \qquad \text{(setting 3)}$$

corresponding to homoscedastic error within each trial, trial-specific covariate-dependent variance, and completely homoscedastic error, respectively. To investigate robustness against model misspecifications, we computed the estimators under four scenarios: all nuisance models are correctly specified (experiment 1), only $\mu(a,x,d)$ is misspecified (experiment 2), only $\zeta(d \mid x)$ and $\pi(x)$ are misspecified (experiment 3), and $\mu(a,x,d)$, $\zeta(d \mid x)$, and $\pi(x)$ are misspecified (experiment 4). All nuisance models except the propensity score were fitted with the super learner ensemble learning algorithm (van der Laan et al., 2007), using random forest, generalized additive model, lasso, and the null model as base learners. Model misspecifications were performed by replacing the original X with elementwise absolute values $|X|$. We simulated datasets of sizes $n \in \{1250, 2500\}$. For each sample size, we generated 1000 datasets. All estimators were obtained with 5-fold crossfitting. The standard errors were computed by plugging in the nuisance parameter estimates, including the weight functions.

For the target population ATE θ, summary statistics calculated from the simulation are displayed in Table 1 and Tables S1–S2 in the Supplementary Material. Under setting 1, the estimators have no Monte-Carlo bias except in experiment 4 when all nuisance parameter models are misspecified. Under experiment 1 of no model misspecification, the plug-in standard error estimates lead to the desired confidence interval coverage. Results from settings 2–3 show similar conclusions, except in setting 3 with sample size $n = 1250$. The anomaly is attributed to occasional volatile behavior of super learning with sample splitting, resulting in inflated plug-in standard error estimates. We refer to the summary statistics based on medians in Table S3. As expected in experiment 1, the estimators $\hat{\theta}_{\text{oracle}}$ and $\hat{\theta}_{\text{optimal}}$ have lower Monte-Carlo mean squared errors compared to the estimators $\hat{\theta}_{\text{overlap}}$ and $\hat{\theta}_{\text{constant}}$ in settings 1–2. No meaningful improvement of the standard error using $\hat{\theta}_{\text{oracle}}$ and $\hat{\theta}_{\text{optimal}}$ was observed in setting 3. Similar evidence was found in Table S4, where we calculated the proportion of plug-in standard error estimates of $\hat{\theta}_{\text{oracle}}$ and $\hat{\theta}_{\text{optimal}}$ being smaller than that of $\hat{\theta}_{\text{overlap}}$ and $\hat{\theta}_{\text{constant}}$. We highlight the importance of crossfitting for nominal coverage of the confidence interval. In Tables S5–S7, we show that in all settings, the plug-in standard error in experiment 1 underestimated the uncertainty of the estimator without crossfitting.

For comparison, we also computed the 5-fold crossfitted estimators $\hat{\theta}^{\ddagger}$ and $\hat{\theta}^{\dagger}$ that are asymptotically efficient under alternative probability models $\mathcal{P}^{\ddagger}$ and $\mathcal{P}^{\dagger}$. The nuisance

Table 1. *Summary of simulation results for $\hat{\theta}$ in setting 1 with crossfitting.*

n	Experiment	Estimator	Mean	Bias	RMSE	SE	Coverage
1250	1	$\hat{\theta}_{\text{oracle}}$	3.11	6.68	9.81	9.42	94.8
		$\hat{\theta}_{\text{overlap}}$	3.12	7.26	10.08	9.87	94.9
		$\hat{\theta}_{\text{constant}}$	3.12	7.30	10.06	9.85	94.9
		$\hat{\theta}_{\text{optimal}}$	3.11	6.54	10.04	9.60	94.6
	2	$\hat{\theta}_{\text{oracle}}$	3.09	−19.68	11.82	10.25	93.5
		$\hat{\theta}_{\text{overlap}}$	3.09	−21.12	12.79	10.54	93.7
		$\hat{\theta}_{\text{constant}}$	3.09	−20.89	12.67	10.51	93.6
		$\hat{\theta}_{\text{optimal}}$	3.09	−21.35	12.80	10.51	93.3
	3	$\hat{\theta}_{\text{oracle}}$	3.11	1.83	10.12	9.19	93.6
		$\hat{\theta}_{\text{overlap}}$	3.11	1.64	10.72	9.65	93.2
		$\hat{\theta}_{\text{constant}}$	3.11	1.80	10.66	9.62	92.8
		$\hat{\theta}_{\text{optimal}}$	3.11	1.59	10.57	9.37	93.6
	4	$\hat{\theta}_{\text{oracle}}$	2.96	−151.97	19.45	9.87	62.7
		$\hat{\theta}_{\text{overlap}}$	2.93	−172.82	22.06	10.16	56.1
		$\hat{\theta}_{\text{constant}}$	2.94	−170.43	21.75	10.12	57.0
		$\hat{\theta}_{\text{optimal}}$	2.95	−161.72	21.18	10.19	61.4
2500	1	$\hat{\theta}_{\text{oracle}}$	3.12	7.30	6.60	6.49	95.3
		$\hat{\theta}_{\text{overlap}}$	3.12	8.18	6.84	6.81	95.0
		$\hat{\theta}_{\text{constant}}$	3.12	8.21	6.83	6.80	94.9
		$\hat{\theta}_{\text{optimal}}$	3.11	7.14	6.63	6.54	95.4
	2	$\hat{\theta}_{\text{oracle}}$	3.10	−5.62	7.53	7.07	93.4
		$\hat{\theta}_{\text{overlap}}$	3.10	−6.18	7.61	7.23	93.3
		$\hat{\theta}_{\text{constant}}$	3.10	−5.96	7.60	7.21	93.3
		$\hat{\theta}_{\text{optimal}}$	3.10	−5.70	7.55	7.10	92.8
	3	$\hat{\theta}_{\text{oracle}}$	3.11	3.05	6.62	6.29	94.9
		$\hat{\theta}_{\text{overlap}}$	3.11	3.31	7.00	6.60	93.7
		$\hat{\theta}_{\text{constant}}$	3.11	3.44	6.97	6.58	93.7
		$\hat{\theta}_{\text{optimal}}$	3.11	2.95	6.66	6.34	94.8
	4	$\hat{\theta}_{\text{oracle}}$	2.95	−153.11	16.81	6.73	38.7
		$\hat{\theta}_{\text{overlap}}$	2.93	−174.87	18.85	6.84	29.1
		$\hat{\theta}_{\text{constant}}$	2.94	−172.32	18.61	6.83	29.9
		$\hat{\theta}_{\text{optimal}}$	2.95	−158.23	17.32	6.82	38.2

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table 2. *Summary of simulation results for $\hat{\theta}^{\dagger}$ and $\hat{\theta}^{\ddagger}$ in setting* 1 *with crossfitting.*

n	Estimator	Mean	Bias	RMSE	SE	Coverage
1250	$\hat{\theta}^{\dagger}$	−0.01	−3113.81	311.39	8.43	0.0
	$\hat{\theta}^{\ddagger}$	0.00	−3105.59	310.56	8.90	0.0
2500	$\hat{\theta}^{\dagger}$	−0.01	−3113.98	311.40	5.92	0.0
	$\hat{\theta}^{\ddagger}$	0.00	−3105.55	310.56	6.26	0.0

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

parameter $\mu(a, x)$ was estimated with super learning. The propensity score $e(a \mid x)$ appearing in $\hat{\theta}^{\ddagger}$ was estimated by $\hat{e}(a \mid x) = \sum_{d \in [m]} e(a \mid x, d)\hat{\zeta}(d \mid x)$. Specifically for $\hat{\theta}^{\dagger}$, the oracle was substituted for the weight function; that is, $w^{\dagger}(a, x, d) = \sigma^{-2}(x, d)$. Summary statistics for these estimators can be found in Table 2 and Tables S8–S9 in the Supplementary Material. The data generating mechanism here is compatible with the testable implication (1) of the CATE transportability Assumption 2(v) in the observed data distribution. However, stronger transportability assumptions discussed in §3.4 are violated, so that estimators $\hat{\theta}^{\dagger}$ and $\hat{\theta}^{\ddagger}$ suffer from severe bias in this setting.

For the target population projected CATE, we chose $Z = \{X_1\}$ and the cubic polynomial basis $\psi(x_1) = (1, x_1, x_1^2, x_1^3)^{\mathrm{T}}$. We computed the crossfitted least-squares estimator

$$\hat{\beta}_{\bullet} = (\hat{\beta}_{1,\bullet}, \hat{\beta}_{2,\bullet}, \hat{\beta}_{3,\bullet}, \hat{\beta}_{4,\bullet})^{\mathrm{T}}$$

with different weight functions. Summary statistics of individual $\hat{\beta}_{j,\bullet}$ can be found in Tables S10-S13 in the Supplementary Material, demonstrating the expected behaviors of these estimators. Again for sample size $n = 1250$ in setting 3, the behavior of the estimators under experiment 1 was surprising; see Table S14 for summary statistics based on medians. We used t-bootstrap to approximate the 95%-quantile of the Gaussian process supremum. In Table S15 in the Supplementary Material, we show that the 95%-uniform confidence interval for the projected CATE $\beta^{\mathrm{T}}\psi(x_1)$ over $x_1 \in [-1, 1]$ achieves the nominal coverage under experiment 1 with no model misspecification across all settings.

6. Real data example

STEP-1 (ClinicalTrials.gov ID NCT03548935, Wilding et al., 2021) is a multicenter randomized controlled trial comparing the weight loss effect of once-weekly semaglutide, 2.4 mg to placebo in addition to a lifestyle intervention. The main inclusion criteria of STEP-1 are body-mass index (BMI) of at least 30 (or at least 27 with weight-related co-morbidities) and no diabetes. As an illustration of our method, we regroup the trial data into four regions and take the study population recruited in the United States as the target population. The three source populations are defined by the subjects recruited from the European Union area (Belgium, Denmark, Finland, France, Germany, and Poland), the United Kingdom, and East Asia (Japan and Taiwan). We choose the target population based on the assumption that the demographic composition of the United States is the most varied in these four regions.

The outcome Y is the percentage change in body weight from baseline (week 0) to week 68. The target parameter is the ATE $\theta = E\{Y(1) - Y(0) \mid G = 1\}$ in the study

population of STEP-1 recruited in the United States ($G = 1$) of semaglutide ($A = 1$) versus placebo ($A = 0$). The analysis assumes cross-regional CATE transportability, conditioning on appropriate baseline characteristics. The baseline covariates X used in the standardization are body weight, age, sex, BMI, waist circumference, smoking status, diabetic history, and hemoglobin A1c. Note that the interpretation of the ATE defined on the original scale of body weight and that defined on the change percentage differ. However, the transportability assumption is the same for both scales, since the baseline body weight is included in X.

We perform a complete-case analysis where subjects with missing body weight measurements at week 68 are removed. The percentage of missing outcomes and the number of complete cases per region are reported in Table 3. All nuisance parameters except the propensity score are fitted with super learning, using gradient-boosted trees, random forest, a generalized additive model, lasso, and the null model as base learners. The propensity score within each source region is computed as the proportion of subjects receiving the active treatment. In particular, the outcome model is fitted on each combination of treatment and source region, such that the conditional means of the body weight change under the same treatment are allowed to vary freely across regions. We computed three estimators for the transported ATE using constant weights $\hat{\theta}_{\text{constant}}$, the overlap weights $\hat{\theta}_{\text{overlap}}$, and the learned optimal weights $\hat{\theta}_{\text{optimal}}$. All estimators were computed with 10-fold crossfitting, and plug-in standard errors were used for the construction of 95%-confidence intervals.

The results are displayed in Table 4. Since outcome and treatment information is directly available in the target population, we also report the augmented inverse probability weighting (AIPW) estimator $\hat{\theta}_{\text{AIPW}}$ based only on the target population for reference. The transported ATE estimator with learned optimal weights $\hat{\theta}_{\text{optimal}}$ gave a weight reduction of 12.57 percentage points from baseline. The estimates $\hat{\theta}_{\text{constant}}$ and $\hat{\theta}_{\text{overlap}}$ both indicated a reduction of 12.75 percentage points up to rounding. These two estimators are practically identical, since the overlap weights are nearly a constant in a well-implemented multi-site clinical trial. The plug-in standard error of $\hat{\theta}_{\text{optimal}}$ is slightly higher than that of $\hat{\theta}_{\text{overlap}}$ and $\hat{\theta}_{\text{constant}}$. However, this may be observed even in situations where variance reduction should be expected; see Table S4 in the Supplementary Material. The point estimates of the transported estimators roughly agree, and they are about 0.5 percentage point higher than the estimate from the AIPW estimator. Besides possible failure to account for all shifted effect modifiers, the discrepancy between the point estimates $\hat{\theta}_{\text{optimal}}$ and $\hat{\theta}_{\text{AIPW}}$ may result from the deletion of missing observations, since patients' adherence to treatment plans and dropout rates differ among the regions.

The exact effect estimates given by the crossfitted estimators depend on the number of splits. In the sensitivty analysis, we compute the same estimators for the target population ATE with 5-fold crossfitting, 2-fold crossfittting, and no crossfitting. The results are displayed in Table S16 of the Supplementary Material. The point estimates are mostly similar to the preceding results. The standard errors of the transported estimators increase as the number of crossfitting folds decreases. Notably, the plug-in standard errors of the transported estimators without crossfitting appear much lower than those obtained with crossfitting. On the other hand, the standard error of the AIPW estimator of barely changes. See §S1 in the Supplementary Material for further discussions.

To investigate treatment effect heterogeneity in the study population from the United

Table 3. *Percentage of missing outcomes and complete cases by region.*

	US	Europe	UK	Asia
Missing (%)	8.3	4.7	16.5	1.5
Complete (N)	700	367	182	133

Table 4. *Results for the target population ATE with* 10*-fold crossfitting.*

Estimator	Estimate	Standard error	95%-confidence interval
$\hat{\theta}_{\text{overlap}}$	−12.75	0.73	(−14.18, −11.32)
$\hat{\theta}_{\text{constant}}$	−12.75	0.73	(−14.18, −11.31)
$\hat{\theta}_{\text{optimal}}$	−12.57	0.75	(−14.03, −11.10)
$\hat{\theta}_{\text{AIPW}}$	−13.21	0.66	(−14.50, −11.92)

States, we calculated three target population projected CATEs onto the basis formed by baseline body weight, age, and BMI, respectively. In all cases, a polynomial basis of order 3 was applied. The nuisance parameters were estimated identically as in the estimation of the target population ATE, and only the estimated optimal weights were used. The results are plotted in Fig. 1. The percentage reduction in body weight tends to be smaller with a higher body weight at baseline, while there is no clear trend with respect to BMI. The treatment effect is less pronounced for elder individuals up to around 60 years old. We also observe a plateau of weight loss effect for those above 60, but the estimated projected CATE shows much statistical uncertainty. The discrepancy in sex is most striking, with females on average losing nearly 6.5 percentage points more body weight from semaglutide than males.

The outcome scale is paramount to the interpretation of effect heterogeneity. In the present example, the definition of the outcome, percentage change in body weight, explicitly involves the body weight at baseline. Fig. S1 in the Supplementary Material displays the target population projected CATE estimates using body weight at week 68 as the outcome. In contrast, the treatment effect under this outcome appears to increase with both baseline body weight and BMI. An informed treatment decision should preferably be based on multiple outcome scales of clinical value.

7. Discussion

In this work, we study efficient estimation of the target population ATE using multiple data sources. CATE transportability allows for identifiability of the target parameter but does not constrain the conditional counterfactual distributions nor the conditional treatment-specific means across the sources. However, if the outcome is bounded, for instance, binary or positive, the CATE transportability we assume induces potential variational dependence in the counterfactual distributions. This is mostly innocuous when the conditional effect size is small, but it may introduce undesired implicit assumptions when the effect is large. A less important technical difficulty lies in the study of the semiparametric efficiency bound when constraints on the model are not explicitly stated.

Certain conditional effect measures do not suffer from this problem. However, not all measures are suitable for transportability. As was pointed out by Colnet et al. (2023), the causal odds ratio fails to disentangle the risk under placebo, even under a complete

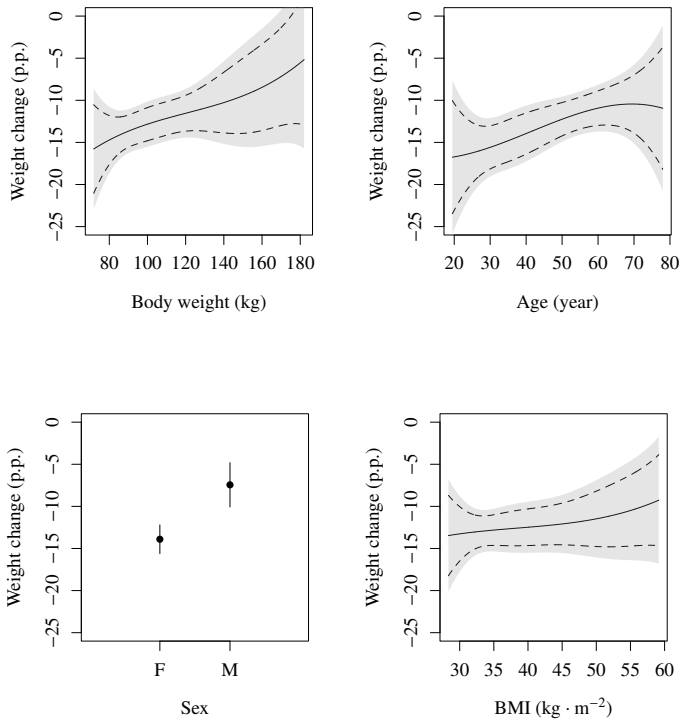

Figure 1. Target population projected CATE estimates onto baseline body weight, age, sex, and BMI, respectively. For body weight, age, and BMI, the solid lines are estimates of the projected CATE, while the pointwise and uniform 95%-confidence intervals are drawn with dashed lines and shadow. For sex, the point estimates and 95%-confidence intervals are displayed as solid dots and error bars. F: female; M: male.

lack of treatment effect heterogeneity. Moreover, noncollapsibility of the odds ratio leads to a mismatch between conditional and marginal effect measures. For positive outcomes, transportability on the ratio between conditional treatment-specific means presents a possibility. This assumption has been adopted for the estimation of target population causal mean ratio from a single source population (Wang et al., 2024a).

When estimating the CATE, one strategy is to decompose the outcome regression model as $\mu(a, x, d) = \mu(0, x, d) + a\delta(x)$ under the constraint in (1). For example, the nuisance parameters $\mu(0, x, d)$ and $\delta(x)$ can be jointly fitted to obtain an estimate of CATE. In practice, we may be inclined to use a meta-learner for the estimation of CATE. Shyr et al. (2024) proposed a multi-study R-learner precisely to leverage the overlap between the study populations. The samples are given the same weight in the multi-study R-learner, and it is not robust against the misspecification of nuisance models. Inspired by the efficient influence function in (5), we give an intuitive construction for the multi-study DR-learner by constructing pseudo-outcomes $\ell_{\hat{\eta}}(O_i)$ for observations O_i with $G_i = 0$. Regressing $\ell_{\hat{\eta}}(O_i)$ on X_i with $G_i = 0$ produces a fitted CATE. We leave the investigation of finite sample performance of the multi-study DR-learner for future work. An alternative line of work is higher-order meta-learners for CATE combining ideas from the minimax rate results from Kennedy et al. (2024).

Conflict of interest

Zehao Su is funded by a research gift from Novo Nordisk A/S to the Section of Biostatistics, University of Copenhagen. Henrik Ravn is employed by Novo Nordisk A/S.

References

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. (2018). Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework. *The Annals of Statistics*, 46(6B):3643–3675.

Benkeser, D. and van der Laan, M. (2016). The Highly Adaptive Lasso Estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 689–696.

Chen, R., Chen, G., and Yu, M. (2023). A generalizability score for aggregate causal effect. *Biostatistics*, 24(2):309–326.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691.

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

Colnet, B., Josse, J., Varoquaux, G., and Scornet, E. (2023). Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize? *arXiv*: 2303.16008v2.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2024). Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1):165–191.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. (2023). Estimating heterogeneous

treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211.

Dahabreh, I. J., Robertson, S. E., Petito, L. C., Hernán, M. A., and Steingrimsson, J. A. (2023). Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics*, 79(2):1057–1072.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernán, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39(14):1999–2014.

Fan, J. and Yao, Q. (1998). Efficient Estimation of Conditional Variance Functions in Stochastic Regression. *Biometrika*, 85(3):645–660.

Hines, O., Diaz-Ordaz, K., and Vansteelandt, S. (2024). Optimally weighted average derivative effects. *arXiv*: 2308.05456v2.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Josey, K. P., Berkowitz, S. A., Ghosh, D., and Raghavan, S. (2021). Transporting experimental results with entropy balancing. *Statistics in Medicine*, 40(19):4310–4326.

Kennedy, E. H., Balakrishnan, S., Robins, J. M., and Wasserman, L. (2024). Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793–816.

Lee, D., Yang, S., and Wang, X. (2022). Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *Journal of Causal Inference*, 10(1):415–440.

Li, F., Hong, H., and Stuart, E. A. (2023a). A note on semiparametric efficient generalization of causal effects from randomized trials to target populations. *Communications in Statistics - Theory and Methods*, 52(16):5767–5798.

Li, S. and Luedtke, A. (2023). Efficient estimation under data fusion. *Biometrika*, 110(4):1041–1054.

Li, X., Miao, W., Lu, F., and Zhou, X.-H. (2023b). Improving efficiency of inference in clinical trials with external control data. *Biometrics*, 79(1):394–403.

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics*, 17(4):363–374.

Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289.

Shyr, C., Ren, B., Patil, P., and Parmigiani, G. (2024). Multi-study R-learner for estimating heterogeneous treatment effects across studies using statistical machine learning. *arXiv*: 2306.01086v3.

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

van der Vaart, A. W. and Wellner, J. A. (2023). *Weak convergence and empirical processes: With applications to statistics*. Springer Series in Statistics. Springer.

VanderWeele, T. J. and Robins, J. M. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, 18(5):561–568.

Wang, G., Levis, A., Steingrimsson, J., and Dahabreh, I. (2024a). Causal inference under transportability assumptions for conditional relative effect measures. *arXiv*:2402.02702v2.

Wang, G., Levis, A., Steingrimsson, J., and Dahabreh, I. (2024b). Efficient estimation of subgroup treatment effects using multi-source data. *arXiv*:2402.02684v1.

Wilding, J. P., Batterham, R. L., Calanna, S., Davies, M., Van Gaal, L. F., Lingvay, I., McGowan, B. M., Rosenstock, J., Tran, M. T., Wadden, T. A., Wharton, S., Yokote, K., Zeuthen, N., and Kushner, R. F. (2021). Once-weekly semaglutide in adults with overweight or obesity. *New England Journal of Medicine*, 384(11):989–1002.

Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted learning: Causal inference for observational and experimental data*, Springer Series in Statistics, pages 459–474. Springer.

Supplementary material for "Efficient estimation of target-population average treatment effect from multi-source data"

Zehao Su
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
zehao.su@sund.ku.dk

Helene C. Rytgaard
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
hely@sund.ku.dk

Henrik Ravn
Novo Nordisk A/S,
Vandtårnsvej 108, DK-2860 Søborg, Denmark
hnrv@novonordisk.com

Frank Eriksson
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
eriksson@sund.ku.dk

S1. Further details on the simulated and the real data example

We used the implementation of super learning from R-package SuperLearner (Polley et al., 2024). The base learners and their corresponding names in SuperLearner are the null model (SL.mean), lasso (SL.glmnet), generalized additive model (SL.gam), random forest (SL.ranger), and gradient boosting trees (SL.xgboost). Since SuperLearner does not support categorical outcome with more than two levels, to estimate η, we fitted ensemble models $\check{\eta}_k$ treating the binary indicator $I(D = d)$ as outcome for each $d \in [m]$. To ensure the estimated probabilities summed up to 1, we used the normalized average

$$\hat{\eta}_k(d \mid x) = \frac{\check{\eta}_k(d \mid x)}{\sum_{d' \in [m]} \check{\eta}_k(d' \mid x)}.$$

The conditional means $\hat{\mu}_k$ were fitted separately for each combination of $a \in \{0, 1\}$ and $d \in [m]$.

For the uniform confidence interval, we used t-bootstrap (Belloni et al., 2015) to approximate the distribution of the Gaussian process supremum $T = \sup_{z \in \mathcal{Z}} |\mathbb{T}(z)|$. We calculated the plug-in estimate $\hat{\Omega}^\beta$ of the covariance matrix Ω^β. In the simulation, the support of $Z = \{X_1\}$ is the compact set $\mathcal{Z}_1 = [-1, 1]$. We used an equidistant grid $\tilde{\mathcal{Z}}_1$ of size 1000 to capture $\mathcal{Z}_1$. Let $\{\xi_1, \ldots, \xi_B\}$ be $B = 1000$ i.i.d. copies of a 4-dimensional normal random variable with mean zero and identity covariance matrix. The dimension of the random noise should match the dimension of the the cubic basis $\psi(z) = (1, x_1, x_1, x_1^3)^\mathrm{T}$. Then we simulated the Gaussian supremum T by

$$T_b = \max_{z \in \tilde{\mathcal{Z}}_1} \left| \frac{\psi^\mathrm{T}(z)(\hat{\Omega}^\beta)^{1/2}}{\{\psi^\mathrm{T}(z)\hat{\Omega}^\beta\psi(z)\}^{1/2}} \xi_b \right|$$

for $b \in [B]$, and the theoretical quantile $c_T(0.95)$ was approximated by the 95%-empirical quantile of the sample $\{T_1, \ldots, T_B\}$.

For the estimation of the projected CATE in the real data example, we formed a cubic basis of baseline body weight, age, and BMI after scaling these variables with $1/100$, $1/50$, and $1/50$, respectively. The inverse of the sample covariance matrix using the scaled variables was more stable numerically compared to the inverse using raw variables. The projected CATE and the associated confidence intervals were calculated and presented on the original scale.

In the simulation study, we saw that the plug-in standard error for the non-crossfitted estimators underestimates their variability. We hypothesized that using nonparametric base learners in super learning for certain nuisance parameters undermines consistency of the standard error estimator. To investigate the impact of crossfitting on the plug-in standard error estimates of the target population ATE in the real data example, we changed the base learners in super learning for different combinations of nuisance parameters. Thus, we re-calculated the transported ATE estimators replacing the full set of base learners above with only the null model and a generalized linear model (SL.glm) for $\{\mu\}$, $\{\eta\}$, $\{\pi\}$, $\{\mu, \eta\}$, $\{\mu, \pi\}$, $\{\eta, \pi\}$, and $\{\mu, \eta, \pi\}$ in turn, keeping the rest of the others unchanged. However, if the true nuisance parameters are not generalized linear models, we should not expect the estimators to be asymptotically normal. Nevertheless, the difference between the standard error estimates obtained with and without crossfitting may still inform which nuisance parameters are responsible for under-estimating the standard error.

The results are displayed in Table S17. As commented in the main text, the weight functions in the estimators $\hat{\theta}_{\text{constant}}$ and $\hat{\theta}_{\text{overlap}}$ do not change with the source population. Hence, the standard errors for these two estimators do not depend on model specification for η. This partly explains why the standard errors for the non-crossfitted estimators are particularly optimistic when we used generalized linear models for $\{\eta\}$. We are not able to single out any nuisance parameters that can explain the under-estimation of the standard error. We notice a recurring pattern that the estimator $\hat{\theta}_{\text{optimal}}$ has just slightly higher plug-in standard errors than the other two estimators. This observation suggests that the learned optimal weights might have been constant across the source populations.

S2. Proofs

S2.1. *Proof of Lemma 1*

We start with the g-formula representation. We see that for all $x \in \mathcal{X}_1$ and $d \in \mathcal{D}_x$

$$
\begin{aligned}
&E\{Y(1) - Y(0) \mid X = x, G = 1\} \\
&\quad = E\{Y(1) - Y(0) \mid X = x, D = d\} \\
&\quad = E\{Y(1) \mid A = 1, X = x, D = d\} - E\{Y(0) \mid A = 0, X = x, D = d\} \\
&\quad = E(Y \mid A = 1, X = x, D = d) - E(Y \mid A = 0, X = x, D = d) \\
&\quad = \mu(1, x, d) - \mu(0, x, d) \\
&\quad = \delta(x).
\end{aligned}
$$

The target parameter

$$
\begin{aligned}
\theta = E\{Y(1) - Y(0) \mid G = 1\} &= E[E\{Y(1) - Y(0) \mid G = 1, X\} \mid G = 1] \quad \text{(Iterated expectation)} \\
&= E\{\delta(X) \mid G = 1\}.
\end{aligned}
$$

Below we first show an inverse probability weighting representation via the identification formula above. For any $h(x, d)$ such that $E\{h(X, D) \mid X, G = 0\} = 1$, we have

$$
\frac{1}{\alpha} E\left\{ \frac{(1 - G)\pi(X)}{1 - \pi(X)} h(X, D) \frac{2A - 1}{e(A, X, D)} Y \right\}
$$

Table S1. *Summary of simulation results for $\hat{\theta}$ in setting 2 with crossfitting.*

n	Experiment	Estimator	Mean	Bias	RMSE	SE	Coverage
1250	1	$\hat{\theta}_{\text{oracle}}$	3.11	5.40	11.91	11.41	95.1
		$\hat{\theta}_{\text{overlap}}$	3.11	5.38	12.24	12.18	95.0
		$\hat{\theta}_{\text{constant}}$	3.11	5.46	12.21	12.13	95.1
		$\hat{\theta}_{\text{optimal}}$	3.11	5.91	12.35	11.66	94.4
	2	$\hat{\theta}_{\text{oracle}}$	3.09	−18.98	13.33	12.05	93.9
		$\hat{\theta}_{\text{overlap}}$	3.09	−21.00	15.30	12.76	94.4
		$\hat{\theta}_{\text{constant}}$	3.09	−20.79	15.13	12.69	94.2
		$\hat{\theta}_{\text{optimal}}$	3.09	−21.18	14.22	12.34	93.9
	3	$\hat{\theta}_{\text{oracle}}$	3.11	−2.53	12.05	11.36	94.2
		$\hat{\theta}_{\text{overlap}}$	3.10	−5.98	13.01	12.09	93.7
		$\hat{\theta}_{\text{constant}}$	3.10	−5.66	12.93	12.04	93.6
		$\hat{\theta}_{\text{optimal}}$	3.10	−2.85	12.59	11.55	93.8
	4	$\hat{\theta}_{\text{oracle}}$	2.97	−138.70	19.50	11.93	76.3
		$\hat{\theta}_{\text{overlap}}$	2.94	−172.23	24.30	12.62	68.1
		$\hat{\theta}_{\text{constant}}$	2.94	−169.86	23.95	12.55	68.7
		$\hat{\theta}_{\text{optimal}}$	2.96	−150.93	21.47	12.24	73.4
2500	1	$\hat{\theta}_{\text{oracle}}$	3.12	7.70	7.99	7.86	95.1
		$\hat{\theta}_{\text{overlap}}$	3.12	8.58	8.42	8.43	94.8
		$\hat{\theta}_{\text{constant}}$	3.12	8.62	8.40	8.39	94.8
		$\hat{\theta}_{\text{optimal}}$	3.12	7.85	8.03	7.94	95.0
	2	$\hat{\theta}_{\text{oracle}}$	3.10	−4.18	8.77	8.33	94.0
		$\hat{\theta}_{\text{overlap}}$	3.10	−5.75	9.07	8.75	94.0
		$\hat{\theta}_{\text{constant}}$	3.10	−5.52	9.04	8.71	94.0
		$\hat{\theta}_{\text{optimal}}$	3.10	−4.73	8.81	8.41	93.8
	3	$\hat{\theta}_{\text{oracle}}$	3.11	2.25	8.02	7.83	95.2
		$\hat{\theta}_{\text{overlap}}$	3.11	1.12	8.88	8.36	93.5
		$\hat{\theta}_{\text{constant}}$	3.11	1.32	8.82	8.33	93.6
		$\hat{\theta}_{\text{optimal}}$	3.11	2.34	8.10	7.89	95.0
	4	$\hat{\theta}_{\text{oracle}}$	2.97	−137.98	16.16	8.20	60.8
		$\hat{\theta}_{\text{overlap}}$	2.93	−174.52	19.54	8.54	47.9
		$\hat{\theta}_{\text{constant}}$	2.94	−171.97	19.30	8.51	49.0
		$\hat{\theta}_{\text{optimal}}$	2.96	−147.72	17.06	8.29	57.5

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S2. *Summary of simulation results for $\hat{\theta}$ in setting 3 with crossfitting.*

n	Experiment	Estimator	Mean	Bias	RMSE	SE	Coverage
1250	1	$\hat{\theta}_{\text{oracle}}$	3.22	114.51	329.51	16.17	94.7
		$\hat{\theta}_{\text{overlap}}$	3.22	114.51	329.51	16.17	94.7
		$\hat{\theta}_{\text{constant}}$	3.22	112.07	321.81	15.99	94.5
		$\hat{\theta}_{\text{optimal}}$	3.20	95.09	271.40	14.85	94.6
	2	$\hat{\theta}_{\text{oracle}}$	3.09	−20.98	10.66	8.85	93.3
		$\hat{\theta}_{\text{overlap}}$	3.09	−20.98	10.66	8.85	93.3
		$\hat{\theta}_{\text{constant}}$	3.09	−20.75	10.59	8.83	93.3
		$\hat{\theta}_{\text{optimal}}$	3.09	−21.19	11.31	9.07	92.9
	3	$\hat{\theta}_{\text{oracle}}$	3.18	76.55	224.03	13.28	93.7
		$\hat{\theta}_{\text{overlap}}$	3.18	76.55	224.03	13.28	93.7
		$\hat{\theta}_{\text{constant}}$	3.18	74.58	217.57	13.12	94.0
		$\hat{\theta}_{\text{optimal}}$	3.17	60.85	177.07	12.23	93.1
	4	$\hat{\theta}_{\text{oracle}}$	2.93	−172.89	20.38	8.39	43.5
		$\hat{\theta}_{\text{overlap}}$	2.93	−172.89	20.38	8.39	43.5
		$\hat{\theta}_{\text{constant}}$	2.94	−170.50	20.11	8.38	44.4
		$\hat{\theta}_{\text{optimal}}$	2.94	−169.39	20.59	8.73	46.5
2500	1	$\hat{\theta}_{\text{oracle}}$	3.11	7.14	5.65	5.56	95.0
		$\hat{\theta}_{\text{overlap}}$	3.11	7.14	5.65	5.56	95.0
		$\hat{\theta}_{\text{constant}}$	3.11	7.16	5.65	5.56	94.7
		$\hat{\theta}_{\text{optimal}}$	3.11	7.15	5.67	5.61	94.9
	2	$\hat{\theta}_{\text{oracle}}$	3.10	−6.97	6.55	6.09	91.9
		$\hat{\theta}_{\text{overlap}}$	3.10	−6.97	6.55	6.09	91.9
		$\hat{\theta}_{\text{constant}}$	3.10	−6.76	6.55	6.08	92.1
		$\hat{\theta}_{\text{optimal}}$	3.10	−6.03	6.55	6.11	92.0
	3	$\hat{\theta}_{\text{oracle}}$	3.11	3.77	5.69	5.37	94.2
		$\hat{\theta}_{\text{overlap}}$	3.11	3.77	5.69	5.37	94.2
		$\hat{\theta}_{\text{constant}}$	3.11	3.85	5.68	5.37	94.3
		$\hat{\theta}_{\text{optimal}}$	3.11	3.78	5.72	5.41	94.6
	4	$\hat{\theta}_{\text{oracle}}$	2.93	−175.59	18.51	5.68	13.3
		$\hat{\theta}_{\text{overlap}}$	2.93	−175.59	18.51	5.68	13.3
		$\hat{\theta}_{\text{constant}}$	2.93	−173.06	18.27	5.68	14.1
		$\hat{\theta}_{\text{optimal}}$	2.95	−162.82	17.35	5.84	22.6

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S3. *Extra summary of simulation results for $\hat{\theta}$ in setting 3 with 5-fold crossfitting and sample size $n = 1250$.*

Experiment	Estimator	Median	Bias	MAD	SE
1	$\hat{\theta}_{\text{oracle}}$	3.11	4.32	5.40	7.96
	$\hat{\theta}_{\text{overlap}}$	3.11	4.32	5.40	7.96
	$\hat{\theta}_{\text{constant}}$	3.11	4.39	5.35	7.96
	$\hat{\theta}_{\text{optimal}}$	3.11	2.32	5.36	8.07
2	$\hat{\theta}_{\text{oracle}}$	3.08	−26.55	6.21	8.63
	$\hat{\theta}_{\text{overlap}}$	3.08	−26.55	6.21	8.63
	$\hat{\theta}_{\text{constant}}$	3.08	−26.37	6.20	8.61
	$\hat{\theta}_{\text{optimal}}$	3.08	−28.52	6.13	8.80
3	$\hat{\theta}_{\text{oracle}}$	3.11	1.58	5.48	7.71
	$\hat{\theta}_{\text{overlap}}$	3.11	1.58	5.48	7.71
	$\hat{\theta}_{\text{constant}}$	3.11	1.63	5.41	7.71
	$\hat{\theta}_{\text{optimal}}$	3.11	2.41	5.50	7.82
4	$\hat{\theta}_{\text{oracle}}$	2.93	−176.63	5.69	8.13
	$\hat{\theta}_{\text{overlap}}$	2.93	−176.63	5.69	8.13
	$\hat{\theta}_{\text{constant}}$	2.93	−173.60	5.73	8.13
	$\hat{\theta}_{\text{optimal}}$	2.93	−173.08	5.71	8.41

Median: median of estimates; Bias: Monte-Carlo bias between median of estimates and truth, 10^{-3}; MAD: median average deviation, 10^{-2}; SE: median of standard error estimates, 10^{-2}.

Table S4. *Comparison of plug-in standard error estimates in simulations.*

		$\text{SE}_{\text{oracle}}$		$\text{SE}_{\text{optimal}}$	
Setting	n	$\text{SE}_{\text{constant}}$	$\text{SE}_{\text{overlap}}$	$\text{SE}_{\text{constant}}$	$\text{SE}_{\text{overlap}}$
1	1250	99.2	99.2	90.3	91.3
	2500	100.0	100.0	99.3	99.5
2	1250	99.7	99.8	95.8	96.8
	2500	100.0	100.0	99.8	99.8
3	1250	61.2	100.0	10.2	9.6
	2500	63.2	100.0	9.1	7.8

SE: standard error. Values represent the proportions (in percentages) of the upper standard error being smaller than the lower standard error, such as $\text{SE}_{\text{oracle}} \leqslant \text{SE}_{\text{constant}}$.

Table S5. *Summary of simulation results for $\hat{\theta}$ in setting 1 without crossfitting.*

n	Experiment	Estimator	Mean	Bias	RMSE	SE	Coverage
1250	1	$\hat{\theta}_{\text{oracle}}$	3.10	−10.97	27.51	8.68	92.0
		$\hat{\theta}_{\text{overlap}}$	3.09	−13.50	35.75	9.19	92.3
		$\hat{\theta}_{\text{constant}}$	3.09	−13.61	36.30	9.19	92.4
		$\hat{\theta}_{\text{optimal}}$	3.10	−11.04	27.03	8.38	90.0
	2	$\hat{\theta}_{\text{oracle}}$	3.07	−42.75	11.28	8.84	87.1
		$\hat{\theta}_{\text{overlap}}$	3.06	−46.01	11.60	9.05	88.2
		$\hat{\theta}_{\text{constant}}$	3.06	−45.48	11.56	9.04	88.3
		$\hat{\theta}_{\text{optimal}}$	3.06	−50.58	11.81	8.43	84.9
	3	$\hat{\theta}_{\text{oracle}}$	3.09	−15.89	32.74	8.61	90.2
		$\hat{\theta}_{\text{overlap}}$	3.09	−20.26	44.07	9.15	90.7
		$\hat{\theta}_{\text{constant}}$	3.09	−20.26	44.58	9.15	90.7
		$\hat{\theta}_{\text{optimal}}$	3.09	−16.13	32.90	8.31	87.8
	4	$\hat{\theta}_{\text{oracle}}$	2.95	−162.37	19.01	8.49	50.6
		$\hat{\theta}_{\text{overlap}}$	2.92	−184.06	20.97	8.65	43.9
		$\hat{\theta}_{\text{constant}}$	2.93	−181.59	20.74	8.64	44.7
		$\hat{\theta}_{\text{optimal}}$	2.94	−171.11	19.94	8.11	42.9
2500	1	$\hat{\theta}_{\text{oracle}}$	3.11	2.49	6.47	5.95	92.5
		$\hat{\theta}_{\text{overlap}}$	3.11	3.54	6.71	6.23	92.9
		$\hat{\theta}_{\text{constant}}$	3.11	3.57	6.71	6.22	92.9
		$\hat{\theta}_{\text{optimal}}$	3.11	2.84	6.48	5.81	92.2
	2	$\hat{\theta}_{\text{oracle}}$	3.09	−21.40	7.71	6.40	88.5
		$\hat{\theta}_{\text{overlap}}$	3.08	−23.09	7.86	6.55	89.2
		$\hat{\theta}_{\text{constant}}$	3.09	−22.68	7.83	6.53	89.1
		$\hat{\theta}_{\text{optimal}}$	3.08	−26.12	7.79	6.16	86.5
	3	$\hat{\theta}_{\text{oracle}}$	3.11	−0.85	6.57	5.80	92.5
		$\hat{\theta}_{\text{overlap}}$	3.11	−0.35	6.97	6.07	90.5
		$\hat{\theta}_{\text{constant}}$	3.11	−0.24	6.94	6.06	90.4
		$\hat{\theta}_{\text{optimal}}$	3.11	−0.44	6.59	5.65	91.2
	4	$\hat{\theta}_{\text{oracle}}$	2.95	−157.11	17.14	6.12	29.7
		$\hat{\theta}_{\text{overlap}}$	2.93	−178.71	19.17	6.22	21.5
		$\hat{\theta}_{\text{constant}}$	2.93	−176.20	18.93	6.21	22.2
		$\hat{\theta}_{\text{optimal}}$	2.95	−160.81	17.51	5.93	28.7

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S6. *Summary of simulation results for $\hat{\theta}$ in setting 2 without crossfitting.*

n	Experiment	Estimator	Mean	Bias	RMSE	SE	Coverage
1250	1	$\hat{\theta}_{\text{oracle}}$	3.09	−16.74	37.31	10.44	91.2
		$\hat{\theta}_{\text{overlap}}$	3.09	−19.89	49.04	11.29	91.6
		$\hat{\theta}_{\text{constant}}$	3.09	−20.07	49.81	11.27	91.4
		$\hat{\theta}_{\text{optimal}}$	3.09	−15.27	34.75	10.05	89.0
	2	$\hat{\theta}_{\text{oracle}}$	3.07	−40.56	12.74	10.43	89.3
		$\hat{\theta}_{\text{overlap}}$	3.06	−45.21	13.16	10.92	89.6
		$\hat{\theta}_{\text{constant}}$	3.06	−44.72	13.11	10.89	89.4
		$\hat{\theta}_{\text{optimal}}$	3.06	−46.18	13.07	9.97	86.3
	3	$\hat{\theta}_{\text{oracle}}$	3.08	−24.54	44.40	10.57	91.1
		$\hat{\theta}_{\text{overlap}}$	3.08	−31.77	60.63	11.47	89.9
		$\hat{\theta}_{\text{constant}}$	3.08	−31.74	61.34	11.44	89.8
		$\hat{\theta}_{\text{optimal}}$	3.08	−23.93	44.24	10.17	88.3
	4	$\hat{\theta}_{\text{oracle}}$	2.96	−149.38	19.05	10.33	65.8
		$\hat{\theta}_{\text{overlap}}$	2.92	−184.23	22.09	10.73	58.2
		$\hat{\theta}_{\text{constant}}$	2.93	−181.77	21.86	10.70	59.0
		$\hat{\theta}_{\text{optimal}}$	2.95	−159.20	19.99	9.83	60.2
2500	1	$\hat{\theta}_{\text{oracle}}$	3.11	2.95	7.83	7.15	92.7
		$\hat{\theta}_{\text{overlap}}$	3.11	4.05	8.26	7.65	92.8
		$\hat{\theta}_{\text{constant}}$	3.11	4.09	8.24	7.62	92.9
		$\hat{\theta}_{\text{optimal}}$	3.11	3.58	7.85	6.98	91.8
	2	$\hat{\theta}_{\text{oracle}}$	3.09	−18.85	8.84	7.54	90.2
		$\hat{\theta}_{\text{overlap}}$	3.09	−21.70	9.22	7.91	90.8
		$\hat{\theta}_{\text{constant}}$	3.09	−21.31	9.18	7.88	90.8
		$\hat{\theta}_{\text{optimal}}$	3.09	−22.03	8.87	7.30	88.7
	3	$\hat{\theta}_{\text{oracle}}$	3.11	−1.58	7.93	7.16	92.4
		$\hat{\theta}_{\text{overlap}}$	3.11	−2.29	8.81	7.62	91.4
		$\hat{\theta}_{\text{constant}}$	3.11	−2.11	8.75	7.59	91.2
		$\hat{\theta}_{\text{optimal}}$	3.11	−0.95	8.00	6.95	91.5
	4	$\hat{\theta}_{\text{oracle}}$	2.97	−142.22	16.47	7.46	52.5
		$\hat{\theta}_{\text{overlap}}$	2.93	−178.14	19.81	7.74	38.3
		$\hat{\theta}_{\text{constant}}$	2.93	−175.62	19.57	7.71	39.1
		$\hat{\theta}_{\text{optimal}}$	2.96	−149.93	17.16	7.20	46.8

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S7. *Summary of simulation results for $\hat{\theta}$ in setting 3 without crossfitting.*

n	Experiment	Estimator	Mean	Bias	RMSE	SE	Coverage
1250	1	$\hat{\theta}_{\text{oracle}}$	3.10	−10.02	27.62	7.61	92.1
		$\hat{\theta}_{\text{overlap}}$	3.10	−10.02	27.62	7.61	92.1
		$\hat{\theta}_{\text{constant}}$	3.10	−10.09	28.05	7.62	92.0
		$\hat{\theta}_{\text{optimal}}$	3.10	−8.31	22.03	7.26	91.0
	2	$\hat{\theta}_{\text{oracle}}$	3.06	−47.15	10.79	7.64	85.1
		$\hat{\theta}_{\text{overlap}}$	3.06	−47.15	10.79	7.64	85.1
		$\hat{\theta}_{\text{constant}}$	3.06	−46.59	10.72	7.64	85.0
		$\hat{\theta}_{\text{optimal}}$	3.05	−55.80	11.98	7.27	81.7
	3	$\hat{\theta}_{\text{oracle}}$	3.09	−14.07	34.02	7.57	90.8
		$\hat{\theta}_{\text{overlap}}$	3.09	−14.07	34.02	7.57	90.8
		$\hat{\theta}_{\text{constant}}$	3.09	−14.09	34.42	7.58	90.7
		$\hat{\theta}_{\text{optimal}}$	3.10	−12.39	28.50	7.22	89.0
	4	$\hat{\theta}_{\text{oracle}}$	2.92	−184.26	20.45	7.21	30.5
		$\hat{\theta}_{\text{overlap}}$	2.92	−184.26	20.45	7.21	30.5
		$\hat{\theta}_{\text{constant}}$	2.93	−181.80	20.21	7.21	31.5
		$\hat{\theta}_{\text{optimal}}$	2.93	−180.39	20.56	6.96	30.0
2500	1	$\hat{\theta}_{\text{oracle}}$	3.11	2.58	5.53	5.14	93.0
		$\hat{\theta}_{\text{overlap}}$	3.11	2.58	5.53	5.14	93.0
		$\hat{\theta}_{\text{constant}}$	3.11	2.59	5.54	5.14	93.1
		$\hat{\theta}_{\text{optimal}}$	3.11	2.71	5.53	5.01	92.3
	2	$\hat{\theta}_{\text{oracle}}$	3.08	−24.59	6.92	5.53	87.0
		$\hat{\theta}_{\text{overlap}}$	3.08	−24.59	6.92	5.53	87.0
		$\hat{\theta}_{\text{constant}}$	3.08	−24.18	6.90	5.52	87.0
		$\hat{\theta}_{\text{optimal}}$	3.08	−29.26	7.05	5.29	85.3
	3	$\hat{\theta}_{\text{oracle}}$	3.11	0.16	5.64	4.98	91.7
		$\hat{\theta}_{\text{overlap}}$	3.11	0.16	5.64	4.98	91.7
		$\hat{\theta}_{\text{constant}}$	3.11	0.21	5.63	4.98	91.7
		$\hat{\theta}_{\text{optimal}}$	3.11	0.43	5.65	4.86	90.9
	4	$\hat{\theta}_{\text{oracle}}$	2.93	−179.60	18.87	5.17	8.4
		$\hat{\theta}_{\text{overlap}}$	2.93	−179.60	18.87	5.17	8.4
		$\hat{\theta}_{\text{constant}}$	2.93	−177.09	18.63	5.18	9.6
		$\hat{\theta}_{\text{optimal}}$	2.94	−166.13	17.63	5.07	13.3

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S8. *Summary of simulation results for $\hat{\theta}^{\dagger}$ and $\hat{\theta}^{\ddagger}$ in setting* 2 *with crossfitting.*

n	Estimator	Mean	Bias	RMSE	SE	Coverage
1250	$\hat{\theta}^{\dagger}$	−0.00	−3111.88	311.22	10.47	0.0
	$\hat{\theta}^{\ddagger}$	0.00	−3105.80	310.59	11.25	0.0
2500	$\hat{\theta}^{\dagger}$	−0.00	−3111.75	311.19	7.32	0.0
	$\hat{\theta}^{\ddagger}$	0.00	−3105.38	310.54	7.89	0.0

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S9. *Summary of simulation results for* $\hat{\theta}^{\dagger}$ *and* $\hat{\theta}^{\ddagger}$ *in setting* 3 *with crossfitting.*

n	Estimator	Mean	Bias	RMSE	SE	Coverage
1250	$\hat{\theta}^{\dagger}$	0.00	−3105.60	310.56	7.06	0.0
	$\hat{\theta}^{\ddagger}$	0.00	−3105.60	310.56	7.06	0.0
2500	$\hat{\theta}^{\dagger}$	0.00	−3106.01	310.60	4.97	0.0
	$\hat{\theta}^{\ddagger}$	0.00	−3106.01	310.60	4.97	0.0

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

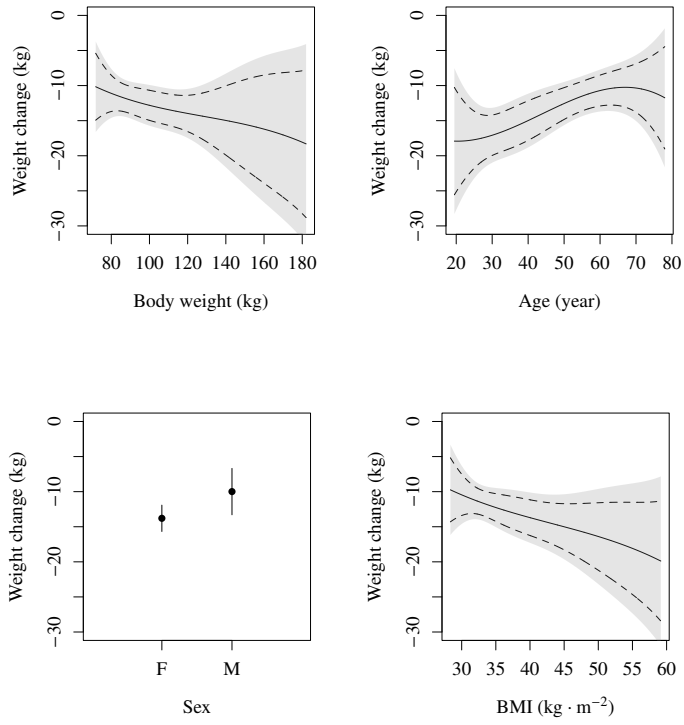

Figure S1. Body weight loss at week 68 conditional on baseline body weight, age, sex, and BMI, respectively. For body weight, age, and BMI, the solid lines are estimates of the projected CATE, while the pointwise and uniform 95%-confidence intervals are drawn with dashed lines and shadow. For sex, the point estimates and 95%-confidence intervals are displayed as solid dots and error bars. F: female; M: male.

Table S10. *Summary of simulation results for $\hat{\beta}_1$ in experiment 1.*

Crossfitting	Setting	n	Estimator	Mean	Bias	RMSE	SE	Coverage
5-fold	1	1250	$\hat{\beta}_{1,\text{oracle}}$	3.00	6.70	14.43	13.19	94.6
			$\hat{\beta}_{1,\text{overlap}}$	3.00	5.39	14.52	13.86	94.0
			$\hat{\beta}_{1,\text{constant}}$	3.00	5.48	14.51	13.83	94.0
			$\hat{\beta}_{1,\text{optimal}}$	3.00	5.82	14.74	13.45	93.9
		2500	$\hat{\beta}_{1,\text{oracle}}$	3.01	10.26	8.61	8.96	95.2
			$\hat{\beta}_{1,\text{overlap}}$	3.01	10.85	9.17	9.46	95.8
			$\hat{\beta}_{1,\text{constant}}$	3.01	10.93	9.15	9.44	95.9
			$\hat{\beta}_{1,\text{optimal}}$	3.00	10.03	8.65	9.05	96.2
	2	1250	$\hat{\beta}_{1,\text{oracle}}$	3.00	8.60	18.40	16.52	94.4
			$\hat{\beta}_{1,\text{overlap}}$	3.00	6.94	18.32	17.69	94.3
			$\hat{\beta}_{1,\text{constant}}$	3.00	7.05	18.28	17.61	94.1
			$\hat{\beta}_{1,\text{optimal}}$	3.00	8.75	19.38	16.91	94.3
		2500	$\hat{\beta}_{1,\text{oracle}}$	3.01	11.76	10.80	11.25	95.3
			$\hat{\beta}_{1,\text{overlap}}$	3.01	11.82	11.72	12.13	96.3
			$\hat{\beta}_{1,\text{constant}}$	3.01	11.92	11.68	12.08	96.1
			$\hat{\beta}_{1,\text{optimal}}$	3.01	11.75	10.85	11.38	95.5
	3	1250	$\hat{\beta}_{1,\text{oracle}}$	2.76	−232.51	756.44	27.07	94.0
			$\hat{\beta}_{1,\text{overlap}}$	2.76	−232.51	756.44	27.07	94.0
			$\hat{\beta}_{1,\text{constant}}$	2.77	−226.90	738.80	26.71	94.0
			$\hat{\beta}_{1,\text{optimal}}$	2.80	−191.26	623.07	24.46	93.6
		2500	$\hat{\beta}_{1,\text{oracle}}$	3.00	9.40	7.22	7.47	95.5
			$\hat{\beta}_{1,\text{overlap}}$	3.00	9.40	7.22	7.47	95.5
			$\hat{\beta}_{1,\text{constant}}$	3.00	9.45	7.23	7.47	95.6
			$\hat{\beta}_{1,\text{optimal}}$	3.00	9.02	7.25	7.54	95.9
None	1	1250	$\hat{\beta}_{1,\text{oracle}}$	2.99	−2.49	16.65	11.73	92.5
			$\hat{\beta}_{1,\text{overlap}}$	2.99	−1.19	17.33	12.46	92.1
			$\hat{\beta}_{1,\text{constant}}$	2.99	−1.11	17.46	12.45	91.9
			$\hat{\beta}_{1,\text{optimal}}$	2.99	−2.29	16.32	11.24	90.8
		2500	$\hat{\beta}_{1,\text{oracle}}$	3.00	6.38	8.37	8.13	94.3
			$\hat{\beta}_{1,\text{overlap}}$	3.00	7.09	8.92	8.57	94.5
			$\hat{\beta}_{1,\text{constant}}$	3.00	7.15	8.90	8.55	94.6
			$\hat{\beta}_{1,\text{optimal}}$	3.00	6.57	8.34	7.91	93.4
	2	1250	$\hat{\beta}_{1,\text{oracle}}$	3.00	0.65	19.23	14.64	92.3
			$\hat{\beta}_{1,\text{overlap}}$	3.00	1.25	22.12	15.90	92.5
			$\hat{\beta}_{1,\text{constant}}$	3.00	1.40	22.25	15.86	92.4
			$\hat{\beta}_{1,\text{optimal}}$	2.99	−1.72	15.98	13.92	90.9
		2500	$\hat{\beta}_{1,\text{oracle}}$	3.00	7.83	10.50	10.17	94.0
			$\hat{\beta}_{1,\text{overlap}}$	3.00	8.27	11.38	10.94	94.2
			$\hat{\beta}_{1,\text{constant}}$	3.00	8.34	11.34	10.90	94.5
			$\hat{\beta}_{1,\text{optimal}}$	3.00	8.46	10.48	9.89	93.2
	3	1250	$\hat{\beta}_{1,\text{oracle}}$	2.99	−1.38	13.96	9.92	91.7
			$\hat{\beta}_{1,\text{overlap}}$	2.99	−1.38	13.96	9.92	91.7
			$\hat{\beta}_{1,\text{constant}}$	2.99	−1.32	14.08	9.93	91.8
			$\hat{\beta}_{1,\text{optimal}}$	2.99	−2.09	12.67	9.39	91.1
		2500	$\hat{\beta}_{1,\text{oracle}}$	3.00	5.74	7.01	6.79	94.1
			$\hat{\beta}_{1,\text{overlap}}$	3.00	5.74	7.01	6.79	94.1
			$\hat{\beta}_{1,\text{constant}}$	3.00	5.78	7.01	6.79	94.2
			$\hat{\beta}_{1,\text{optimal}}$	3.00	5.92	6.97	6.59	93.2

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S11. *Summary of simulation results for $\hat{\beta}_2$ in experiment 1.*

Crossfitting	Setting	n	Estimator	Mean	Bias	RMSE	SE	Coverage
5-fold	1	1250	$\hat{\beta}_{2,\text{oracle}}$	0.95	15.96	40.87	38.90	95.0
			$\hat{\beta}_{2,\text{overlap}}$	0.96	21.22	41.93	40.96	95.0
			$\hat{\beta}_{2,\text{constant}}$	0.96	20.78	41.86	40.88	95.0
			$\hat{\beta}_{2,\text{optimal}}$	0.96	19.37	41.83	39.76	95.0
		2500	$\hat{\beta}_{2,\text{oracle}}$	0.94	−0.27	26.20	26.33	94.2
			$\hat{\beta}_{2,\text{overlap}}$	0.94	0.22	27.89	27.77	94.6
			$\hat{\beta}_{2,\text{constant}}$	0.94	0.35	27.83	27.72	94.6
			$\hat{\beta}_{2,\text{optimal}}$	0.94	0.38	26.33	26.59	94.8
	2	1250	$\hat{\beta}_{2,\text{oracle}}$	0.95	12.40	50.63	48.12	95.3
			$\hat{\beta}_{2,\text{overlap}}$	0.96	20.01	51.78	51.57	94.8
			$\hat{\beta}_{2,\text{constant}}$	0.96	19.50	51.63	51.35	94.8
			$\hat{\beta}_{2,\text{optimal}}$	0.95	15.94	52.29	49.25	94.9
		2500	$\hat{\beta}_{2,\text{oracle}}$	0.94	−1.36	32.55	32.77	94.0
			$\hat{\beta}_{2,\text{overlap}}$	0.94	−1.71	35.57	35.30	94.7
			$\hat{\beta}_{2,\text{constant}}$	0.94	−1.54	35.40	35.15	94.7
			$\hat{\beta}_{2,\text{optimal}}$	0.94	−0.14	32.86	33.14	94.6
	3	1250	$\hat{\beta}_{2,\text{oracle}}$	1.81	869.87	2727.18	90.97	94.7
			$\hat{\beta}_{2,\text{overlap}}$	1.81	869.87	2727.18	90.97	94.7
			$\hat{\beta}_{2,\text{constant}}$	1.79	849.83	2663.63	89.67	94.7
			$\hat{\beta}_{2,\text{optimal}}$	1.66	724.37	2246.26	81.29	95.8
		2500	$\hat{\beta}_{2,\text{oracle}}$	0.94	0.27	21.98	21.96	94.4
			$\hat{\beta}_{2,\text{overlap}}$	0.94	0.27	21.98	21.96	94.4
			$\hat{\beta}_{2,\text{constant}}$	0.94	0.35	21.97	21.96	94.5
			$\hat{\beta}_{2,\text{optimal}}$	0.94	1.24	22.08	22.17	94.3
None	1	1250	$\hat{\beta}_{2,\text{oracle}}$	0.96	17.90	52.41	35.42	93.4
			$\hat{\beta}_{2,\text{overlap}}$	0.95	14.58	47.95	37.76	93.7
			$\hat{\beta}_{2,\text{constant}}$	0.95	14.37	48.29	37.76	93.6
			$\hat{\beta}_{2,\text{optimal}}$	0.96	19.67	48.91	33.85	92.7
		2500	$\hat{\beta}_{2,\text{oracle}}$	0.94	−1.26	24.92	23.81	93.2
			$\hat{\beta}_{2,\text{overlap}}$	0.94	−0.15	26.56	25.10	93.3
			$\hat{\beta}_{2,\text{constant}}$	0.94	−0.02	26.50	25.06	93.2
			$\hat{\beta}_{2,\text{optimal}}$	0.94	0.01	24.63	23.14	93.1
	2	1250	$\hat{\beta}_{2,\text{oracle}}$	0.95	7.47	48.76	43.82	93.5
			$\hat{\beta}_{2,\text{overlap}}$	0.95	9.68	55.11	47.83	93.5
			$\hat{\beta}_{2,\text{constant}}$	0.95	9.19	55.27	47.74	93.4
			$\hat{\beta}_{2,\text{optimal}}$	0.96	24.82	47.47	41.59	92.8
		2500	$\hat{\beta}_{2,\text{oracle}}$	0.94	−1.64	30.92	29.53	93.3
			$\hat{\beta}_{2,\text{overlap}}$	0.94	−1.22	33.85	31.75	93.2
			$\hat{\beta}_{2,\text{constant}}$	0.94	−1.06	33.70	31.63	93.2
			$\hat{\beta}_{2,\text{optimal}}$	0.94	−1.17	30.60	28.69	93.1
	3	1250	$\hat{\beta}_{2,\text{oracle}}$	0.95	12.43	40.58	30.17	93.7
			$\hat{\beta}_{2,\text{overlap}}$	0.95	12.43	40.58	30.17	93.7
			$\hat{\beta}_{2,\text{constant}}$	0.95	12.31	40.96	30.23	93.7
			$\hat{\beta}_{2,\text{optimal}}$	0.95	14.33	36.64	28.31	92.8
		2500	$\hat{\beta}_{2,\text{oracle}}$	0.94	−0.57	21.00	19.93	93.4
			$\hat{\beta}_{2,\text{overlap}}$	0.94	−0.57	21.00	19.93	93.4
			$\hat{\beta}_{2,\text{constant}}$	0.94	−0.49	20.99	19.93	93.4
			$\hat{\beta}_{2,\text{optimal}}$	0.94	0.75	20.68	19.33	93.0

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S12. *Summary of simulation results for $\hat{\beta}_3$ in experiment 1.*

Crossfitting	Setting	n	Estimator	Mean	Bias	RMSE	SE	Coverage
5-fold	1	1250	$\hat{\beta}_{3,\text{oracle}}$	0.01	−2.67	35.10	30.72	94.2
			$\hat{\beta}_{3,\text{overlap}}$	0.01	2.80	38.04	32.47	94.3
			$\hat{\beta}_{3,\text{constant}}$	0.01	2.65	37.88	32.40	94.3
			$\hat{\beta}_{3,\text{optimal}}$	0.01	−0.61	37.68	31.57	94.1
		2500	$\hat{\beta}_{3,\text{oracle}}$	−0.00	−12.90	19.67	20.33	95.4
			$\hat{\beta}_{3,\text{overlap}}$	−0.00	−12.22	20.79	21.48	95.9
			$\hat{\beta}_{3,\text{constant}}$	−0.00	−12.35	20.75	21.43	95.7
			$\hat{\beta}_{3,\text{optimal}}$	−0.00	−12.65	19.83	20.55	95.7
	2	1250	$\hat{\beta}_{3,\text{oracle}}$	−0.00	−13.51	40.12	38.22	94.7
			$\hat{\beta}_{3,\text{overlap}}$	0.00	−9.41	41.61	40.95	94.9
			$\hat{\beta}_{3,\text{constant}}$	0.00	−9.47	41.48	40.78	94.9
			$\hat{\beta}_{3,\text{optimal}}$	−0.00	−13.10	41.79	39.24	94.0
		2500	$\hat{\beta}_{3,\text{oracle}}$	−0.00	−16.36	24.85	25.65	95.5
			$\hat{\beta}_{3,\text{overlap}}$	−0.00	−14.15	26.61	27.61	95.8
			$\hat{\beta}_{3,\text{constant}}$	−0.00	−14.33	26.51	27.50	95.7
			$\hat{\beta}_{3,\text{optimal}}$	−0.00	−16.00	24.99	25.98	96.3
	3	1250	$\hat{\beta}_{3,\text{oracle}}$	1.37	1360.26	4266.54	116.66	94.1
			$\hat{\beta}_{3,\text{overlap}}$	1.37	1360.26	4266.54	116.66	94.1
			$\hat{\beta}_{3,\text{constant}}$	1.34	1328.58	4166.95	114.62	94.2
			$\hat{\beta}_{3,\text{optimal}}$	1.13	1120.81	3513.90	101.29	94.1
		2500	$\hat{\beta}_{3,\text{oracle}}$	0.00	−10.78	16.37	16.91	95.6
			$\hat{\beta}_{3,\text{overlap}}$	0.00	−10.78	16.37	16.91	95.6
			$\hat{\beta}_{3,\text{constant}}$	0.00	−10.89	16.38	16.91	95.7
			$\hat{\beta}_{3,\text{optimal}}$	0.00	−9.66	16.45	17.09	95.3
None	1	1250	$\hat{\beta}_{3,\text{oracle}}$	−0.02	−29.09	122.85	28.06	92.5
			$\hat{\beta}_{3,\text{overlap}}$	−0.03	−42.86	154.62	29.99	92.3
			$\hat{\beta}_{3,\text{constant}}$	−0.03	−43.52	157.16	30.00	92.4
			$\hat{\beta}_{3,\text{optimal}}$	−0.02	−30.14	120.16	26.78	91.7
		2500	$\hat{\beta}_{3,\text{oracle}}$	−0.00	−13.93	18.64	18.30	94.8
			$\hat{\beta}_{3,\text{overlap}}$	−0.00	−13.27	19.65	19.30	94.8
			$\hat{\beta}_{3,\text{constant}}$	−0.00	−13.36	19.61	19.27	94.8
			$\hat{\beta}_{3,\text{optimal}}$	−0.00	−13.41	18.45	17.75	94.3
	2	1250	$\hat{\beta}_{3,\text{oracle}}$	−0.05	−58.44	160.20	34.99	92.5
			$\hat{\beta}_{3,\text{overlap}}$	−0.06	−72.11	210.91	38.31	92.7
			$\hat{\beta}_{3,\text{constant}}$	−0.06	−73.22	214.39	38.26	92.6
			$\hat{\beta}_{3,\text{optimal}}$	−0.03	−45.74	125.73	32.73	91.6
		2500	$\hat{\beta}_{3,\text{oracle}}$	−0.00	−16.95	23.54	22.99	94.7
			$\hat{\beta}_{3,\text{overlap}}$	−0.00	−15.39	25.13	24.70	94.3
			$\hat{\beta}_{3,\text{constant}}$	−0.00	−15.50	25.04	24.60	94.4
			$\hat{\beta}_{3,\text{optimal}}$	−0.01	−17.03	23.37	22.28	93.8
	3	1250	$\hat{\beta}_{3,\text{oracle}}$	−0.02	−29.88	119.87	23.88	92.8
			$\hat{\beta}_{3,\text{overlap}}$	−0.02	−29.88	119.87	23.88	92.8
			$\hat{\beta}_{3,\text{constant}}$	−0.02	−30.34	121.85	23.93	92.8
			$\hat{\beta}_{3,\text{optimal}}$	−0.01	−22.01	93.61	22.27	91.2
		2500	$\hat{\beta}_{3,\text{oracle}}$	0.00	−11.90	15.50	15.26	94.8
			$\hat{\beta}_{3,\text{overlap}}$	0.00	−11.90	15.50	15.26	94.8
			$\hat{\beta}_{3,\text{constant}}$	−0.00	−11.97	15.51	15.26	95.0
			$\hat{\beta}_{3,\text{optimal}}$	0.00	−11.95	15.33	14.78	94.2

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S13. *Summary of simulation results for $\hat{\beta}_4$ in experiment 1.*

Crossfitting	Setting	n	Estimator	Mean	Bias	RMSE	SE	Coverage
5-fold	1	1250	$\hat{\beta}_{4,\text{oracle}}$	0.11	−18.93	67.05	61.31	95.7
			$\hat{\beta}_{4,\text{overlap}}$	0.10	−25.60	70.97	64.73	95.9
			$\hat{\beta}_{4,\text{constant}}$	0.10	−24.83	70.68	64.60	95.8
			$\hat{\beta}_{4,\text{optimal}}$	0.10	−23.87	71.00	62.91	95.4
		2500	$\hat{\beta}_{4,\text{oracle}}$	0.13	5.39	40.92	40.77	94.5
			$\hat{\beta}_{4,\text{overlap}}$	0.13	5.38	43.16	43.03	95.1
			$\hat{\beta}_{4,\text{constant}}$	0.13	5.12	43.06	42.95	95.1
			$\hat{\beta}_{4,\text{optimal}}$	0.13	4.14	41.10	41.19	94.8
	2	1250	$\hat{\beta}_{4,\text{oracle}}$	0.12	−5.75	77.51	75.23	95.9
			$\hat{\beta}_{4,\text{overlap}}$	0.11	−13.73	79.78	80.58	95.8
			$\hat{\beta}_{4,\text{constant}}$	0.11	−12.95	79.51	80.25	95.8
			$\hat{\beta}_{4,\text{optimal}}$	0.12	−8.64	80.28	77.15	95.6
		2500	$\hat{\beta}_{4,\text{oracle}}$	0.13	7.59	50.81	50.71	94.1
			$\hat{\beta}_{4,\text{overlap}}$	0.13	9.50	54.96	54.62	94.8
			$\hat{\beta}_{4,\text{constant}}$	0.13	9.15	54.70	54.40	94.8
			$\hat{\beta}_{4,\text{optimal}}$	0.13	6.59	51.40	51.34	94.9
	3	1250	$\hat{\beta}_{4,\text{oracle}}$	−2.41	−2536.78	7955.51	219.02	96.0
			$\hat{\beta}_{4,\text{overlap}}$	−2.41	−2536.78	7955.51	219.02	96.0
			$\hat{\beta}_{4,\text{constant}}$	−2.35	−2477.96	7769.95	215.28	95.9
			$\hat{\beta}_{4,\text{optimal}}$	−1.98	−2099.37	6552.01	190.95	95.7
		2500	$\hat{\beta}_{4,\text{oracle}}$	0.13	4.73	34.14	33.99	95.0
			$\hat{\beta}_{4,\text{overlap}}$	0.13	4.73	34.14	33.99	95.0
			$\hat{\beta}_{4,\text{constant}}$	0.13	4.56	34.13	34.00	95.1
			$\hat{\beta}_{4,\text{optimal}}$	0.13	3.47	34.36	34.35	94.6
None	1	1250	$\hat{\beta}_{4,\text{oracle}}$	0.12	−6.78	169.19	56.09	94.5
			$\hat{\beta}_{4,\text{overlap}}$	0.14	18.10	196.45	59.84	94.3
			$\hat{\beta}_{4,\text{constant}}$	0.14	19.17	199.60	59.87	94.4
			$\hat{\beta}_{4,\text{optimal}}$	0.12	−7.30	161.31	53.40	93.4
		2500	$\hat{\beta}_{4,\text{oracle}}$	0.12	−1.27	38.39	36.69	93.7
			$\hat{\beta}_{4,\text{overlap}}$	0.12	−1.42	40.54	38.71	93.9
			$\hat{\beta}_{4,\text{constant}}$	0.12	−1.63	40.45	38.64	93.9
			$\hat{\beta}_{4,\text{optimal}}$	0.12	−3.39	37.89	35.60	93.1
	2	1250	$\hat{\beta}_{4,\text{oracle}}$	0.16	35.46	197.78	69.03	94.3
			$\hat{\beta}_{4,\text{overlap}}$	0.18	52.05	259.32	75.54	94.7
			$\hat{\beta}_{4,\text{constant}}$	0.18	53.98	263.47	75.44	94.9
			$\hat{\beta}_{4,\text{optimal}}$	0.13	2.73	120.05	64.33	93.5
		2500	$\hat{\beta}_{4,\text{oracle}}$	0.12	−0.63	47.56	45.49	93.8
			$\hat{\beta}_{4,\text{overlap}}$	0.12	0.71	51.51	48.90	93.7
			$\hat{\beta}_{4,\text{constant}}$	0.12	0.45	51.28	48.72	93.8
			$\hat{\beta}_{4,\text{optimal}}$	0.12	−0.58	46.93	44.10	93.1
	3	1250	$\hat{\beta}_{4,\text{oracle}}$	0.13	5.51	154.66	47.83	94.1
			$\hat{\beta}_{4,\text{overlap}}$	0.13	5.51	154.66	47.83	94.1
			$\hat{\beta}_{4,\text{constant}}$	0.13	6.19	157.16	47.94	94.1
			$\hat{\beta}_{4,\text{optimal}}$	0.12	−3.39	121.97	44.52	93.7
		2500	$\hat{\beta}_{4,\text{oracle}}$	0.12	−1.59	32.22	30.70	94.2
			$\hat{\beta}_{4,\text{overlap}}$	0.12	−1.59	32.22	30.70	94.2
			$\hat{\beta}_{4,\text{constant}}$	0.12	−1.74	32.21	30.71	93.9
			$\hat{\beta}_{4,\text{optimal}}$	0.12	−3.91	31.64	29.73	93.4

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-3}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S14. *Extra summary of simulation results for $\hat{\beta}$ in setting 3 and experiment 1 with* 5*-fold crossfitting and sample size* $n = 1250$*.*

Estimator	Median	Bias	MAD	SE
$\hat{\beta}_{1,\text{oracle}}$	3.00	3.21	6.92	10.72
$\hat{\beta}_{1,\text{overlap}}$	3.00	3.21	6.92	10.72
$\hat{\beta}_{1,\text{constant}}$	3.00	3.72	6.88	10.72
$\hat{\beta}_{1,\text{optimal}}$	3.00	2.22	6.99	10.93
$\hat{\beta}_{2,\text{oracle}}$	0.95	15.51	22.02	31.61
$\hat{\beta}_{2,\text{overlap}}$	0.95	15.51	22.02	31.61
$\hat{\beta}_{2,\text{constant}}$	0.95	15.10	22.26	31.63
$\hat{\beta}_{2,\text{optimal}}$	0.96	21.56	21.85	32.20
$\hat{\beta}_{3,\text{oracle}}$	0.01	−6.11	16.15	24.64
$\hat{\beta}_{3,\text{overlap}}$	0.01	−6.11	16.15	24.64
$\hat{\beta}_{3,\text{constant}}$	0.01	−6.30	16.15	24.65
$\hat{\beta}_{3,\text{optimal}}$	0.01	−2.78	16.50	25.15
$\hat{\beta}_{4,\text{oracle}}$	0.11	−14.50	33.47	49.20
$\hat{\beta}_{4,\text{overlap}}$	0.11	−14.50	33.47	49.20
$\hat{\beta}_{4,\text{constant}}$	0.11	−12.23	33.58	49.20
$\hat{\beta}_{4,\text{optimal}}$	0.09	−36.00	33.73	50.30

Median: median of estimates; Bias: Monte-Carlo bias between median of estimates and truth, 10^{-3}; MAD: median average deviation, 10^{-2}; SE: median of standard error estimates, 10^{-2}.

Table S15. 95%*-uniform coverage of projected CATE in the simulation study.*

			Crossfitting	
Setting	n	Estimator	5-fold	None
1	1250	$\hat{\beta}_{\text{oracle}}$	94.3	88.4
		$\hat{\beta}_{\text{overlap}}$	93.1	88.8
		$\hat{\beta}_{\text{constant}}$	93.4	88.9
		$\hat{\beta}_{\text{optimal}}$	93.8	86.3
	2500	$\hat{\beta}_{\text{oracle}}$	94.8	91.7
		$\hat{\beta}_{\text{overlap}}$	94.6	91.4
		$\hat{\beta}_{\text{constant}}$	94.8	91.9
		$\hat{\beta}_{\text{optimal}}$	94.8	90.5
2	1250	$\hat{\beta}_{\text{oracle}}$	94.1	88.7
		$\hat{\beta}_{\text{overlap}}$	92.9	89.4
		$\hat{\beta}_{\text{constant}}$	93.4	88.7
		$\hat{\beta}_{\text{optimal}}$	94.2	85.1
	2500	$\hat{\beta}_{\text{oracle}}$	94.7	91.2
		$\hat{\beta}_{\text{overlap}}$	94.7	91.3
		$\hat{\beta}_{\text{constant}}$	94.9	91.3
		$\hat{\beta}_{\text{optimal}}$	95.4	89.5
3	1250	$\hat{\beta}_{\text{oracle}}$	93.5	89.3
		$\hat{\beta}_{\text{overlap}}$	93.5	89.0
		$\hat{\beta}_{\text{constant}}$	93.8	88.4
		$\hat{\beta}_{\text{optimal}}$	93.1	86.9
	2500	$\hat{\beta}_{\text{oracle}}$	94.4	92.1
		$\hat{\beta}_{\text{overlap}}$	94.4	91.9
		$\hat{\beta}_{\text{constant}}$	94.4	92.1
		$\hat{\beta}_{\text{optimal}}$	95.7	90.3

Table S16. *Results for the target population ATE with 5-fold crossfitting, 2-fold crossfitting, and without crossfitting.*

Crossfitting	Estimator	Estimate	Standard error	95%-confidence interval
5-fold	$\hat{\theta}_{\text{overlap}}$	−13.02	0.72	(−14.42, −11.62)
	$\hat{\theta}_{\text{constant}}$	−13.02	0.72	(−14.42, −11.61)
	$\hat{\theta}_{\text{optimal}}$	−12.93	0.76	(−14.41, −11.45)
	$\hat{\theta}_{\text{AIPW}}$	−13.18	0.66	(−14.47, −11.90)
2-fold	$\hat{\theta}_{\text{overlap}}$	−12.84	0.80	(−14.41, −11.26)
	$\hat{\theta}_{\text{constant}}$	−12.82	0.80	(−14.39, −11.24)
	$\hat{\theta}_{\text{optimal}}$	−12.69	0.89	(−14.44, −10.94)
	$\hat{\theta}_{\text{AIPW}}$	−13.19	0.66	(−14.48, −11.90)
None	$\hat{\theta}_{\text{overlap}}$	−13.02	0.37	(−13.75, −12.29)
	$\hat{\theta}_{\text{constant}}$	−13.00	0.37	(−13.74, −12.27)
	$\hat{\theta}_{\text{optimal}}$	−12.92	0.35	(−13.61, −12.24)
	$\hat{\theta}_{\text{AIPW}}$	−13.21	0.65	(−14.48, −11.94)

$$
\begin{aligned}
&= \frac{1}{\alpha} E\left\{ \frac{(1-G)\pi(X)}{1-\pi(X)} h(X,D) \frac{2A-1}{e(A,X,D)} \mu(A,X,D) \right\} \\
&= \frac{1}{\alpha} E\left[\frac{(1-G)\pi(X)}{1-\pi(X)} h(X,D) \{\mu(1,X,D) - \mu(0,X,D)\} \right] \\
&= \frac{1}{\alpha} E\left\{ \frac{(1-G)\pi(X)}{1-\pi(X)} h(X,D) \delta(X) \right\} \\
&= \frac{1}{\alpha} E\left\{ \frac{(1-G)\pi(X)}{1-\pi(X)} \delta(X) \right\} \\
&= \frac{1}{\alpha} E\{\pi(X)\delta(X)\} \\
&= E\{\delta(X) \mid G = 1\}.
\end{aligned}
$$

The alternative representation in the lemma is immediate.

S2.2. *Proof of Lemma 2*

Let $L_2^0(P_0)$ denote the space of mean-zero $L_2(P_0)$-functions. Consider the linear subspace of $L_2^0(P_0)$

$$\dot{\mathcal{P}} = \dot{\mathcal{P}}_y \oplus \dot{\mathcal{P}}_a \oplus \dot{\mathcal{P}}_d \oplus \dot{\mathcal{P}}_g \oplus \dot{\mathcal{P}}_x,$$

where $\Lambda_1 \oplus \Lambda_2$ denotes the direct sum of the spaces Λ_1 and Λ_2, and

$$
\begin{aligned}
\dot{\mathcal{P}}_y = \big\{ & (1-g)h(y,a,x,d) : E_0\{h(Y,A,X,D) \mid A,X,D\} = 0, \\
& E_0\{Yh(Y,A,X,D) \mid A=1,X,D=d\} - E_0\{Yh(Y,A,X,D) \mid A=0,X,D=d\} = \\
& E_0\{Yh(Y,A,X,D) \mid A=1,X,D=d'\} - E_0\{Yh(Y,A,X,D) \mid A=0,X,D=d'\} \big\}, \\
\dot{\mathcal{P}}_a = \big\{ & (1-g)h(a,x,d) : E_0\{h(A,X,D) \mid X,D\} = 0 \big\}, \\
\dot{\mathcal{P}}_d = \big\{ & (1-g)h(d,x) : E_0\{h(D,X) \mid X, G=0\} = 0 \big\}, \\
\dot{\mathcal{P}}_g = \big\{ & h(g,x) : E_0\{h(G,X) \mid X\} = 0 \big\}, \\
\dot{\mathcal{P}}_x = \big\{ & h(x) : E_0\{h(X)\} = 0 \big\}.
\end{aligned}
$$

The proof of Lemma 2 consists of two parts. In the first part (Lemma S1), we show the orthocomplement of the space $\dot{\mathcal{P}}$ and how to calculate the projection of functions in $\dot{\mathcal{P}}_y$ onto this space. In the second part (Lemma 2), we show that $\dot{\mathcal{P}}$ is exactly the tangent space of the model

Table S17. *Sensitivity analysis for the target population ATE with* 10*-fold crossfitting and without crossfitting.*

Model	Crossfitting	Estimator	Estimate	Standard error	95%-confidence interval
μ	10-fold	$\hat{\theta}_{\text{overlap}}$	−12.81	0.73	(−14.25, −11.38)
		$\hat{\theta}_{\text{constant}}$	−12.80	0.73	(−14.24, −11.37)
		$\hat{\theta}_{\text{optimal}}$	−12.73	0.74	(−14.18, −11.27)
	None	$\hat{\theta}_{\text{overlap}}$	−12.96	0.47	(−13.88, −12.04)
		$\hat{\theta}_{\text{constant}}$	−12.94	0.47	(−13.86, −12.02)
		$\hat{\theta}_{\text{optimal}}$	−12.94	0.45	(−13.81, −12.06)
η	10-fold	$\hat{\theta}_{\text{overlap}}$	−12.77	0.72	(−14.19, −11.35)
		$\hat{\theta}_{\text{constant}}$	−12.76	0.72	(−14.18, −11.34)
		$\hat{\theta}_{\text{optimal}}$	−12.70	0.75	(−14.17, −11.23)
	None	$\hat{\theta}_{\text{overlap}}$	−12.99	0.35	(−13.67, −12.30)
		$\hat{\theta}_{\text{constant}}$	−12.97	0.35	(−13.66, −12.28)
		$\hat{\theta}_{\text{optimal}}$	−12.82	0.34	(−13.49, −12.15)
π	10-fold	$\hat{\theta}_{\text{overlap}}$	−12.80	0.70	(−14.17, −11.43)
		$\hat{\theta}_{\text{constant}}$	−12.80	0.70	(−14.17, −11.42)
		$\hat{\theta}_{\text{optimal}}$	−12.72	0.72	(−14.12, −11.32)
	None	$\hat{\theta}_{\text{overlap}}$	−12.98	0.51	(−13.99, −11.98)
		$\hat{\theta}_{\text{constant}}$	−12.97	0.51	(−13.98, −11.96)
		$\hat{\theta}_{\text{optimal}}$	−12.85	0.49	(−13.82, −11.89)
μ, η	10-fold	$\hat{\theta}_{\text{overlap}}$	−12.76	0.72	(−14.18, −11.34)
		$\hat{\theta}_{\text{constant}}$	−12.75	0.73	(−14.17, −11.33)
		$\hat{\theta}_{\text{optimal}}$	−12.68	0.75	(−14.15, −11.22)
	None	$\hat{\theta}_{\text{overlap}}$	−12.96	0.46	(−13.87, −12.05)
		$\hat{\theta}_{\text{constant}}$	−12.94	0.46	(−13.85, −12.04)
		$\hat{\theta}_{\text{optimal}}$	−12.95	0.45	(−13.83, −12.07)
μ, π	10-fold	$\hat{\theta}_{\text{overlap}}$	−12.85	0.70	(−14.21, −11.48)
		$\hat{\theta}_{\text{constant}}$	−12.84	0.70	(−14.21, −11.47)
		$\hat{\theta}_{\text{optimal}}$	−12.80	0.71	(−14.19, −11.42)
	None	$\hat{\theta}_{\text{overlap}}$	−12.90	0.65	(−14.18, −11.61)
		$\hat{\theta}_{\text{constant}}$	−12.88	0.65	(−14.16, −11.60)
		$\hat{\theta}_{\text{optimal}}$	−12.94	0.62	(−14.16, −11.72)
η, π	10-fold	$\hat{\theta}_{\text{overlap}}$	−12.78	0.70	(−14.15, −11.42)
		$\hat{\theta}_{\text{constant}}$	−12.78	0.70	(−14.15, −11.41)
		$\hat{\theta}_{\text{optimal}}$	−12.81	0.73	(−14.25, −11.38)
	None	$\hat{\theta}_{\text{overlap}}$	−12.96	0.51	(−13.95, −11.96)
		$\hat{\theta}_{\text{constant}}$	−12.94	0.51	(−13.94, −11.94)
		$\hat{\theta}_{\text{optimal}}$	−12.88	0.48	(−13.82, −11.94)
μ, η, π	10-fold	$\hat{\theta}_{\text{overlap}}$	−12.82	0.70	(−14.18, −11.46)
		$\hat{\theta}_{\text{constant}}$	−12.81	0.70	(−14.18, −11.44)
		$\hat{\theta}_{\text{optimal}}$	−12.82	0.71	(−14.20, −11.44)
	None	$\hat{\theta}_{\text{overlap}}$	−12.88	0.65	(−14.16, −11.60)
		$\hat{\theta}_{\text{constant}}$	−12.87	0.66	(−14.15, −11.58)
		$\hat{\theta}_{\text{optimal}}$	−12.89	0.64	(−14.15, −11.63)

Model: the model(s) changed from using all base learners to using only the null model and a generalized linear model.

$\mathcal{P}$ at P_0 under the local regularity conditions (Assumption 3), and the construction of the tangent space depends on the projection results from Lemma S1.

Lemma S1. *The linear space*

$$\Lambda = \left\{(1-g)\frac{2a-1}{e_0(a \mid x,d)} q(x,d)\{y - \mu_0(a,x,d)\} : E_0\{q(X,D) \mid X, G=0\} = 0\right\}$$

is the orthocomplement of $\dot{\mathcal{P}}$ *in* $L_2^0(P_0)$. *Moreover, it is also the orthocomplement of* $\dot{\mathcal{P}}_y$ *in* $\tilde{\dot{\mathcal{P}}}_y = \{(1-g)h(y,a,x,d) : E_0\{h(Y,A,X,D) \mid A,X,D\} = 0\}$.

Proof. It is trivial to see that $\Lambda \subset \tilde{\dot{\mathcal{P}}}_y$. Hence, Λ is orthogonal to $\dot{\mathcal{P}}_\bullet$ for $\bullet \in \{a,d,g,x\}$, because $\tilde{\dot{\mathcal{P}}}_y$ is orthogonal to $\dot{\mathcal{P}}_\bullet$ by the decomposition of the Hilbert space $L_2^0(P_0)$. To prove the lemma, we need to show that any element in Λ is orthogonal to all elements in $\dot{\mathcal{P}}_y$. For any

$$\begin{aligned}
\lambda_q(y,a,x,d,g) &= (1-g)\frac{2a-1}{e_0(a \mid x,d)} q(x,d)\{y - \mu_0(a,x,d)\} \in \Lambda,\\
\ell_h(y,a,x,d,g) &= (1-g)h(y,a,x,d) \in \dot{\mathcal{P}}_y,
\end{aligned}$$

their inner product is

$$\begin{aligned}
&\langle \lambda_q(Y,A,X,D,G), \ell_h(Y,A,X,D,G)\rangle_{L_2(P_0)}\\
&\quad = E_0\left[(1-G)\frac{2A-1}{e_0(A \mid X,D)} q(X,D)\{Y - \mu_0(A,X,D)\}h(Y,A,X,D)\right]\\
&\quad = E_0\left[(1-G)\frac{2A-1}{e_0(A \mid X,D)} q(X,D) Y h(Y,A,X,D)\right]\\
&\quad = E_0\left[(1-G)\frac{2A-1}{e_0(A \mid X,D)} q(X,D) E_0\{Y h(Y,A,X,D) \mid A,X,D\}\right],
\end{aligned}$$

and if we integrate over A and D conditionally on $G=0$ and X inside the outermost expectation, the inner product is

$$= E_0\left[(1-G)\sum_{d\in[m]} \zeta_0(d \mid X) q(X,d)\kappa_h(X)\right],$$

where $\kappa_h(x) = E_0\{Yh(Y,A,X,D) \mid A=1, X=x, D=d\} - E_0\{Yh(Y,A,X,D) \mid A=0, X=x, D=d\}$ does not depend on the value of d, so

$$= E_0\big[(1-G)\kappa_h(X) E_0\{q(X,D) \mid X, G=0\}\big] = 0.$$

It follows that $\Lambda \perp \dot{\mathcal{P}}$, or equivalently stated, $\Lambda \subset \dot{\mathcal{P}}^\perp$.

Every element $\tilde{\ell}_h = (1-g)h(y,a,x,d) \in \tilde{\dot{\mathcal{P}}}_y$ can be written as

$$\tilde{\ell}_h = \Pi\{\tilde{\ell}_h \mid \Lambda\} + \Pi\{\tilde{\ell}_h \mid \Lambda^\perp\}.$$

Since the space $L_2^0(P_0)$ decomposes as the direct sum

$$\tilde{\dot{\mathcal{P}}}_y \oplus \dot{\mathcal{P}}_a \oplus \dot{\mathcal{P}}_d \oplus \dot{\mathcal{P}}_g \oplus \dot{\mathcal{P}}_x,$$

to prove the lemma, it remains to show that $\Pi\{\tilde{\ell}_h \mid \Lambda^\perp\} \in \dot{\mathcal{P}}_y$.

To this end, we proceed to derive the projected space $\Pi\{\tilde{\dot{\mathcal{P}}}_y \mid \Lambda^\perp\}$. Before doing so, consider the larger linear subspace

$$\tilde{\Lambda} = \left\{(1-g)\frac{2a-1}{e_0(a \mid x,d)} r(x,d)\{y - \mu_0(a,x,d)\} : r(x,d) \text{ arbitrary}\right\} \subset L_2^0(P_0).$$

For any function $r(x,d)$, the projection of

$$\tilde{\lambda}_r(y,a,x,d,g) = (1-g)\frac{2a-1}{e_0(a\,|\,x,d)}r(x,d)\{y-\mu_0(a,x,d)\} \in \tilde{\Lambda}$$

onto the subspace Λ is

$$\Pi\{\tilde{\lambda}_r \,|\, \Lambda\} = (1-g)\frac{2a-1}{e_0(a\,|\,x,d)}\left[r(x,d) - w_0(x,d)\frac{E_0\{r(X,D)\,|\,X=x,G=0\}}{E_0\{w_0(X,D)\,|\,X=x,G=0\}}\right] \{y-\mu_0(a,x,d)\}.$$

The legitimacy of the projection can be established by checking for every $\lambda_q \in \Lambda$, the inner product

$$\begin{aligned}
&\langle \tilde{\lambda}_r - \Pi\{\tilde{\lambda}_r \,|\, \Lambda\}, \lambda_q(Y,A,X,D,G)\rangle_{L_2(P_0)} \\
&\quad = E_0\left[(1-G)\frac{\{Y-\mu_0(A,X,D)\}^2}{\{e_0(A\,|\,X,D)\}^2}w_0(X,D)\frac{E_0\{r(X,D)\,|\,X,G=0\}}{E_0\{w_0(X,D)\,|\,X,G=0\}}q(X,D)\right] \\
&\quad = E_0\left[(1-G)\frac{V_0(A,X,D)}{\{e_0(A\,|\,X,D)\}^2}w_0(X,D)\frac{E_0\{r(X,D)\,|\,X,G=0\}}{E_0\{w_0(X,D)\,|\,X,G=0\}}q(X,D)\right] \\
&\quad = E_0\left[(1-G)\frac{E_0\{r(X,D)\,|\,X,G=0\}}{E_0\{w_0(X,D)\,|\,X,G=0\}}q(X,D)\right] \\
&\quad = E_0\left[(1-G)\frac{E_0\{r(X,D)\,|\,X,G=0\}}{E_0\{w_0(X,D)\,|\,X,G=0\}}E_0\{q(X,D)\,|\,X,G=0\}\right] \\
&\quad = 0,
\end{aligned}$$

and that indeed

$$E_0\left[r(X,D) - w_0(X,D)\frac{E_0\{r(X,D)\,|\,X,G=0\}}{E_0\{w_0(X,D)\,|\,X,G=0\}}\;\middle|\; X, G=0\right] = 0.$$

Take an arbitrary element $\tilde{\ell}_h \in \tilde{\bar{\mathcal{P}}}_y$. Then suppose the projection onto $\tilde{\Lambda}$ is

$$\Pi\{\tilde{\ell}_h \,|\, \tilde{\Lambda}\} = (1-g)\frac{2a-1}{e_0(a\,|\,x,d)}r_h(x,d)\{y-\mu_0(a,x,d)\},$$

so that $r_h(x,d)$ fulfills the equation

$$E_0\Bigg(\frac{2A-1}{e_0(A\,|\,X,D)}\{Y-\mu_0(A,X,D)\} \left[h(Y,A,X,D) - \frac{2A-1}{e_0(A\,|\,X,D)}r_h(X,D)\{Y-\mu_0(A,X,D)\}\right]\Bigg|\, X, D\Bigg) = 0.$$

Direct calculation yields the solution

$$\begin{aligned}
r_h(x,d) &= w_0(x,d)\big[E_0\{Yh(Y,A,X,D)\,|\,A=1,X=x,D=d\} \\
&\qquad - E_0\{Yh(Y,A,X,D)\,|\,A=0,X=x,D=d\}\big] \\
&= w_0(x,d)\{\mathrm{cov}_0(Y,h\,|\,A=1,X=x,D=d) - \mathrm{cov}_0(Y,h\,|\,A=0,X=x,D=d)\}.
\end{aligned}$$

Then further projecting $\Pi\{\ell_h \,|\, \tilde{\Lambda}\}$ onto Λ, we determine that

$$\Pi\{\ell_h \,|\, \Lambda\} = (1-g)\frac{2a-1}{e_0(a\,|\,x,d)}q_h(x,d)\{y-\mu_0(a,x,d)\},$$

where

$$q_h(x,d) = r_h(x,d) - w_0(x,d)\frac{E_0\{r_h(X,D) \mid X=x, G=0\}}{E_0\{w_0(X,D) \mid X=x, G=0\}}.$$

In the following, we verify that

$$\Pi\{\tilde{\ell}_h \mid \Lambda^\perp\} = (1-g)\left[h(y,a,x,d) - \frac{2a-1}{e_0(a \mid x,d)} q_h(x,d)\{y - \mu_0(a,x,d)\}\right],$$

where

$$s_h(y,a,x,d) = h(y,a,x,d) - \frac{2a-1}{e_0(a \mid x,d)} q_h(x,d)\{y - \mu_0(a,x,d)\},$$

is indeed an element of $\dot{\mathcal{P}}_y$. It is trivial that $E_0\{s_h(Y,A,X,D) \mid A,X,D\} = 0$. Furthermore,

$$\begin{aligned}
&E_0\{Ys_h \mid A=1, X=x, D=d\} - E_0\{Ys_h \mid A=0, X=x, D=d\} \\
&\quad = E_0\{Yh(Y,A,X,D) \mid A=1, X=x, D=d\} - E_0\{Yh(Y,A,X,D) \mid A=0, X=x, D=d\} \\
&\qquad - q_h(x,d) \sum_{a\in\{0,1\}} \frac{E_0[Y\{Y - \mu_0(a,X,D)\} \mid A=a, X=x, D=d]}{e_0(a \mid x,d)} \\
&\quad = E_0\{Yh(Y,A,X,D) \mid A=1, X=x, D=d\} - E_0\{Yh(Y,A,X,D) \mid A=0, X=x, D=d\} \\
&\qquad - \frac{q_h(x,d)}{w_0(x,d)} \\
&\quad = \frac{E_0\{r_h(X,D) \mid X=x, G=0\}}{E_0\{w_0(X,D) \mid X=x, G=0\}}
\end{aligned}$$

is constant in the level of d. This ascertains that $\Pi\{\tilde{\ell}_h \mid \Lambda^\perp\} \in \dot{\mathcal{P}}_y$, and the proof is complete □

Proof of Lemma 2. The observed data distribution is

$$p_0(y,a,x,d,g) = \left[p_0(y \mid a,x,d)e_0(a \mid x,d)\zeta_0(d \mid x)\{1-\pi_0(x)\}\right]^{(1-g)}\{\pi_0(x)\}^g p_0(x).$$

Describing the structure of the maximal tangent space. We claim that the tangent space of the model $\mathcal{P}$ at P_0 is $\dot{\mathcal{P}}$. To see that the tangent space must have this structure, we consider an arbitrary, smooth one-dimensional submodel $\{P_\varepsilon\} \subset \mathcal{P}$ such that $P_\varepsilon|_{\varepsilon=0} = P_0$, whose score function at P_0 is

$$\frac{\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon} p_\varepsilon(y,a,x,d,g)\right|_{\varepsilon=0}}{p_0(y,a,x,d,g)} = h(y,a,x,d,g) \in L_2^0(P_0).$$

By the structure of the observed data density, the score function must be decomposable as the sum

$$h(y,a,x,d,g) = (1-g)\{h(y,a,x,d) + h(a,x,d) + h(d,x)\} + h(g,x) + h(x)$$

such that

$$\begin{aligned}
E_0\{h(Y,A,X,D) \mid A=a, X=x, D=d\} &= 0, \\
E_0\{h(A,X,D) \mid X=x, D=d\} &= 0, \\
E_0\{h(D,X) \mid X=x, G=0\} &= 0, \\
E_0\{h(G,X) \mid X=x\} &= 0, \\
E_0\{h(X)\} &= 0.
\end{aligned}$$

The restriction on the tangent space $\dot{\mathcal{P}}_y$ comes from the conditional moment restriction on the observed data distribution that $E_0(Y \mid A=1, X=x, D=d) - E_0(Y \mid A=0, X=x, D=d) = E_0(Y \mid A=1, X=x, D=d') - E_0(Y \mid A=0, X=x, D=d')$, i.e.

$$\int y\big\{p_0(y \mid 1,x,d) - p_0(y \mid 0,x,d)\big\}\mathrm{d}y = \int y\big\{p_0(y \mid 1,x,d') - p_0(y \mid 0,x,d')\big\}\mathrm{d}y,$$

for any $d, d' \in \mathcal{D}_x$. Differentiating both sides of the equation

$$\int y\{p_\varepsilon(y \mid 1, x, d) - p_\varepsilon(y \mid 0, x, d)\}\mathrm{d}y = \int y\{p_\varepsilon(y \mid 1, x, d') - p_\varepsilon(y \mid 0, x, d')\}\mathrm{d}y,$$

with respect to ε, we have that

$$\begin{aligned}&\int y\{p_0(y \mid 1, x, d)h(y, A = 1, x, D = d) - p_0(y \mid 0, x, d)h(y, A = 0, x, D = d)\}\mathrm{d}y \\ &= \int y\{p_0(y \mid 1, x, d')h(y, A = 1, x, D = d') - p_0(y \mid 0, x, d')h(y, A = 0, x, D = d')\}\mathrm{d}y,\end{aligned}$$

or equivalently, defining $\upsilon_h(a, x, d) = E_0\{Yh(Y, A, X, D) \mid A = a, X = x, D = d\}$,

$$\upsilon_h(1, x, d) - \upsilon_h(0, x, d) = \upsilon_h(1, x, d') - \upsilon_h(0, x, d'). \tag{S1}$$

We also have $\kappa_h(x) = \upsilon_h(1, x, d) - \upsilon_h(0, x, d)$ for any $x \in \mathcal{X}$ and $d \in \mathcal{D}_x$.

Constructing the maximal tangent space. We now show that we can construct parametric submodels, so that the closed linear span of the scores of these submodels is exactly $\dot{\mathcal{P}}$. The construction of $\dot{\mathcal{P}}_\bullet$ is standard for $\bullet \in \{a, x, d, g\}$ and omitted here. To construct the tangent subspace $\dot{\mathcal{P}}_y$, we only need to find parametric submodels whose score functions form a dense subset of $\dot{\mathcal{P}}_y$. For any $\ell_h = (1 - g)h(y, a, x, d) \in \dot{\mathcal{P}}_y$, consider the bounded version

$$\ell_{h,M}(y, a, x, d, g) = (1 - g)\big[h_M - E_0\{h_M \mid A = a, X = x, D = d\}\big] \in \tilde{\dot{\mathcal{P}}}_y,$$

where $h_M = hI(|h| \leqslant M)$, for some finite M. Then because $\ell_{h,M} \in \tilde{\dot{\mathcal{P}}}_y$, its projection onto $\dot{\mathcal{P}}_y$ is its projection onto $\Lambda^\perp$. Following the proof of Lemma S1, this is

$$\begin{aligned}\ell^\perp_{h,M}(y, a, x, d, g) &= \Pi\{\ell_{h,M} \mid \dot{\mathcal{P}}_y\} = \Pi\{\ell_{h,M} \mid \Lambda^\perp\} \\ &= (1 - g)\bigg[\{h_M - E_0(h_M \mid A = a, X = x, D = d)\} \\ &\qquad - \frac{2a - 1}{e_0(a \mid x, d)} q_{h,M}(x, d)\{y - \mu_0(a, x, d)\}\bigg], \end{aligned} \tag{S2}$$

where

$$\begin{aligned}r_{h,M}(x, d) &= w_0(x, d)\big[\mathrm{cov}_{P_0}(Y, h_M \mid A = 1, X = x, D = d) \\ &\qquad - \mathrm{cov}_{P_0}(Y, h_M \mid A = 0, X = x, D = d)\big], \\ q_{h,M}(x, d) &= r_{h,M}(x, d) - w_0(x, d)\frac{E_0\{r_{h,M}(X, D) \mid X = x, G = 0\}}{E_0\{w_0(X, D) \mid X = x, G = 0\}}.\end{aligned}$$

To simplify presentation, for sequences (a_n) and (b_n), we write $a_n \lesssim b_n$ if there is a universal constant $C > 0$ such that $a_n \leqslant Cb_n$. From Assumption 3, we bound the following quantities:

$$\begin{aligned}|r_{h,M}(x, d)| &\leqslant w_0(x, d) \sum_{a \in \{0,1\}} V_0^{1/2}(a, x, d)\{\mathrm{var}_{P_0}(h_M \mid A = a, X = x, D = d)\}^{1/2} \\ &\leqslant w_0(x, d) \sum_{a \in \{0,1\}} V_0^{1/2}(a, x, d)\{E_0(h_M^2 \mid A = a, X = x, D = d)\}^{1/2} \\ &\lesssim 1,\end{aligned} \tag{S3}$$

so that $|q_{h,M}(x, d)| \lesssim 1$ and $|\ell^\perp_{h,M}| \lesssim 1$ is bounded by some constant dependent on M. Then consider the parametric submodel $\{P_\varepsilon(h, M) : \varepsilon \in \Gamma\}$ with density

$$p_\varepsilon(y, a, x, d, g) = \big[p_\varepsilon(y \mid a, x, d)p_\varepsilon(a \mid x, d)p_\varepsilon(d \mid x)\big]^{(1-g)} p_\varepsilon(g \mid x)p_\varepsilon(x),$$

with

$$
\begin{aligned}
p_\varepsilon(y \mid a, x, d) &= p_0(y \mid a, x, d)\{1 + \varepsilon \ell^\perp_{h,M}(y, x, a, d, 0)\},\\
p_\varepsilon(a \mid x, d) &= \frac{\chi\{\varepsilon h(a, x, d)\} e_0(a \mid x, d)}{\sum_{a'} \chi\{\varepsilon h(a', x, d)\} e_0(a' \mid x, d)},\\
p_\varepsilon(d \mid x) &= \frac{\chi\{\varepsilon h(d, x)\} \zeta_0(d \mid x)}{\sum_{d'} \chi\{\varepsilon h(d', x)\} \zeta_0(d' \mid x)},\\
p_\varepsilon(g \mid x) &= \frac{\chi\{\varepsilon h(g, x)\} \{\pi_0(x)\}^g \{1 - \pi_0(x)\}^{(1-g)}}{\sum_{g'} \chi\{\varepsilon h(g', x)\} \{\pi_0(x)\}^{g'} \{1 - \pi_0(x)\}^{(1-g')}},\\
p_\varepsilon(x) &= \frac{\chi\{\varepsilon h(x)\} p_0(x)}{\int \chi\{\varepsilon h(x')\} p_0(x') \mathrm{d}x'},
\end{aligned}
$$

and

$$
\begin{aligned}
E_0\{h(Y, A, X, D) \mid A = a, X = x, D = d\} &= 0,\\
E_0\{h(A, X, D) \mid X = x, D = d\} &= 0,\\
E_0\{h(D, X) \mid X = x, G = 0\} &= 0,\\
E_0\{h(G, X) \mid X = x\} &= 0,\\
E_0\{h(X)\} &= 0,
\end{aligned}
$$

where $\chi(x) = 2\{1 + \exp(-2x)\}^{-1}$ (see, for example, Bickel et al., 1993, p. 53), Γ is an open neighborhood around 0 such that $p_\varepsilon(o) \geqslant 0$. Such a set Γ exists because $\ell^\perp_{h,M}$ is bounded. We verify that $\{P_\varepsilon(h, M) : \varepsilon \in \Gamma\} \subset \mathcal{P}$. It is obvious that $P_0(h, M) = P_0$. Additionally, for any $x \in \mathcal{X}_1$ and $d, d' \in \mathcal{D}_x$,

$$
\begin{aligned}
&E_{P_\varepsilon}(Y \mid A = 1, X = x, D = d) - E_{P_\varepsilon}(Y \mid A = 0, X = x, D = d)\\
&\qquad = E_{P_\varepsilon}(Y \mid A = 1, X = x, D = d') - E_{P_\varepsilon}(Y \mid A = 0, X = x, D = d'), \qquad \text{(S4)}
\end{aligned}
$$

because

$$
\begin{aligned}
&E_{P_\varepsilon}(Y \mid A = 1, X = x, D = d) - E_{P_\varepsilon}(Y \mid A = 0, X = x, D = d)\\
&\quad = \{E_0(Y \mid A = 1, X = x, D = d) - E_0(Y \mid A = 0, X = x, D = d)\}\\
&\qquad + \varepsilon\{E_0(Y\ell^\perp_{h,M} \mid A = 1, X = x, D = d) - E_0(Y\ell^\perp_{h,M} \mid A = 0, X = x, D = d)\}
\end{aligned}
$$

is a quantity not depending on d.

Furthermore, $\{\ell^\perp_{h,M} : \ell_h \in \dot{\mathcal{P}}_y, M < \infty\}$ is dense in $\dot{\mathcal{P}}_y$ in the $L_2(P_0)$-sense, which we show below. We bound the $L_2(P_0)$-distance between ℓ_h and $\ell^\perp_{h,M}$ by

$$
\begin{aligned}
\|\ell_h - \ell^\perp_{h,M}\|_{P_0} &\leqslant \|(1 - G)(h - h_M)\|_{P_0} && \text{(S5)}\\
&\quad + \|(1 - G)E_0(h_M \mid A, X, D)\|_{P_0} && \text{(S6)}\\
&\quad + \left\| (1 - G)\frac{2A - 1}{e_0(A \mid X, D)} q_{h,M}(X, D)\{Y - \mu_0(A, X, D)\} \right\|_{P_0}. && \text{(S7)}
\end{aligned}
$$

We argue that every term in the display above tends to zero as M tends to infinity. The limit of the norm (S5) is zero because bounded functions are dense in $L_2(P_0)$. We have $|h_M| \leqslant |h|$, so by dominated convergence, $\lim_{M\to\infty} E_0(h_M \mid A = a, X = x, D = d) = 0$, because $\lim_{M\to\infty} h_M = h$ and $E_0(h \mid A = a, X = x, D = d) = 0$ by definition. Since

$$
\{E_0(h_M \mid A = a, X = x, D = d)\}^2 \leqslant E_0(h_M^2 \mid A = a, X = x, D = d)
$$

$$\leqslant E_0(h^2 \mid A = a, X = x, D = d),$$

P_0-almost surely and $(1-g)h \in L_2(P_0)$, dominated convergence now shows that $\lim_{M\uparrow\infty} \|(1 - G)E_0(h_M \mid A, X, D)\|_{P_0} = 0$, so the limit of (S6) is zero.

Since

$$\begin{aligned}&|\{y - \mu_0(a, x, d)\}\{h_M - E_0(h_M \mid A = a, X = x, D = d)\}| \\ &\qquad \lesssim |h_M| + E_0(|h_M| \mid A = a, X = x, D = d)| \lesssim |h| + E_0(|h| \mid A = a, X = x, D = d)\end{aligned}$$

and the rightmost function is integrable with respect to $P_0(Y \mid A, X, D)$, $P_0(A, X, D \mid G = 0)$-almost surely, we have

$$\lim_{M\uparrow\infty} \operatorname{cov}_{P_0}(Y, h_M \mid A = a, X = x, D = d) = E_0(Yh \mid A = a, X = x, D = d)$$

from $\lim_{M\uparrow\infty}\{h_M - E_0(h_M \mid A = a, X = x, D = d)\} = h$. Then we have

$$\lim_{M\uparrow\infty} E_0\{r_{h,M}(X, D) \mid X = x, G = 0\} = \kappa_h(x)E_0\{w_0(X, D) \mid X = x, G = 0\}. \tag{S8}$$

Moreover,

$$\begin{aligned}r^2_{h,M}(x, d) &= w_0^2(x, d)[\operatorname{cov}_{P_0}(Y, h_M \mid A = 1, X = x, D = d) \\ &\qquad - \operatorname{cov}_{P_0}(Y, h_M \mid A = 0, X = x, D = d)]^2 \\ &\lesssim \sum_{a\in\{0,1\}} \operatorname{cov}^2_{P_0}(Y, h_M \mid A = a, X = x, D = d) \\ &\leqslant \sum_{a\in\{0,1\}} V_0(a, x, d)E_0(h_M^2 \mid A = a, X = x, D = d) \\ &\lesssim \sum_{a\in\{0,1\}} E_0(h^2 \mid A = a, X = x, D = d) \\ &\lesssim \sum_{a\in\{0,1\}} e_0(a \mid x, d)E_0(h^2 \mid A = a, X = x, D = d) \\ &= E_0(h^2 \mid X = x, D = d),\end{aligned}$$

so that

$$\begin{aligned}q^2_{h,M}(x, d) &= \left[r_{h,M}(x, d) + w_0(x, d)\frac{E_0\{r_{h,M}(X, D) \mid X = x, G = 0\}}{E_0\{w_0(X, D) \mid X = x, G = 0\}}\right]^2 \\ &\leqslant 2r^2_{h,M}(x, d) + 2w_0^2(x, d)\left[\frac{E_0\{r_{h,M}(X, D) \mid X = x, G = 0\}}{E_0\{w_0(X, D) \mid X = x, G = 0\}}\right]^2 \\ &\lesssim E_0(h^2 \mid X = x, D = d) + E_0\{r^2_{h,M}(X, D) \mid X = x, G = 0\} \\ &\lesssim E_0(h^2 \mid X = x, D = d) + E_0(h^2 \mid X = x, G = 0) \in L_1(P_0).\end{aligned}$$

Expression (S7) can be bounded by $O(\|(1 - G)q_{h,M}(X, D)\|_{P_0})$. Another application of dominated convergence yields $\lim_{M\uparrow\infty} \|(1 - G)q_{h,M}(X, D)\|_{P_0} = 0$, because by (S8),

$$\lim_{M\uparrow\infty} q^2_{h,M}(x, d) = \left\{\lim_{M\uparrow\infty} q_{h,M}(x, d)\right\}^2 = 0,$$

so the limit of (S7) is zero. The denseness follows from $\lim_{M\uparrow\infty} \|\ell_h - \ell^{\perp}_{h,M}\|_{P_0} = 0$ for any $\ell_h \in \dot{\mathcal{P}}_y$.

Therefore, the closed linear span of $\{\ell^{\perp}_{h,M} : \ell_h \in \dot{\mathcal{P}}_y, M < \infty\}$ is exactly $\dot{\mathcal{P}}_y$.

Calculating the pathwise derivative. We next compute the pathwise derivative of θ_ε, the target parameter on P_ε, along the submodel $\{P_\varepsilon : \varepsilon \in \Gamma\}$ at P_0. To express the target parameter as a function of only the observed data and its density, we notice that

$$\begin{aligned}
\theta_\varepsilon &= E_{P_\varepsilon}\{\delta_\varepsilon(X) \mid G = 1\} \\
&= \frac{1}{\alpha_\varepsilon} \int \delta_\varepsilon(x)\pi_\varepsilon(x)p_\varepsilon(x)\mathrm{d}x \\
&= \frac{1}{\alpha_\varepsilon} \int \sum_{d\in[m]} \zeta_\varepsilon(d \mid x)\delta_\varepsilon(x)\pi_\varepsilon(x)p_\varepsilon(x)\mathrm{d}x \\
&= \frac{1}{\alpha_\varepsilon} \int \sum_{d\in[m]} \zeta_\varepsilon(d \mid x)\{\mu_\varepsilon(1,x,d) - \mu_\varepsilon(0,x,d)\}\pi_\varepsilon(x)p_\varepsilon(x)\mathrm{d}x \\
&= \frac{1}{\alpha_\varepsilon} \int \sum_{d\in[m]} \zeta_\varepsilon(d \mid x) \int y\{p_\varepsilon(y \mid 1,x,d) - p_\varepsilon(y \mid 0,x,d)\}\mathrm{d}y\pi_\varepsilon(x)p_\varepsilon(x)\mathrm{d}x.
\end{aligned}$$

The pathwise derivative of θ_ε evaluated at the true model is

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_\varepsilon\bigg|_{\varepsilon=0} \\
&\quad = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \frac{\int \sum_{d\in[m]} \zeta_\varepsilon(d \mid x) \int y\{p_\varepsilon(y \mid 1,x,d) - p_\varepsilon(y \mid 0,x,d)\}\mathrm{d}y\pi_\varepsilon(x)p_\varepsilon(x)\mathrm{d}x}{\int \pi_\varepsilon(x)p_\varepsilon(x)\mathrm{d}x}\bigg|_{\varepsilon=0},
\end{aligned}$$

which by the product rule is

$$\begin{aligned}
&= \frac{1}{\alpha_0}\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \int \sum_{d\in[m]} \zeta_\varepsilon(d \mid x) \int y\{p_\varepsilon(y \mid 1,x,d) - p_\varepsilon(y \mid 0,x,d)\}\mathrm{d}y\pi_\varepsilon(x)p_\varepsilon(x)\mathrm{d}x\bigg|_{\varepsilon=0} \\
&\qquad - \frac{\theta_0}{\alpha_0}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\int \pi_\varepsilon(x)p_\varepsilon(x)\mathrm{d}x\bigg|_{\varepsilon=0} \\
&= \frac{1}{\alpha_0} \int \sum_{d\in[m]} \zeta_0(d \mid x) \int y\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\{p_0(y \mid 1,x,d;\varepsilon) - p_0(y \mid 0,x,d;\varepsilon)\}\bigg|_{\varepsilon=0}\right) \qquad \text{(S9)} \\
&\qquad\qquad \mathrm{d}y\pi_0(x)p_0(x)\mathrm{d}x \\
&\qquad + \frac{1}{\alpha_0}\int \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \sum_{d\in[m]} \zeta_\varepsilon(d \mid x)\bigg|_{\varepsilon=0} \delta_0(x)\pi_0(x)p_0(x)\mathrm{d}x \qquad \text{(S10)} \\
&\qquad + \frac{1}{\alpha_0}\int \delta_0(x)\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\pi_\varepsilon(x)p_\varepsilon(x)\bigg|_{\varepsilon=0}\right)\mathrm{d}x \qquad \text{(S11)} \\
&\qquad - \frac{\theta_0}{\alpha_0}\int \left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\pi_\varepsilon(x)p_\varepsilon(x)\bigg|_{\varepsilon=0}\right)\mathrm{d}x. \qquad \text{(S12)}
\end{aligned}$$

We study the expressions separately. We have

$$\begin{aligned}
\text{(S9)} &= \frac{1}{\alpha_0} \int \sum_{d\in[m]} \zeta_0(d \mid x) \int y\big\{p_0(y \mid 1,x,d)h(y, A = 1, x, D = d) \\
&\qquad\qquad - p_0(y \mid 0,x,d)h(y, A = 0, x, D = d)\big\}\mathrm{d}y\pi_0(x)p_0(x)\mathrm{d}x \\
&= \frac{1}{\alpha_0} \int \sum_{d\in[m]} \zeta_0(d \mid x)\kappa_h(x)\pi_0(x)p_0(x)\mathrm{d}x \\
&= \frac{1}{\alpha_0} \int \kappa_h(x)\pi_0(x)p_0(x)\mathrm{d}x,
\end{aligned}$$

$$\text{(S10)} = \frac{1}{\alpha}\int\left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon}1\Big|_{\varepsilon=0}\right)\delta_0(x)\pi_0(x)p_0(x)\mathrm{d}x = 0.$$

We observe that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\pi_\varepsilon(x)p_\varepsilon(x)\Big|_{\varepsilon=0} = \pi_0(x)p_0(x)\{h(G=1,x)+h(x)\},$$

and therefore

$$\text{(S11)} + \text{(S12)} = \frac{1}{\alpha_0}\int\{\delta_0(x)-\theta_0\}\pi_0(x)p_0(x)\{h(G=1,x)+h(x)\}\mathrm{d}x.$$

Collecting the terms, the pathwise derivative is

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_\varepsilon\Big|_{\varepsilon=0} = \frac{1}{\alpha_0}\int\kappa_h(x)\pi_0(x)p_0(x)\mathrm{d}x + \frac{1}{\alpha_0}\int\{\delta_0(x)-\theta_0\}\pi_0(x)p_0(x)\{h(G=1,x)+h(x)\}\mathrm{d}x$$

Finding the efficient influence function. We claim that the efficient influence function of θ_0 in the model $\mathcal{P}$ is as displayed in Lemma 2.

The inner product of φ_{w_0} and any score $h(o) \in \dot{\mathcal{P}}$ of the model $\mathcal{P}$ at P_0 is

$$\begin{aligned}
&E_0\{\varphi_{w_0}(O)h(O)\} \\
&\quad= E_0\left[\frac{(1-G)(2A-1)\pi_0(X)}{\alpha_0 e_0(A\mid X,D)\{1-\pi_0(X)\}}\frac{w_0(X,D)}{\sum_{d\in[m]}\zeta_0(d\mid X)w_0(X,d)}\right. \qquad \text{(S13)}\\
&\qquad\qquad\left.\{Y-\mu_0(A,X,D)\}h(Y,A,X,D)\right] \\
&\quad\quad + E_0\left[\frac{G}{\alpha_0}\{\delta_0(X)-\theta_0\}\{h(G,X)+h(X)\}\right]. \qquad \text{(S14)}
\end{aligned}$$

Since the score $h(y,a,x,d)$ must satisfy $E_0\{h(Y,A,X,D)\mid A,X,D\}=0$, it follows that

$$\begin{aligned}
\text{(S13)} &= E_0\left\{\frac{(1-G)(2A-1)\pi_0(X)}{\alpha_0 e_0(A\mid X,D)\{1-\pi_0(X)\}}\frac{w_0(X,D)}{\sum_{d\in[m]}\zeta_0(d\mid X)w_0(X,d)}Yh(Y,A,X,D)\right\} \\
&= E_0\left[\frac{(1-G)(2A-1)\pi_0(X)}{\alpha_0 e_0(A\mid X,D)\{1-\pi_0(X)\}}\frac{w_0(X,D)}{\sum_{d\in[m]}\zeta_0(d\mid X)w_0(X,d)}\upsilon_h(A,X,D)\right] \\
&= E_0\left[\frac{(1-G)\pi_0(X)}{\alpha_0\{1-\pi_0(X)\}}\frac{w_0(X,D)}{\sum_{d\in[m]}\zeta_0(d\mid X)w_0(X,d)}\{\upsilon_h(1,X,D)-\upsilon_h(0,X,D)\}\right] \\
&= E_0\left\{\frac{(1-G)\pi_0(X)}{\alpha_0\{1-\pi_0(X)\}}\kappa_h(X)E_0\left(\frac{w_0(X,D)}{\sum_{d\in[m]}\zeta_0(d\mid X)w_0(X,d)}\;\middle|\; X,G=0\right)\right\} \\
&= E_0\left\{\frac{(1-G)\pi_0(X)}{\alpha_0\{1-\pi_0(X)\}}\kappa_h(X)\right\} \\
&= E_0\left\{\frac{\pi_0(X)}{\alpha_0}\kappa_h(X)\right\} \\
&= \frac{1}{\alpha_0}\int\kappa_h(x)\pi_0(x)p_0(x)\mathrm{d}x.
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
\text{(S14)} &= E_0\left[\frac{G}{\alpha_0}\{\delta_0(X)-\theta_0\}\{h(G,X)+h(X)\}\right] \\
&= \frac{1}{\alpha_0}\int\{\delta_0(x)-\theta_0\}\{h(G=1,x)+h(x)\}\pi_0(x)p_0(x)\mathrm{d}x.
\end{aligned}$$

Therefore, it is clear that

$$E_0\{\varphi_{w_0}(O)h(O)\} = (S13) + (S14) = \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_\varepsilon\Big|_{\varepsilon=0},$$

so φ_{w_0} is an influence function. To show that φ_{w_0} is the efficient influence function, it remains to verify that φ_{w_0} lies in the tangent space $\dot{\mathcal{P}}$.

The function φ_{w_0} can be decomposed into the sum

$$\varphi_{w_0}(o) = (1-g)h^*(y,a,x,d) + h^*(g,x) + h^*(x),$$

where

$$\begin{aligned}
h^*(y,a,x,d) &= \frac{(2a-1)\pi_0(x)}{\alpha_0 e_0(a\,|\,x,d)\{1-\pi_0(x)\}} \frac{w_0(x,d)}{\sum_{d'\in[m]}\zeta_0(d'\,|\,x)w_0(x,d')}\{y-\mu_0(a,x,d)\},\\
h^*(g,x) &= \frac{g-\pi_0(x)}{\alpha_0}\{\delta_0(x)-\theta_0\},\\
h^*(x) &= \frac{\pi_0(x)}{\alpha_0}\{\delta_0(x)-\theta_0\}.
\end{aligned}$$

It is trivial to verify that $E_0\{h^*(Y,A,X,D)\,|\,A,X,D\} = 0$, $E_0\{h^*(G,X)\,|\,X\} = 0$, as well as $E_0\{h^*(X)\} = 0$, so $h^*(G,X) \in \dot{\mathcal{P}}_g$ and $h^*(X) \in \dot{\mathcal{P}}_x$. Now for any $d \in \mathcal{D}_x$,

$$\begin{aligned}
&E_0\{Yh^*(Y,A,X,D)\,|\,A=1,X=x,D=d\} - E_0\{Yh^*(Y,A,X,D)\,|\,A=0,X=x,D=d\}\\
&\quad= E_0[\{Y-\mu_0(A,X,D)\}h^*(Y,A,X,D)\,|\,A=1,X=x,D=d]\\
&\qquad- E_0[\{Y-\mu_0(A,X,D)\}h^*(Y,A,X,D)\,|\,A=0,X=x,D=d]\\
&\quad= \frac{\pi_0(x)}{\alpha_0\{1-\pi_0(x)\}}\frac{w_0(x,d)}{\sum_{d'\in[m]}\zeta_0(d'\,|\,x)w_0(x,d')}\left\{\frac{V_0(1,x,d)}{e_0(1\,|\,x,d)}+\frac{V_0(0,x,d)}{e_0(0\,|\,x,d)}\right\}\\
&\quad= \frac{\pi_0(x)}{\alpha_0\{1-\pi_0(x)\}}\frac{1}{\sum_{d'\in[m]}\zeta_0(d'\,|\,x)w_0(x,d')},
\end{aligned}$$

which does not depend on the value of d. This shows that

$$\begin{aligned}
&E_0\{Yh^*(Y,A,X,D)\,|\,A=1,X=x,D=d\} - E_0\{Yh^*(Y,A,X,D)\,|\,A=0,X=x,D=d\}\\
&= E_0\{Yh^*(Y,A,X,D)\,|\,A=1,X=x,D=d'\} - E_0\{Yh^*(Y,A,X,D)\,|\,A=0,X=x,D=d'\},
\end{aligned}$$

for any $d, d' \in \mathcal{D}_x$, and thus $(1-g)h^*(y,a,x,d) \in \dot{\mathcal{P}}_y$. □

S2.3. *Proof of Corollary 1*

The first part of Corollary 1 is a direct result from the proof of Lemma 2. To prove the second part of Corollary 1, note that the space of the influence functions of the parameter θ_0 can be characterized by the translation $\varphi_{w_0} + \dot{\mathcal{P}}^\perp$, where φ_{w_0} is the efficient influence function in Lemma 2. Therefore, if we choose

$$\begin{aligned}
u_{\tilde{w}}(o) = &\frac{(1-g)\pi_0(x)}{\alpha_0\{1-\pi_0(x)\}}\frac{2a-1}{e_0(a\,|\,x,d)}\\
&\left\{\frac{\tilde{w}(x,d)}{\sum_{d'\in[m]}\zeta_0(d'\,|\,x)\tilde{w}(x,d')} - \frac{w_0(x,d)}{\sum_{d'\in[m]}\zeta_0(d'\,|\,x)w_0(x,d')}\right\}\{y-\mu_0(a,x,d)\},
\end{aligned}$$

for $\tilde{w}$ as stated in Corollary 1, then $u_{\tilde{w}} \in \dot{\mathcal{P}}^\perp$, and $\tilde{\varphi} = \varphi_{w_0} + u_{\tilde{w}}$ is an influence function of θ_0.

S2.4. *Proof of Theorem 1*

We quote a lemma from Kennedy (2024). For any function f of the data and $k \in [K]$, let

$$\mathbb{P}_{n,k} f = \frac{K}{n} \sum_{i \in \mathcal{I}_k} f(O_i).$$

Lemma S2. *Let f_k be a random function which only depends on the sample $\mathcal{O}_{-k} = \{O_i : i \notin \mathcal{I}_k\}$. Then $(\mathbb{P}_{n,k} - P_0) f_k = O_{P_0}(n^{-1/2} \|f_k\|_{P_0})$.*

Proof. By the Markov inequality conditional on $\mathcal{O}_{-k}$, for $t > 0$,

$$\mathrm{pr}\left\{ \frac{|(n/K)^{1/2}(\mathbb{P}_{n,k} - P_0) f_k|^2}{\|f_k\|_{P_0}^2} \geqslant t \;\middle|\; \mathcal{O}_{-k} \right\} \leqslant \frac{P_0(f_k - P_0 f_k)^2}{t \|f_k\|_{P_0}^2} \leqslant \frac{1}{t}.$$

Marginally, we have $\mathrm{pr}\{|(n/K)^{1/2}(\mathbb{P}_{n,k} - P_0) f_k| / \|f_k\|_{P_0} \geqslant t^{1/2}\} \leqslant t^{-1}$. Therefore, we also have $|(n/K)^{1/2}(\mathbb{P}_{n,k} - P_0) f_k| / \|f_k\|_{P_0} = O_{P_0}(1)$, and the lemma follows, since K does not depend on n. □

Define

$$\ell_{\bar{\eta}}(o) = \frac{1-g}{\alpha_0} \frac{\pi_0(x)}{1-\pi_0(x)} \frac{\bar{w}(x,d)}{\sum_{d' \in [m]} \zeta_0(d' \,|\, x) \bar{w}(x,d')} \frac{2a-1}{e_0(a \,|\, x, d)} \{y - \mu_0(a,x,d)\} + \frac{g}{\alpha_0} \delta_0(x),$$

$$\ell_{\hat{\eta}_k}(o) = \frac{1-g}{\hat{\alpha}} \frac{\hat{\pi}_k(x)}{1-\hat{\pi}_k(x)} \frac{\hat{w}_k(x,d)}{\sum_{d' \in [m]} \hat{\zeta}_k(d' \,|\, x) \hat{w}_k(x,d')} \frac{2a-1}{\hat{e}_k(a \,|\, x, d)} \{y - \hat{\mu}_k(a,x,d)\} + \frac{g}{\hat{\alpha}} \hat{\delta}_k(x).$$

Lemma S3. *If Assumption 4 is satisfied, then $\|\ell_{\hat{\eta}_k}\|_{P_0} = O_{P_0}(1)$ for every $k \in [K]$.*

Proof. Let

$$H_k(a,x,d) = \frac{\hat{\pi}_k(x)}{1-\hat{\pi}_k(x)} \frac{\hat{w}_k(x,d)}{\sum_{d' \in [m]} \hat{\zeta}_k(d' \,|\, x) \hat{w}_k(x,d')} \frac{1}{\hat{e}_k(a \,|\, x, d)} \lesssim 1.$$

We also have

$$|\hat{\delta}_k(x)| \leqslant \sum_{d \in [m]} \sum_{a \in \{0,1\}} \frac{\left|\hat{w}_k(x,d) \hat{\zeta}_k(d \,|\, x) \hat{\mu}_k(a,x,d)\right|}{\left|\sum_{d' \in [m]} \hat{w}_k(x,d') \hat{\zeta}_k(d' \,|\, x)\right|} \lesssim 1.$$

Then

$$\begin{aligned}
&\|\ell_{\hat{\eta}_k}\|_{P_0}^2 \\
&\quad \lesssim P_0[(1-G) H_k^2(A,X,D) \{Y - \hat{\mu}_k(A,X,D)\}^2] + P_0\{\pi_0(X) \hat{\delta}_k^2(X)\} \\
&\quad \lesssim P_0[(1-G)\{Y - \mu_0(A,X,D)\}^2] + P_0[(1-G)(\hat{\mu}_k - \bar{\mu})^2(A,X,D)] \\
&\qquad + P_0[(1-G)(\bar{\mu}^2 + \mu_0^2)(A,X,D)] + 1 \\
&\quad \lesssim P_0[(1-G) V_0(A,X,D)] + \max_{d \in [m]} \max_{a \in \{0,1\}} \|(\hat{\mu}_k - \bar{\mu})(a,X,d) I\{\zeta_0(d \,|\, X) > 0\}\|_{P_0}^2 + 1 \\
&\quad \lesssim o_{P_0}(1) + 1 = O_{P_0}(1).
\end{aligned}$$

□

Lemma S4. *If Assumptions 4 and 5b(i) are satisfied, then* $\|\ell_{\hat{\eta}_k} - \ell_{\bar{\eta}}\|_{P_0} \xrightarrow{\mathrm{p}} 0$ *for every* $k \in [K]$.

Proof. We first show consistency of $\hat{\delta}_k$. By the triangular inequality,

$$
\begin{aligned}
&\|(\hat{\delta}_k - \delta_0)(X)\|_{P_0} \\
&\quad\leqslant \sum_{a\in\{0,1\}}\sum_{d\in[m]} \left\{\left(P_0\left[\frac{\{\hat{w}_k(X,d)(\hat{\zeta}_k-\zeta_0)(d\,|\,X)\hat{\mu}_k(a,X,d)\}^2}{\{\sum_{d'\in[m]}\hat{w}_k(X,d')\hat{\zeta}_k(d'\,|\,X)\}^2}\right]\right)^{1/2}\right. \\
&\qquad + \sum_{a\in\{0,1\}}\sum_{d\in[m]} \left(P_0\left[\frac{\{(\hat{w}_k-\bar{w})(X,d)\zeta_0(d\,|\,X)\hat{\mu}_k(a,X,d)\}^2}{\{\sum_{d'\in[m]}\hat{w}_k(X,d')\hat{\zeta}_k(d'\,|\,X)\}^2}\right]\right)^{1/2} \\
&\qquad + \sum_{a\in\{0,1\}}\sum_{d\in[m]} \left(P_0\left[\left\{\frac{1}{\sum_{d'\in[m]}\hat{w}_k(X,d')\hat{\zeta}_k(d'\,|\,X)}\right.\right.\right. \\
&\qquad\qquad \left.\left.\left. - \frac{1}{\sum_{d'\in[m]}\bar{w}(X,d')\zeta_0(d'\,|\,X)}\right\}^2 \right.\right. \\
&\qquad\qquad \left.\left.\{\bar{w}(X,d)\zeta_0(d\,|\,X)\hat{\mu}_k(a,X,d)\}^2\right]\right)^{1/2} \\
&\qquad + \sum_{a\in\{0,1\}}\sum_{d\in[m]} \left(P_0\left[\frac{\{\bar{w}(X,d)\zeta_0(d\,|\,X)(\hat{\mu}_k-\mu_0)(a,X,d)\}^2}{\{\sum_{d'\in[m]}\bar{w}(X,d')\zeta_0(d'\,|\,X)\}^2}\right]\right)^{1/2} \\
&\quad\lesssim \max_{d\in[m]}\|(\hat{\zeta}_k-\zeta_0)(d\,|\,X)\|_{P_0} + \max_{d\in[m]}\|(\hat{w}_k-\bar{w})(X,d)I\{\eta_0(d\,|\,X)>0\}\|_{P_0} \\
&\qquad + \max_{d\in[m]}\max_{a\in\{0,1\}}\|(\hat{\mu}_k-\mu_0)(a,X,d)I\{\eta_0(d\,|\,X)>0\}\|_{P_0} = o_{P_0}(1).
\end{aligned}
$$

We write $\ell_{\hat{\eta}_k} - \ell_{\bar{\eta}}$ as a sum such that

$$
\begin{aligned}
\ell_{\hat{\eta}_k} - \ell_{\bar{\eta}} = {}& \frac{1-g}{\hat{\alpha}_k}\frac{\hat{\pi}_k(x)}{1-\hat{\pi}_k(x)}\frac{\hat{w}_k(x,d)}{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,x)\hat{w}_k(x,d')}\frac{2a-1}{\hat{e}_k(a\,|\,x,d)}(\mu_0-\hat{\mu})(a,x,d) \\
&- \frac{1-g}{\hat{\alpha}_k}\frac{\hat{\pi}_k(x)}{1-\hat{\pi}_k(x)}\frac{\hat{w}_k(x,d)}{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,x)\hat{w}_k(x,d')}\frac{2a-1}{\hat{e}_k(a\,|\,x,d)e_0(a\,|\,x,d)} \\
&\qquad (\hat{e}_k-e_0)(a\,|\,x,d)\{y-\mu_0(a,x,d)\} \\
&+ \frac{1-g}{\hat{\alpha}_k}\frac{\hat{\pi}_k(x)}{1-\hat{\pi}_k(x)}\frac{(\hat{w}_k-\bar{w})(x,d)}{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,x)\hat{w}_k(x,d')}\frac{2a-1}{e_0(a\,|\,x,d)}\{y-\mu_0(a,x,d)\} \\
&+ \frac{1-g}{\hat{\alpha}_k}\frac{\hat{\pi}_k(x)}{1-\hat{\pi}_k(x)}\frac{\bar{w}(x,d)\sum_{d'\in[m]}(\zeta_0-\hat{\zeta}_k)(d'\,|\,x)\hat{w}_k(x,d')}{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,x)\hat{w}_k(x,d')\sum_{d'\in[m]}\zeta_0(d'\,|\,x)\bar{w}(x,d')} \\
&\qquad \frac{2a-1}{e_0(a\,|\,x,d)}\{y-\mu_0(a,x,d)\} \\
&+ \frac{1-g}{\hat{\alpha}_k}\frac{\hat{\pi}_k(x)}{1-\hat{\pi}_k(x)}\frac{\bar{w}(x,d)\sum_{d'\in[m]}\zeta_0(d'\,|\,x)(\bar{w}-\hat{w}_k)(x,d')}{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,x)\hat{w}_k(x,d')\sum_{d'\in[m]}\zeta_0(d'\,|\,x)\bar{w}(x,d')} \\
&\qquad \frac{2a-1}{e_0(a\,|\,x,d)}\{y-\mu_0(a,x,d)\} \\
&+ \frac{1-g}{\hat{\alpha}_k}\frac{(\hat{\pi}_k-\pi_0)(x)}{\{1-\pi_0(x)\}\{1-\hat{\pi}_k(x)\}}\frac{\bar{w}(x,d)}{\sum_{d'\in[m]}\zeta_0(d'\,|\,x)\bar{w}(x,d')} \\
&\qquad \frac{2a-1}{e_0(a\,|\,x,d)}\{y-\mu_0(a,x,d)\}
\end{aligned}
$$

$$+\frac{(1-g)(\alpha_0-\hat{\alpha}_k)}{\alpha_0\hat{\alpha}_k}\frac{\pi_0(x)}{1-\pi_0(x)}\frac{\bar{w}(x,d)}{\sum_{d'\in[m]}\zeta_0(d'\,|\,x)\bar{w}(x,d')}\frac{2a-1}{e_0(a\,|\,x,d)}\{y-\mu_0(a,x,d)\}$$

$$+\frac{g}{\hat{\alpha}_k}(\hat{\delta}-\delta_0)(x)+\frac{g(\alpha_0-\hat{\alpha}_k)}{\alpha_0\hat{\alpha}_k}\delta_0(x)$$

By the triangular inequality,

$$\begin{aligned}
&\|\ell_{\hat{\eta}_k}-\ell_{\bar{\eta}}\|_{P_0}\\
&\leqslant\left(P_0\left[\frac{1-\pi_0(X)}{\hat{\alpha}_k^2}\frac{\hat{\pi}_k^2(X)}{\{1-\hat{\pi}_k(X)\}^2}\sum_{d\in[m]}\frac{\hat{w}_k^2(X,d)\zeta_0(d\,|\,X)}{\{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,X)\hat{w}_k(X,d')\}^2}\right.\right.\\
&\qquad\left.\left.\sum_{a\in\{0,1\}}\frac{e_0(a\,|\,X,d)}{\hat{e}_k^2(a\,|\,X,d)}(\mu_0-\hat{\mu})^2(a,X,d)\right]\right)^{1/2}\\
&\quad+\left(P_0\left[\frac{1-\pi_0(X)}{\hat{\alpha}_k^2}\frac{\hat{\pi}_k^2(X)}{\{1-\hat{\pi}_k(X)\}^2}\sum_{d\in[m]}\frac{\hat{w}_k^2(X,d)\zeta_0(d\,|\,X)}{\{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,X)\hat{w}_k(X,d')\}^2}\right.\right.\\
&\qquad\left.\left.\sum_{a\in\{0,1\}}\frac{(\hat{e}_k-e_0)^2(a\,|\,X,d)}{\hat{e}_k^2(a\,|\,X,d)e_0(a\,|\,X,d)}V_0(a,X,d)\right]\right)^{1/2}\\
&\quad+\left(P_0\left[\frac{1-\pi_0(X)}{\hat{\alpha}_k^2}\frac{\hat{\pi}_k^2(X)}{\{1-\hat{\pi}_k(X)\}^2}\sum_{d\in[m]}\frac{(\hat{w}_k-\bar{w})^2(X,d)\zeta_0(d\,|\,X)w_0^{-1}(X,d)}{\{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,X)\hat{w}_k(X,d')\}^2}\right]\right)^{1/2}\\
&\quad+\left(P_0\left[\frac{1-\pi_0(X)}{\hat{\alpha}_k^2}\frac{\hat{\pi}_k^2(X)}{\{1-\hat{\pi}_k(X)\}^2}\left\{\sum_{d'\in[m]}(\zeta_0-\hat{\zeta}_k)(d'\,|\,X)\hat{w}_k(X,d')\right\}^2\right.\right.\\
&\qquad\left.\left.\frac{\sum_{d\in[m]}\bar{w}^2(X,d)\zeta_0(d\,|\,X)w_0^{-1}(X,d)}{\{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,X)\hat{w}_k(X,d')\}^2\{\sum_{d'\in[m]}\zeta_0(d'\,|\,X)\bar{w}(X,d')\}^2}\right]\right)^{1/2}\\
&\quad+\left(P_0\left[\frac{1-\pi_0(X)}{\hat{\alpha}_k^2}\frac{\hat{\pi}_k^2(X)}{\{1-\hat{\pi}_k(X)\}^2}\left\{\sum_{d'\in[m]}\zeta_0(d'\,|\,X)(\hat{w}_k-\bar{w})(X,d')\right\}^2\right.\right.\\
&\qquad\left.\left.\frac{\sum_{d\in[m]}\bar{w}^2(X,d)\zeta_0(d\,|\,X)w_0^{-1}(X,d)}{\{\sum_{d'\in[m]}\hat{\zeta}_k(d'\,|\,X)\hat{w}_k(X,d')\}^2\{\sum_{d'\in[m]}\zeta_0(d'\,|\,X)\bar{w}(X,d')\}^2}\right]\right)^{1/2}\\
&\quad+\left(P_0\left[\frac{1}{\hat{\alpha}_k^2}\frac{(\hat{\pi}_k-\pi_0)^2(X)}{\{1-\pi_0(X)\}\{1-\hat{\pi}_k(X)\}^2}\sum_{d\in[m]}\frac{\bar{w}^2(X,d)\zeta_0(d\,|\,X)w_0^{-1}(X,d)}{\{\sum_{d'\in[m]}\zeta_0(d'\,|\,X)\bar{w}(X,d')\}^2}\right]\right)^{1/2}\\
&\quad+\left(P_0\left[\frac{(\alpha_0-\hat{\alpha}_k)^2}{\alpha_0^2\hat{\alpha}_k^2}\frac{\pi_0^2(X)}{1-\pi_0(X)}\sum_{d\in[m]}\frac{\bar{w}^2(X,d)\zeta_0(d\,|\,X)w_0^{-1}(X,d)}{\{\sum_{d'\in[m]}\zeta_0(d'\,|\,X)\bar{w}(X,d')\}^2}\right]\right)^{1/2}\\
&\quad+\left(P_0\left[\frac{\pi_0(X)}{\hat{\alpha}_k^2}(\hat{\delta}-\delta_0)^2(X)\right]\right)^{1/2}+\left(P_0\left[\frac{\pi_0(X)(\alpha_0-\hat{\alpha}_k)^2}{\alpha_0^2\hat{\alpha}_k^2}\delta^2(X)\right]\right)^{1/2}\\
&\lesssim\max_{d\in[m]}\max_{a\in\{0,1\}}\|(\hat{\mu}_k-\mu_0)(a,X,d)I\{\zeta_0(d\,|\,X)>0\}\|_{P_0}\\
&\quad+\max_{d\in[m]}\max_{a\in\{0,1\}}\|(\hat{e}_k-e_0)(a,X,d)I\{\zeta_0(d\,|\,X)>0\}\|_{P_0}\\
&\quad+\max_{d\in[m]}\|(\hat{w}_k-\bar{w})(X,d)I\{\zeta_0(d\,|\,X)>0\}\|_{P_0}
\end{aligned}$$

$$+ \max_{d\in[m]} \|(\hat{\zeta}_k - \zeta_0)(d \mid X)\|_{P_0} + \|(\hat{\pi}_k - \pi_0)(X)\|_{P_0} + \|(\hat{\delta}_k - \delta_0)(X)\|_{P_0} + |\hat{\alpha}_k - \alpha_0|,$$

which converges in probability to zero by assumption and by $\|(\hat{\delta}_k - \delta_0)(X)\|_{P_0} = o_{P_0}(1)$ shown earlier in the proof. □

Proof of Theorem 1. The difference between the estimator and the target parameter decomposes as

$$\hat{\theta} - \theta_0 = \frac{1}{K} \sum_{k\in[K]} \left\{ (\mathbb{P}_{n,k} - P_0)\ell_{\hat{\eta}_k} + \left(P_0\ell_{\hat{\eta}_k} - \frac{\alpha_0}{\hat{\alpha}_k}\theta_0 \right) - \frac{\hat{\alpha}_k - \alpha_0}{\hat{\alpha}_k}\theta_0 \right\}. \tag{S15}$$

The second term in (S15) is, using $\theta_0 = E_0\{\pi_0(X)\delta_0(X)\}/\alpha_0$,

$$\begin{aligned} P_0\ell_{\hat{\eta}_k} - \frac{\alpha_0}{\hat{\alpha}_k}\theta_0 = \frac{1}{\hat{\alpha}_k} P_0 \Bigg\{ & \frac{\{1-\pi_0(X)\}\hat{\pi}_k(X)}{\{1-\hat{\pi}_k(X)\}} \sum_{d\in[m]} \frac{\hat{w}_k(X,d)\zeta_0(d \mid X)}{\sum_{d'\in[m]} \hat{w}_k(X,d')\hat{\zeta}_k(d' \mid X)} \\ & \sum_{a\in\{0,1\}} \frac{(2a-1)e_0(a \mid X,d)}{\hat{e}_k(a \mid X,d)} (\mu_0 - \hat{\mu}_k)(a,X,d) \Bigg\} \\ & + P_0\left[\frac{\pi_0(X)}{\hat{\alpha}_k} (\hat{\delta}_k - \delta_0)(X) \right]. \end{aligned} \tag{S16}$$

Developing from (S16),

$$\begin{aligned} \text{(S16)} &= P_0\Bigg[\frac{\{1-\pi_0(X)\}\hat{\pi}_k(X)}{\hat{\alpha}_k\{1-\hat{\pi}_k(X)\}} \sum_{d\in[m]} \frac{\hat{w}_k(X,d)\zeta_0(d \mid X)}{\sum_{d'\in[m]} \hat{w}_k(X,d')\hat{\zeta}_k(d' \mid X)} \\ &\qquad \sum_{a\in\{0,1\}} \frac{(2a-1)e_0(a \mid X,d)}{\hat{e}_k(a \mid X,d)} (\mu_0 - \hat{\mu}_k)(a,X,d) \Bigg] \\ &= P_0\Bigg[\frac{\{1-\pi_0(X)\}\hat{\pi}_k(X)}{\hat{\alpha}_k\{1-\hat{\pi}_k(X)\}} \frac{\sum_{d\in[m]} \hat{w}_k(X,d)\zeta_0(d \mid X)}{\sum_{d'\in[m]} \hat{w}_k(X,d')\hat{\zeta}_k(d' \mid X)} \\ &\qquad \sum_{a\in\{0,1\}} \frac{(2a-1)(e_0 - \hat{e}_k)(a \mid X,d)}{\hat{e}_k(a \mid X,d)} (\mu_0 - \hat{\mu}_k)(a,X,d) \Bigg] \\ &\quad + P_0\Bigg(\frac{\{1-\pi_0(X)\}\hat{\pi}_k(X)}{\hat{\alpha}_k\{1-\hat{\pi}_k(X)\}} \frac{\sum_{d\in[m]} \hat{w}_k(X,d)\zeta_0(d \mid X)}{\sum_{d'\in[m]} \hat{w}_k(X,d')\hat{\zeta}_k(d' \mid X)} \\ &\qquad [\delta_0(X) - \{\hat{\mu}_k(1,X,d) - \hat{\mu}_k(0,X,d)\}] \Bigg). \end{aligned} \tag{S17}$$

Continuing from the last expression,

$$\begin{aligned} \text{(S17)} = {} & P_0\left[\frac{(\hat{\pi}_k - \pi_0)(X)}{\hat{\alpha}_k\{1-\hat{\pi}_k(X)\}} \frac{\sum_{d\in[m]} \hat{w}_k(X,d)\zeta_0(d \mid X)}{\sum_{d'\in[m]} \hat{w}_k(X,d')\hat{\zeta}_k(d' \mid X)} \sum_{a\in\{0,1\}} (\mu_0 - \hat{\mu}_k)(a,X,d) \right] \\ & + P_0\left(\frac{\pi_0(X)}{\hat{\alpha}_k} \frac{\sum_{d\in[m]} \hat{w}_k(X,d)(\zeta_0 - \hat{\zeta}_k)(d \mid X)}{\sum_{d'\in[m]} \hat{w}_k(X,d')\hat{\zeta}_k(d' \mid X)} \sum_{a\in\{0,1\}} (\mu_0 - \hat{\mu}_k)(a,X,d) \right) \\ & + P_0\Bigg(\frac{\pi_0(X)}{\hat{\alpha}_k} \Bigg[\delta_0(X) - \underbrace{\frac{\sum_{d\in[m]} \hat{w}_k(X,d)\hat{\zeta}_k(d \mid X)}{\sum_{d'\in[m]} \hat{w}_k(X,d')\hat{\zeta}_k(d' \mid X)} \{\hat{\mu}_k(1,X,d) - \hat{\mu}_k(0,X,d)\}}_{=\hat{\delta}_k(X)} \Bigg] \Bigg). \end{aligned}$$

Collecting all relevant terms,

$$
\begin{aligned}
&P_0\ell_{\hat{\eta}_k} - \frac{\alpha_0}{\hat{\alpha}_k}\theta_0 \\
&\quad = P_0\Bigg[\frac{\{1-\pi_0(X)\}\hat{\pi}_k(X)}{\hat{\alpha}_k\{1-\hat{\pi}_k(X)\}}\frac{\sum_{d\in[m]}\hat{w}_k(X,d)\zeta_0(d\mid X)}{\sum_{d'\in[m]}\hat{w}_k(X,d')\hat{\zeta}_k(d'\mid X)} \\
&\qquad\qquad \sum_{a\in\{0,1\}}\frac{(2a-1)(e_0-\hat{e}_k)(a\mid X,d)}{\hat{e}_k(a\mid X,d)}(\mu_0-\hat{\mu}_k)(a,X,d)\Bigg] \\
&\qquad + P_0\Bigg(\frac{\pi_0(X)}{\hat{\alpha}_k}\frac{\sum_{d\in[m]}\hat{w}_k(X,d)(\zeta_0-\hat{\zeta}_k)(d\mid X)}{\sum_{d'\in[m]}\hat{w}_k(X,d')\hat{\zeta}_k(d'\mid X)}\sum_{a\in\{0,1\}}(\mu_0-\hat{\mu}_k)(a,X,d)\Bigg) \\
&\qquad + P_0\Bigg(\frac{(\hat{\pi}_k-\pi_0)(X)}{\hat{\alpha}_k\{1-\hat{\pi}_k(X)\}}\frac{\sum_{d\in[m]}\hat{w}_k(X,d)\zeta_0(d\mid X)}{\sum_{d'\in[m]}\hat{w}_k(X,d')\hat{\zeta}_k(d'\mid X)}\sum_{a\in\{0,1\}}(\mu_0-\hat{\mu}_k)(a,X,d)\Bigg).
\end{aligned}
$$

The representation of the second-order remainder above can be bounded as

$$
\begin{aligned}
&\left|P_0\ell_{\hat{\eta}_k} - \frac{\alpha_0}{\hat{\alpha}_k}\theta_0\right| \\
&\quad \lesssim \max_{d\in[m]}\max_{a\in\{0,1\}} P_0\{|(e_0-\hat{e}_k)(a\mid X,d)||(\mu_0-\hat{\mu}_k)(a,X,d)|\} \\
&\qquad + \max_{d\in[m]}\max_{a\in\{0,1\}} P_0\{|(\zeta_0-\hat{\zeta}_k)(d\mid X)||(\mu_0-\hat{\mu}_k)(a,X,d)|\} \\
&\qquad + \max_{d\in[m]}\max_{a\in\{0,1\}} P_0\{|(\hat{\pi}_k-\pi_0)(X)||(\mu_0-\hat{\mu}_k)(a,X,d)|\} \\
&\quad \leqslant \max_{d\in[m]}\max_{a\in\{0,1\}} \|(\hat{\mu}_k-\mu_0)(a,X,d)I\{\zeta_0(d\mid X)>0\}\|_{P_0} \\
&\qquad \Bigg\{\max_{d\in[m]}\max_{a\in\{0,1\}} \|(\hat{e}_k-e_0)(a\mid X,d)I\{\zeta_0(d\mid X)>0\}\|_{P_0} \\
&\qquad\qquad + \|(\hat{\pi}_k-\pi_0)(X)\|_{P_0} + \max_{d\in[m]}\|(\hat{\zeta}_k-\zeta_0)(d\mid X)\|_{P_0}\Bigg\}.
\end{aligned}
$$

We first show consistency of $\hat{\theta}$. The estimation error is bounded by

$$
|\hat{\theta}-\theta_0| \leqslant \frac{1}{K}\sum_{k\in[K]}\left\{|(\mathbb{P}_{n,k}-P_0)\ell_{\hat{\eta}_k}| + \left|P_0\ell_{\hat{\eta}_k} - \frac{\alpha_0}{\hat{\alpha}_k}\theta_0\right| + \frac{|\hat{\alpha}_k-\alpha_0|}{\hat{\alpha}_k}\theta_0\right\}.
$$

Since the number of splits does not scale with n, we focus on the terms in the braces for every $k \in [K]$. The first term converges in probability to zero by Lemmas S2–S3. The second term converges in probability to zero by Assumption 5a. The third term converges in probability to zero by an application of Slutsky's theorem because $\hat{\alpha}_k \xrightarrow{\text{p}} \alpha_0$. Therefore, we have $\hat{\theta} \xrightarrow{\text{p}} \theta_0$.

To show asymptotic linearly, we further decompose (S15) as

$$
\hat{\theta}-\theta_0 = \mathbb{P}_n\varphi_{\bar{w}} + \frac{1}{K}\sum_{k\in[K]}\left\{(\mathbb{P}_{n,k}-P_0)(\ell_{\hat{\eta}_k}-\ell_{\bar{\eta}}) + \left(P_0\ell_{\hat{\eta}_k} - \frac{\alpha_0}{\hat{\alpha}_k}\theta_0\right) + \frac{(\hat{\alpha}_k-\alpha_0)^2}{\hat{\alpha}_k\alpha_0}\theta_0\right\}.
$$

The second term is $o_{P_0}(n^{-1/2})$ by Lemmas S2 and S4. The third term is $o_{P_0}(n^{-1/2})$ by Assumption 5b. By the central limit theorem, $\hat{\alpha}_k - \alpha_0 = O_{P_0}(n^{-1/2})$, and the last term is $O_P(n^{-1}) = o_{P_0}(n^{-1/2})$ by Slutsky's theorem. □

Remark S1. When $\hat{\mu}_k(1,x,d) - \hat{\mu}_k(0,x,d) = \hat{\mu}_k(1,x,d') - \hat{\mu}_k(0,x,d')$ and $\hat{w}_k(x,d) = \hat{w}_k(x,d')$ for all $d, d' \in [m]$, the error from nuisance model estimation no longer involves

the product term $\max_{d\in[m]} \|(\hat{\zeta}_k - \zeta_0)(d \mid X)\|_{P_0} \max_{a\in\{0,1\}} \|(\hat{\mu}_k - \mu_0)(a, X, d) I\{\zeta_0(d \mid X) > 0\}\|_{P_0}$.

S2.5. *Proof of Proposition 1*

The tangent space at $P_0 \in \mathcal{P}^\dagger$ is

$$\dot{\mathcal{P}}^\dagger = \dot{\mathcal{P}}_y^\dagger \oplus \dot{\mathcal{P}}_a \oplus \dot{\mathcal{P}}_d \oplus \dot{\mathcal{P}}_g \oplus \dot{\mathcal{P}}_x,$$

where $\dot{\mathcal{P}}_\bullet$, $\bullet \in \{a, d, g, x\}$, are as in the proof of Lemma 2, and

$$\dot{\mathcal{P}}_y^\dagger = \big\{(1-g)h(y,a,x,d) : E_0\{h(Y,A,X,D) \mid A,X,D\} = 0, \\ E_0\{Yh(Y,A,X,D) \mid A,X,D=d\} = E_0\{Yh(Y,A,X,D) \mid A,X,D=d'\}\big\}.$$

Along the parametric submodel $\{P_\varepsilon\}$ with score function $h(o)$, the pathwise derivative of

$$\theta_\varepsilon = E_{P_\varepsilon}\{\mu_\varepsilon(1,X) - \mu_\varepsilon(0,X) \mid D = 1\}$$

at $\varepsilon = 0$ is

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_\varepsilon\bigg|_{\varepsilon=0} = \frac{1}{\alpha_0}\int \{\upsilon_h(1,x) - \upsilon_h(0,x)\}\pi_0(x)p_0(x)\mathrm{d}x \\ + \frac{1}{\alpha_0}\int \{\mu(1,x) - \mu(0,x) - \theta_0\}\pi_0(x)p_0(x)\{h(G=1,x) + h(x)\}\mathrm{d}x, \quad \text{(S18)}$$

where $\upsilon_h(a,x) = E_0\{Yh(Y,A,X,D) \mid A = a, X = x, D = d\}$ which does not depend on $d \in \mathcal{D}_x$. We will verify that $\varphi^\dagger$ is an influence function by showing that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_\varepsilon\bigg|_{\varepsilon=0} = E_0\{\varphi^\dagger(O)h(O)\}.$$

For any $h(o) \in \dot{\mathcal{P}}^\dagger$, the inner product

$$E_0\{\varphi^\dagger(O)h(O)\} \\ = E_0\bigg[\frac{(1-G)\pi_0(X)}{\alpha_0\{1-\pi_0(X)\}}\frac{2A-1}{e_0(A \mid X)}\frac{w_0^\dagger(A,X,D)}{\sum_{d\in[m]} w_0^\dagger(A,X,d)\zeta_0(d \mid A,X)} \quad \text{(S19)} \\ \{Y - \mu_0(A,X)\}h(Y,A,X,D)\bigg] \\ + E_0\bigg[\frac{G}{\alpha_0}\{\delta_0(X) - \theta_0\}\{h(G,X) + h(X)\}\bigg]. \quad \text{(S20)}$$

The second term in the display above is equal to the second term in the derivative (S18). The first term is

$$\text{(S19)} = E_0\bigg[\frac{(1-G)\pi_0(X)}{\alpha_0\{1-\pi_0(X)\}}\frac{2A-1}{e_0(A \mid X)}\frac{w_0^\dagger(A,X,D)}{\sum_{d\in[m]} w_0^\dagger(A,X,d)\zeta_0(d \mid A,X)} \\ E_0\{Yh(Y,A,X,D) \mid A,X,D\}\bigg] \\ = E_0\bigg[\frac{(1-G)\pi_0(X)}{\alpha_0\{1-\pi_0(X)\}}\frac{2A-1}{e_0(A \mid X)}\frac{w_0^\dagger(A,X,D)}{\sum_{d\in[m]} w_0^\dagger(A,X,d)\zeta_0(d \mid A,X)}\upsilon_h(A,X)\bigg]$$

$$= E_0\left[\frac{(1-G)\pi_0(X)}{\alpha_0\{1-\pi_0(X)\}}\frac{2A-1}{e_0(A\,|\,X)}\upsilon_h(A,X)\right]$$
$$= E_0\left[\frac{(1-G)\pi_0(X)}{\alpha_0\{1-\pi_0(X)\}}\{\upsilon_h(1,X)-\upsilon_h(0,X)\}\right]$$
$$= \frac{1}{\alpha_0}\int\{\upsilon_h(1,x)-\upsilon_h(0,x)\}\pi_0(x)p_0(x)\mathrm{d}x.$$

Therefore, $\varphi^\dagger$ is a gradient of θ_0. To show that it is the efficient influence function, we check that $\varphi^\dagger \in \dot{\mathcal{P}}^\dagger$. This amounts to verifying that

$$h^*(y,a,x,d) = \frac{1}{\alpha_0}\frac{\pi_0(x)}{1-\pi_0(x)}\frac{2a-1}{e_0(a\,|\,x)}\frac{w_0^\dagger(a,x,d)}{\sum_{d'\in[m]} w_0^\dagger(a,x,d')\zeta(d'\,|\,a,x)}\{y-\mu_0(a,x)\}$$

satisfies $E_0\{Yh^*(Y,A,X,D)\,|\,A,X,D=d\} = E_0\{Yh^*(Y,A,X,D)\,|\,A,X,D=d'\}$. By direct calculation,

$$\begin{aligned}
&E_0\{Yh^*(Y,A,X,D)\,|\,A=a,X=x,D=d\}\\
&\quad= \frac{1}{\alpha_0}E_0\Bigg[\frac{\pi_0(X)}{1-\pi_0(X)}\frac{2A-1}{e_0(A\,|\,X)}\frac{w_0^\dagger(A,X,D)}{\sum_{d'\in[m]} w_0^\dagger(A,X,d')\zeta_0(d'\,|\,A,X)}\\
&\qquad\qquad\{Y-\mu_0(A,X)\}Y\;\Big|\;A=a,X=x,D=d\Bigg]\\
&\quad= \frac{1}{\alpha_0}E_0\Bigg[\frac{\pi_0(X)}{1-\pi_0(X)}\frac{2A-1}{e_0(A\,|\,X)}\frac{w_0^\dagger(A,X,D)}{\sum_{d'\in[m]} w_0^\dagger(A,X,d')\zeta_0(d'\,|\,A,X)}\\
&\qquad\qquad V_0(A,X,D)\;\Big|\;A=a,X=x,D=d\Bigg]\\
&\quad= \frac{1}{\alpha_0}\frac{\pi_0(x)}{1-\pi_0(x)}\frac{2a-1}{\sum_{d'\in[m]} w_0^\dagger(a,x,d')\zeta_0(d'\,|\,a,x)},
\end{aligned}$$

which is constant in d. This observation concludes the proof of the first part.

For the second part, we follow the arguments in the proof of Lemma S1. Define

$$\tilde{\Lambda}^\dagger = \left\{(1-g)\frac{(2a-1)}{e_0(a\,|\,x)}r(a,x,d)\{y-\mu_0(a,x)\} : r(a,x,d)\text{ arbitrary}\right\} \subset L_2^0(P_0).$$

The projection of any

$$\tilde{\lambda}_r = (1-g)\frac{(2a-1)}{e_0(a\,|\,x)}r(a,x,d)\{y-\mu_0(a,x)\} \in \tilde{\Lambda}^\dagger$$

onto $\Lambda^\dagger$ is

$$\begin{aligned}
\Pi\{\tilde{\lambda}_r\,|\,\Lambda^\dagger\} = (1-g)\frac{(2a-1)}{e_0(a\,|\,x)}\Bigg[&r(a,x,d)\\
&- w_0^\dagger(a,x,d)\frac{E_0\{r(A,X,D)\,|\,A=a,X=x,G=0\}}{E_0\{w_0^\dagger(A,X,D)\,|\,A=a,X=x,G=0\}}\Bigg]\{y-\mu_0(a,x)\}.
\end{aligned}$$

Take any $\ell_h(y,a,x,d,g) = (1-g)h(y,a,x,d) \in \tilde{\mathcal{P}}_y$, its projection onto $\tilde{\Lambda}^\dagger$ is

$$\Pi\{\ell_h\,|\,\tilde{\Lambda}^\dagger\} = (1-g)\frac{(2a-1)}{e_0(a\,|\,x)}r_h(a,x,d)\{y-\mu_0(a,x)\},$$

where

$$r_h(a,x,d) = (2a-1)w_0^\dagger(a,x,d)E_0\{Yh(Y,A,X,D) \mid A=a, X=x, D=d\}.$$

Hence,

$$\Pi\{\ell_h \mid \Lambda^\dagger\} = (1-g)\frac{(2a-1)}{e_0(a \mid x)}q_h(a,x,d)\{y-\mu_0(a,x)\},$$

where

$$q_h(a,x,d) = r_h(a,x,d) - w_0^\dagger(a,x,d)\frac{E_0\{r_h(A,X,D) \mid A=a, X=x, G=0\}}{E_0\{w_0^\dagger(A,X,D) \mid A=a, X=x, G=0\}}.$$

Now $\ell_h - \Pi\{\ell_h \mid \Lambda^\dagger\} = (1-g)s_h(y,a,x,d)$, where

$$s_h(y,a,x,d) = h(y,a,x,d) - \frac{(2a-1)}{e_0(a \mid x)}q_h(a,x,d)\{y-\mu_0(a,x)\}.$$

We obviously have $E_0\{s_h(Y,A,X,D) \mid A=a, X=x, D=d\} = 0$, and

$$\begin{aligned}
&E_0\{Ys_h(Y,A,X,D) \mid A=a, X=x, D=d\} \\
&\quad = E_0\{Yh(Y,A,X,D) \mid A=a, X=x, D=d\} - \frac{2a-1}{e_0(a \mid x)}q_h(a,x,d)V_0(a,x,d) \\
&\quad = \frac{E_0\{r_h(A,X,D) \mid A=a, X=x, G=0\}}{E_0\{w_0^\dagger(A,X,D) \mid A=a, X=x, G=0\}},
\end{aligned}$$

which does not depend on d. Therefore, $\Pi\{\ell_h \mid (\Lambda^\dagger)^\perp\} \in \dot{\mathcal{P}}_y^\dagger$. Together with the decomposition $\ell_h = \ell_h - \Pi\{\ell_h \mid \Lambda^\dagger\} + \Pi\{\ell_h \mid \Lambda^\dagger\}$, we conclude that $\dot{\mathcal{P}}^\dagger = (\Lambda^\dagger)^\perp$.

S2.6. *Proof of Proposition 2*

Consider the parametric submodel $\{P_\varepsilon : \varepsilon \in \Gamma\}$ with score function $h(o)$. Suppose

$$\beta_\varepsilon \in \arg\min_{\tilde{\beta}\in\mathbb{R}^q} E_{P_\varepsilon}[\{\delta_\varepsilon(X) - \tilde{\beta}^{\mathrm{T}}\psi(Z)\}^2 \mid G=1].$$

Then β_ε must fulfill the first-order condition

$$E_{P_\varepsilon}\left[\psi(Z)\{\delta_\varepsilon(X) - \beta_\varepsilon^{\mathrm{T}}\psi(Z)\} \;\middle|\; G=1\right] = 0.$$

By the implicit function theorem, there exists a function β_ε of ε such that $\beta_\varepsilon|_{\varepsilon=0} = \beta_0$ and that it is differentiable at $\varepsilon = 0$ with derivative

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\beta_\varepsilon\bigg|_{\varepsilon=0} = \left[E_0\{\psi^{\otimes 2}(Z) \mid G=1\}\right]^{-1}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}E_{P_\varepsilon}\left[\psi(Z)\{\delta_\varepsilon(X) - \beta_0^{\mathrm{T}}\psi(Z)\} \;\middle|\; G=1\right]\bigg|_{\varepsilon=0}.$$

The uniqueness of β_0 is ensured by Assumption 7. The Gateaux derivative

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon}E_{P_\varepsilon}\left[\psi(Z)\{\delta_\varepsilon(X) - \beta_0^{\mathrm{T}}\psi(Z)\} \;\middle|\; G=1\right]\Big|_{\varepsilon=0} \\
&\quad = \frac{1}{\alpha_0}E_0\left[G\psi(Z)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\delta_\varepsilon(X)\bigg|_{\varepsilon=0}\right] + \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\frac{E_{P_\varepsilon}\left[G\psi(Z)\{\delta_0(X) - \beta_0^{\mathrm{T}}\psi(Z)\}\right]}{P_\varepsilon(G=1)}\bigg|_{\varepsilon=0} \\
&\quad = \frac{1}{\alpha_0}\int \psi(z)\kappa_h(x)\{1-\pi_0(x)\}p_0(x)\mathrm{d}x \\
&\qquad + \frac{1}{\alpha_0}\int \Big(\psi(z)\{\delta_0(x) - \beta_0^{\mathrm{T}}\psi(z)\} - \underbrace{E_0[\psi(Z)\{\delta_0(X) - \beta_0^{\mathrm{T}}\psi(Z)\} \mid G=1]}_{=0}\Big)
\end{aligned}$$

$$\pi_0(x)p_0(x)\{h(G=1,x)+h(x)\}\mathrm{d}x$$
$$= E_0\Bigg(\Bigg[\frac{1-G}{\alpha_0}\psi(Z)\frac{\pi_0(X)}{1-\pi_0(X)}\frac{w_0(X,D)}{\sum_{d\in[m]}\zeta_0(d\mid X)w_0(X,d)}\frac{2A-1}{e_0(A\mid X,D)}$$
$$\{Y-\mu_0(A,X,D)\}+\frac{G}{\alpha_0}\{\delta_0(X)-\beta_0^{\mathrm{T}}\psi(Z)\}\Bigg]h(O)\Bigg).$$

This shows that φ_{β_0} is a gradient of β_0. Since $\varphi_{\beta_0}\in\dot{\mathcal{P}}$, it is the efficient influence function of β_0.

S2.7. *Proof of Theorem 2*

We first show consistency of $\hat{\beta}$. Define

$$\hat{\Psi}=\frac{1}{n_1}\sum_{i:G_i=1}\psi^{\otimes 2}(Z_i),\quad \Psi_0=E_0\{\psi^{\otimes 2}(Z)\mid G=1\}.$$

Decompose the error by

$$\begin{aligned}\hat{\beta}-\beta_0&=\frac{1}{K}\sum_{k\in[K]}\hat{\Psi}^{-1}\mathbb{P}_{n,k}\psi\ell_{\hat{\eta}_k}-\beta_0\\&=\frac{1}{K}\sum_{k\in[K]}\hat{\Psi}^{-1}(\mathbb{P}_{n,k}-P_0)\psi\ell_{\hat{\eta}_k}\\&\quad+\frac{1}{K}\sum_{k\in[K]}\hat{\Psi}^{-1}\Big[P_0\psi\ell_{\hat{\eta}_k}-\frac{\alpha_0}{\hat{\alpha}_k}\Psi_0\beta_0\Big]\\&\quad+\frac{1}{K}\sum_{k\in[K]}\Big[\frac{\alpha_0}{\hat{\alpha}_k}\hat{\Psi}^{-1}\Psi_0-\mathrm{Id}\Big]\beta_0.\end{aligned}\tag{S21}$$

By the law of large numbers, every entry of the matrix $\hat{\Psi}_{jk}$ converges in probability to the corresponding entry in $(\Psi_0)_{jk}$ for $j,k\in[q]$. Since the dimension q is finite, we have

$$\|\hat{\Psi}-\Psi_0\|\leqslant\Big\{\sum_{j,k\in[q]}|\hat{\Psi}_{jk}-(\Psi_0)_{jk}|^2\Big\}^{1/2}\xrightarrow{\mathrm{p}}0.$$

By the continuous mapping theorem, $\|\hat{\Psi}^{-1}-\Psi_0^{-1}\|=o_{P_0}(1)$. In the following, we combine the matrix-norm convergence with vector-norm convergence using Slutsky's theorem, since $\|A_nv_n\|=o_{P_0}(1)$ if $\|A_n\|\|v_n\|=o_{P_0}(1)$. The first term of (S21) converges in probability to 0 due to the boundedness of $|\psi_j(z)|$ for $j\in[q]$ and Lemmas S2–S3. Consider the third term of (S21). It converges in probability to 0 because $\hat{\alpha}_k\xrightarrow{\mathrm{p}}\alpha_0$, and by the continuous mapping theorem $\|(\alpha_0/\hat{\alpha}_k)\hat{\Psi}^{-1}\Psi_0-\mathrm{Id}\|\xrightarrow{\mathrm{p}}0$. The second term also converges in probability to 0 because

$$\begin{aligned}&\Big\|P_0\psi\ell_{\hat{\eta}_k}-\frac{\alpha_0}{\hat{\alpha}_k}\Psi_0\beta_0\Big\|\\&\quad=\Big\|P_0\psi(Z)\Big\{\ell_{\hat{\eta}_k}(O)-\frac{G}{\hat{\alpha}_k}\delta_0(X)+\frac{G}{\hat{\alpha}_k}\delta_0(X)-\frac{G}{\hat{\alpha}_k}\psi^{\mathrm{T}}(Z)\beta_0\Big\}\Big\|\\&\quad=\Big\|P_0\psi(Z)\Big\{\ell_{\hat{\eta}_k}(O)-\frac{G}{\hat{\alpha}_k}\delta_0(X)\Big\}\Big\|,\end{aligned}$$

which can be seen to be $o_{P_0}(1)$ by modifying the steps of bounding $|P_0\ell_{\hat{\eta}_k}-(\alpha_0/\hat{\alpha}_k)\theta_0|$ in the proof of Theorem 1 and by using that $|\psi|\leqslant C$. The consistency of $\hat{\beta}$ is established.

To show asymptotic linearity, we further make the following decomposition:

$$\begin{aligned}\hat{\beta} - \beta_0 = (\mathbb{P}_n - P_0)\varphi_{\vec{w}}^{\beta} \\ &+ (\hat{\Psi}^{-1} - \Psi_0^{-1})(\mathbb{P}_n - P_0)\psi(Z)\left\{\ell_{\bar{\eta}}(O) - \frac{G}{\alpha_0}\psi^{\mathrm{T}}(Z)\beta_0\right\} \\ &+ \frac{1}{K}\sum_{k\in[K]} \hat{\Psi}^{-1}(\mathbb{P}_{n,k} - P_0)\psi(\ell_{\hat{\eta}_k} - \ell_{\bar{\eta}}) \\ &+ \frac{1}{K}\sum_{k\in[K]} \hat{\Psi}^{-1}\left[P_0\psi\ell_{\hat{\eta}_k} - \frac{\alpha_0}{\hat{\alpha}_k}\Psi_0\beta_0\right] \\ &+ \frac{1}{K}\sum_{k\in[K]} \frac{\alpha_0 - \hat{\alpha}_k}{\hat{\alpha}_k}(\hat{\Psi}^{-1}\Psi_0 - \mathrm{Id})\beta_0 \\ &+ \frac{1}{K}\sum_{k\in[K]} \frac{(\hat{\alpha}_k - \alpha_0)^2}{\alpha_0\hat{\alpha}_k}\beta_0. \end{aligned} \tag{S22}$$

Following the arguments in the proof of Theorem 1, all terms of (S22) are $o_{P_0}(n^{-1/2})$.

S2.8. *Proof of Corollary 2*

Lemma S5. *Suppose Assumption 4 holds. Then* $\hat{\Omega} \xrightarrow{\mathrm{p}} \bar{\Omega}$ *if* $\hat{\theta} \xrightarrow{\mathrm{p}} \theta_0$.

Proof. Since $\hat{\Omega} = \sum_{k\in[K]} \hat{\Omega}_k/K$ for a finite number of splits, it suffices to show consistency of every $\hat{\Omega}_k$. We make the following decomposition:

$$\begin{aligned}\hat{\Omega}_k &= \mathbb{P}_{n,k}\left\{\ell_{\hat{\eta}_k} - \frac{G}{\hat{\alpha}_k}\hat{\theta}\right\}^2 \\ &= \mathbb{P}_{n,k}\left[\left\{\ell_{\hat{\eta}_k} - \frac{G}{\hat{\alpha}_k}\hat{\theta}\right\} - \left\{\ell_{\bar{\eta}} - \frac{G}{\alpha_0}\theta_0\right\}\right]^2 \\ &\quad + 2\mathbb{P}_{n,k}\left[\left\{\ell_{\hat{\eta}_k} - \frac{G}{\hat{\alpha}_k}\hat{\theta}\right\} - \left\{\ell_{\bar{\eta}} - \frac{G}{\alpha_0}\theta_0\right\}\right]\left\{\ell_{\bar{\eta}} - \frac{G}{\alpha_0}\theta_0\right\} \\ &\quad + \mathbb{P}_{n,k}\left\{\ell_{\bar{\eta}} - \frac{G}{\alpha_0}\theta_0\right\}^2. \end{aligned} \tag{S23}$$

The third term in (S23) converges in probability to $\bar{\Omega}$ by the law of large numbers. By the Cauchy-Schwarz inequality, the second term in (S23) is bounded by the product of the square root of the first term and $\bar{\Omega}^{1/2}$. It remains to show that the first term in (S23) is $o_{P_0}(1)$. We have

$$\mathbb{P}_{n,k}\left[\left\{\ell_{\hat{\eta}_k} - \frac{G}{\hat{\alpha}_k}\hat{\theta}\right\} - \left\{\ell_{\bar{\eta}} - \frac{G}{\alpha_0}\theta_0\right\}\right]^2 \leqslant 2\mathbb{P}_{n,k}(\ell_{\hat{\eta}_k} - \ell_{\bar{\eta}})^2 + 2\hat{\alpha}_k\left(\frac{\hat{\theta}}{\hat{\alpha}_k} - \frac{\theta_0}{\alpha_0}\right)^2 \tag{S24}$$

Applying the Markov inequality conditional on the data $\mathcal{O}_{-k} = \{O_i : i \notin \mathcal{I}_k\}$, for any $t > 0$,

$$\mathrm{pr}\left\{\frac{\mathbb{P}_{n,k}(\ell_{\hat{\eta}_k} - \ell_{\bar{\eta}})^2}{\|\ell_{\hat{\eta}_k} - \ell_{\bar{\eta}}\|^2} \geqslant t \;\middle|\; \mathcal{O}_{-k}\right\} \leqslant \frac{1}{t}.$$

Hence, marginally $\mathbb{P}_{n,k}(\ell_{\hat{\eta}_k} - \ell_{\bar{\eta}})^2 = O_{P_0}(\|\ell_{\hat{\eta}_k} - \ell_{\bar{\eta}}\|^2)$, which is $o_{P_0}(1)$ as shown in the proof of Theorem 1. The second term on the righthand side of (S24) is also $o_{P_0}(1)$ from the consistency of $\hat{\alpha}$ and $\hat{\theta}$ followed by an application of Slutsky's theorem. □

Lemma S6. *Suppose Assumptions 4, 7, and 8 hold. Then* $\|\hat{\Omega}_\beta - \bar{\Omega}_\beta\| \xrightarrow{\mathrm{p}} 0$ *if* $\|\hat{\beta} - \beta\| \xrightarrow{\mathrm{p}} 0$.

Proof. From the proof of Theorem 2, we have $\|\hat{\Psi}^{-1} - \Psi_0^{-1}\| = o_{P_0}(1)$. Let

$$\hat{Q}_k = \mathbb{P}_{n,k}\psi^{\otimes 2}(Z)\left\{\ell_{\hat{\eta}_k}(O) - \frac{G}{\hat{\alpha}_k}\hat{\beta}^{\mathrm{T}}\psi(Z)\right\}^2, \quad Q = E_0\left[\psi^{\otimes 2}(Z)\left\{\ell_{\bar{\eta}}(O) - \frac{G}{\alpha_0}\beta_0^{\mathrm{T}}\psi(Z)\right\}^2\right].$$

Then for $j, j' \in [q]$, every entry $(\hat{Q}_k)_{jj'}$ can be decomposed as

$$\begin{aligned}
&(\hat{Q}_k)_{jj'} \\
&\quad = \mathbb{P}_{n,k}\psi_j(Z)\psi_{j'}(Z)\left\{\ell_{\hat{\eta}_k} - \frac{G}{\hat{\alpha}_k}\hat{\beta}^{\mathrm{T}}\psi(Z)\right\}^2 \\
&\quad = \mathbb{P}_{n,k}\psi_j(Z)\psi_{j'}(Z)\left[\left\{\ell_{\hat{\eta}_k} - \frac{G}{\hat{\alpha}_k}\hat{\beta}^{\mathrm{T}}\psi(Z)\right\} - \left\{\ell_{\bar{\eta}} - \frac{G}{\alpha_0}\beta_0^{\mathrm{T}}\psi(Z)\right\}\right]^2 \\
&\qquad + 2\mathbb{P}_{n,k}\psi_j(Z)\psi_{j'}(Z)\left[\left\{\ell_{\hat{\eta}_k} - \frac{G}{\hat{\alpha}_k}\hat{\beta}^{\mathrm{T}}\psi(Z)\right\} - \left\{\ell_{\bar{\eta}} - \frac{G}{\alpha_0}\beta_0^{\mathrm{T}}\psi(Z)\right\}\right]\left\{\ell_{\bar{\eta}} - \frac{G}{\alpha_0}\beta_0^{\mathrm{T}}\psi(Z)\right\} \\
&\qquad + \mathbb{P}_{n,k}\psi_j(Z)\psi_{j'}(Z)\left\{\ell_{\bar{\eta}} - \frac{G}{\alpha_0}\beta_0^{\mathrm{T}}\psi(Z)\right\}^2.
\end{aligned}$$

Following the proof of Lemma S5, we can show that $|\hat{Q}_{jj'} - Q_{jj'}| = o_{P_0}(1)$ if $\|\hat{\beta} - \beta\| = o_{P_0}(1)$, because $\sup_{z\in\mathcal{Z}_1}|\psi_j(z)| \leqslant C$. Since the dimension q does not depend on n, we have $\|\hat{Q} - Q\| = o_{P_0}(1)$. The consistency of $(\hat{\Omega}_\beta)_k$ can now be established by the continuous mapping theorem, hence the consistency of $\hat{\Omega}_\beta$. □

Proof of Corollary 2. In the following, we treat $z \in \mathcal{Z}_1$ as an indexing parameter. Let $\bar{\omega}_\beta^2(z) = \psi^{\mathrm{T}}(z)\bar{\Omega}_\beta\psi(z)$ and $\hat{\omega}_\beta^2(z) = \psi^{\mathrm{T}}(z)\hat{\Omega}_\beta\psi(z)$. Define three stochastic processes

$$\begin{aligned}
\hat{\mathbb{T}}_n(z) &= \frac{n^{1/2}\psi^{\mathrm{T}}(z)(\hat{\beta} - \beta_0)}{\hat{\omega}_\beta(z)}, \\
\tilde{\mathbb{T}}_n(z) &= \frac{n^{1/2}\psi^{\mathrm{T}}(z)(\hat{\beta} - \beta_0)}{\bar{\omega}_\beta(z)}, \\
\mathbb{T}_n(z) &= \frac{\mathbb{G}_n\psi^{\mathrm{T}}(z)\varphi_{\bar{w}}^\beta}{\bar{\omega}_\beta(z)}
\end{aligned}$$

Define the function class

$$\mathcal{F} = \left\{\frac{\psi^{\mathrm{T}}(z)\varphi_{\bar{w}}^\beta}{\bar{\omega}_\beta(z)} : z \in \mathcal{Z}_1\right\}.$$

By Assumption 8, $\bar{\omega}_\beta^{-1}(z) \leqslant \lambda_{\min}^{-1/2}(\bar{\Omega}_\beta)\|\psi(z)\|^{-1} \lesssim 1$ and $\|\bar{\Omega}_\beta\psi^{\otimes 2}(z)\| \leqslant \lambda_{\max}(\bar{\Omega}_\beta)\|\psi(z)\|^2 \lesssim 1$. The partial derivative satisfies

$$\frac{\partial}{\partial\psi(z)}\frac{\psi^{\mathrm{T}}(z)\varphi_{\bar{w}}^\beta}{\bar{\omega}_\beta(z)} \lesssim \|\varphi_{\bar{w}}^\beta\|.$$

Applying Theorems 2.7.17 and 2.5.6 from van der Vaart and Wellner (2023) in this order, we see that $\mathcal{F}$ is P_0-Donsker, since

$$\left|\frac{\psi^{\mathrm{T}}(z)\varphi_{\bar{w}}^\beta}{\bar{\omega}_\beta(z)} - \frac{\psi^{\mathrm{T}}(z')\varphi_{\bar{w}}^\beta}{\bar{\omega}_\beta(z')}\right| \lesssim \|\varphi_{\bar{w}}^\beta\|\|\psi(z) - \psi(z')\|$$

and $\{\psi(z) : C^{-1} \leqslant \|\psi(z)\| \leqslant C, z \in \mathcal{Z}_1\}$ represents a compact set in $\mathbb{R}^q$. We have

$$\mathbb{T}_n \rightsquigarrow \mathbb{T}, \qquad \text{in } \ell^\infty(\mathcal{Z}_1),$$

where $\mathbb{T}$ is the mean-zero Gaussian process stated in the corollary.

We now show that $\|\hat{\mathbb{T}}_n - \mathbb{T}_n\|_{\mathcal{Z}_1} = o_{P_0}(1)$, which by the continuous mapping theorem (van der Vaart and Wellner, 2023, Theorem 1.3.6), proves the corollary combined with the weak convergence above. The standard error estimator is uniformly consistent, because

$$\begin{aligned}
\left\| \frac{\hat{\omega}_\beta}{\bar{\omega}_\beta} - 1 \right\|_{\mathcal{Z}_1} &\leqslant \left\| \frac{\hat{\omega}_\beta^2}{\bar{\omega}_\beta^2} - 1 \right\|_{\mathcal{Z}_1} \\
&\leqslant \sup_{z \in \mathcal{Z}_1} \left| \frac{\psi^{\mathrm{T}}(z)(\hat{\Omega}_\beta - \bar{\Omega}_\beta)\psi(z)}{\psi^{\mathrm{T}}(z)\bar{\Omega}_\beta\psi(z)} \right| \\
&\leqslant \lambda_{\min}^{-1}(\bar{\Omega}_\beta)\|\hat{\Omega}_\beta - \bar{\Omega}_\beta\| \xrightarrow{\mathrm{p}} 0,
\end{aligned}$$

which follows from the consistency of $\hat{\Omega}_\beta$ in Lemma S6. We bound the difference of the two stochastic processes by

$$\begin{aligned}
\|\hat{\mathbb{T}}_n - \mathbb{T}_n\|_{\mathcal{Z}_1} &= \left\| (\tilde{\mathbb{T}}_n - \mathbb{T}_n)\frac{\bar{\omega}_\beta}{\hat{\omega}_\beta} + \mathbb{T}_n\left(\frac{\bar{\omega}_\beta}{\hat{\omega}_\beta} - 1\right) \right\|_{\mathcal{Z}_1} \\
&\leqslant \|\tilde{\mathbb{T}}_n - \mathbb{T}_n\|_{\mathcal{Z}_1} \left\| \frac{\bar{\omega}_\beta}{\hat{\omega}_\beta} \right\|_{\mathcal{Z}_1} + \|\mathbb{T}_n\|_{\mathcal{Z}_1} \left\| \frac{\bar{\omega}_\beta}{\hat{\omega}_\beta} - 1 \right\|_{\mathcal{Z}_1}.
\end{aligned}$$

Since $\|\bar{\omega}_\beta/\hat{\omega}_\beta\|_{\mathcal{Z}_1} = 1 + o_{P_0}(1) = O_{P_0}(1)$ and $\|\mathbb{T}_n\|_{\mathcal{Z}_1} = O_{P_0}(1)$ from Theorem 2.14.2 in van der Vaart and Wellner (2023), it remains to show $\|\tilde{\mathbb{T}}_n - \mathbb{T}_n\|_{\mathcal{Z}_1} = o_{P_0}(1)$. Clearly,

$$\begin{aligned}
\tilde{\mathbb{T}}_n(z) - \mathbb{T}_n(z) &= \frac{n^{1/2}\psi^{\mathrm{T}}(z)(\hat{\beta} - \beta_0)}{\bar{\omega}_\beta(z)} - \frac{\mathbb{G}_n\psi^{\mathrm{T}}(z)\varphi_{\tilde{w}}^\beta}{\bar{\omega}_\beta(z)} \\
&= \frac{n^{1/2}\psi^{\mathrm{T}}(z)}{\bar{\omega}_\beta(z)} R_n,
\end{aligned}$$

where we have used (S22) from the proof of Theorem 2, and R_n denotes the $o_{P_0}(n^{-1/2})$ terms in (S22). Since $\|\psi(z)\| \lesssim 1$ and $\bar{\omega}_\beta^{-1}(z) \leqslant \lambda_{\min}^{-1/2}(\bar{\Omega}_\beta)\|\psi(z)\|^{-1} \lesssim 1$, we have $\|\tilde{\mathbb{T}}_n - \mathbb{T}_n\|_{\mathcal{Z}_1} = o_{P_0}(1)$. □

S3. Subsidiary results

S3.1. *Asymptotic variance reduction under difference transportability*

We present details for the difference of asymptotic variances when $\tilde{w}(x, d) = \tilde{w}(x) \neq 0$ that corresponds to the discussion in §3 in the main text.

By direct calculation, the difference is

$$\begin{aligned}
&E_0(\varphi_{w_0}^2 - \varphi_{\tilde{w}}^2) \\
&\quad = E_0\Bigg(\frac{(1 - G)\{\pi_0(X)\}^2}{\alpha_0^2\{1 - \pi_0(X)\}^2} \frac{1}{\{e_0(A \mid X, D)\}^2} \left[\left\{ \frac{w_0(X, D)}{\sum_{d \in [m]} \zeta_0(d \mid X) w_0(X, d)} \right\}^2 - 1 \right] \\
&\qquad\qquad \{Y - \mu_0(A, X, D)\}^2 \Bigg)
\end{aligned}$$

$$= E_0\left(\frac{(1-G)\{\pi_0(X)\}^2}{\alpha_0^2\{1-\pi_0(X)\}^2}\frac{V_0(A,X,D)}{\{e_0(A\mid X,D)\}^2}\left[\left\{\frac{w_0(X,D)}{\sum_{d\in[m]}\zeta_0(d\mid X)w_0(X,d)}\right\}^2-1\right]\right)$$

$$= E_0\left(\frac{(1-G)\{\pi_0(X)\}^2}{\alpha_0^2\{1-\pi_0(X)\}^2}\underbrace{\left\{\frac{V_0(1,X,D)}{e_0(1\mid X,D)}+\frac{V_0(0,X,D)}{e_0(0\mid X,D)}\right\}}_{\{w_0(X,D)\}^{-1}}\left[\left\{\frac{w_0(X,D)}{\sum_{d\in[m]}\zeta_0(d\mid X)w_0(X,d)}\right\}^2-1\right]\right)$$

$$= E_0\left(\frac{(1-G)\{\pi_0(X)\}^2}{\alpha_0^2\{1-\pi_0(X)\}^2}\sum_{d\in[m]}\zeta_0(d\mid X)\{w_0(X,d)\}^{-1}\left[\left\{\frac{w_0(X,d)}{\sum_{d'\in[m]}\zeta_0(d'\mid X)w_0(X,d')}\right\}^2-1\right]\right)$$

$$= E_0\left(\frac{\{\pi_0(X)\}^2}{\alpha_0^2\{1-\pi_0(X)\}}\left[\left\{\sum_{d'\in[m]}w_0(X,d')\zeta_0(d'\mid X)\right\}^{-1}-\left\{\sum_{d'\in[m]}\zeta_0(d'\mid X)\{w_0(X,d')\}^{-1}\right\}\right]\right).$$

Alternatively, using Corollary 1, we have that $\varphi_{\tilde{w}}-\varphi_{w_0}$ is orthogonal to the tangent space at P_0 but φ_{w_0} lies in the tangent space, so that $E\{\varphi^2-\tilde{\varphi}^2\}=-E\{\varphi-\tilde{\varphi}\}^2\leqslant 0$ directly by the Pythagorean theorem.

The difference of asymptotic variances for $w_0(x,d)$ and $\tilde{w}(x,d)$ such that $\tilde{w}(x,d)\neq 0$ for some d is

$$E_0(\varphi_{w_0}^2-\varphi_{\tilde{w}}^2)=E_0\left(\frac{\{\pi_0(X)\}^2}{\alpha_0^2\{1-\pi_0(X)\}}\left[\frac{1}{\sum_{d\in[m]}\zeta_0(d\mid X)w_0(X,d)}-\frac{\sum_{d\in[m]}\zeta_0(d\mid X)\{w_0(X,d)\}^{-1}\{\tilde{w}(X,d)\}^2}{\{\sum_{d'\in[m]}\zeta_0(d'\mid X)\tilde{w}(X,d')\}^2}\right]\right).$$

S3.2. *Parametric conditional average treatment effect*

Consider the class of functions of the baseline covariates $\Delta=\{\delta_\beta(x):\beta\in\mathbb{R}^q\}$, where $\delta_\beta(x)$ is a known, smooth function of β, and the true CATE is $\delta_0(x)=\delta_{\beta_0}(x)$ for some unique β_0. Under the transportability assumption, the conditional mean of the outcome can be expressed as $\mu_0(a,x,d)=\mu_0(0,x,d)+a\delta_{\beta_0}(x)$. Define the semiparametric model

$$\mathcal{P}_{\text{sp}}=\{P\in\mathcal{P}:E(Y\mid A=1,X=x,D=d)-E(Y\mid A=0,X=x,D=d)=\delta_\beta(x)\}.$$

Lemma S7. *Suppose there is a universal constant $C>0$ such that $|Y-\mu_0(0,X,D)-A\delta_{\beta_0}(X)|\leqslant C$ under P_0. The efficient score of β_0 under the model $P_0\in\mathcal{P}_{\text{sp}}$ is*

$$\varsigma_{\beta_0}(o)=(1-g)\dot{\delta}_{\beta_0}(x)\left\{a-\frac{\{V_0(1,x,d)\}^{-1}e_0(1\mid x,d)}{\{V_0(1,x,d)\}^{-1}e_0(1\mid x,d)+\{V_0(0,x,d)\}^{-1}e_0(0\mid x,d)}\right\}\{V_0(a,x,d)\}^{-1}\big[y-\{\mu_0(0,x,d)+a\delta_{\beta_0}(x)\}\big]. \quad \text{(S25)}$$

Consequently, the efficient influence function of β_0 is $\varphi_{\beta_0}(o)=[E_0\{\varsigma_{\beta_0}^{\otimes 2}(O)\}]^{-1}\varsigma_{\beta_0}(o)$.

Proof. Define the transformation $U = Y - \{\mu(0, X, D) + A\delta_\beta(X)\}$ if $G = 0$ and denote its density by $p_u(u \mid a, x, d)$. The likelihood of the data can be factorized as

$$\begin{aligned} p_0(o) &= p_0(x)\pi_0(x)^g\{1-\pi_0(x)\}^{1-g}\{\zeta_0(d \mid x)e_0(a \mid x, d)p_0(y \mid a, x, d)\}^{1-g} \\ &= p_0(x)\pi_0(x)^g\{1-\pi_0(x)\}^{1-g}\{\zeta_0(d \mid x)e_0(a \mid x, d)p_{0u}(u \mid a, x, d)\}^{1-g}. \end{aligned}$$

Let $h_u(u, a, x, d) = (\mathrm{d}/\mathrm{d}u)p_{0u}(u \mid a, x, d)/p_{0u}(u \mid a, x, d)$. The nuisance tangent space at $P_0 \in \mathcal{P}_{\text{sp}}$ is

$$\dot{\mathcal{P}}_{\text{sp}} = \dot{\mathcal{P}}_\mu + (\dot{\mathcal{P}}_u \oplus \dot{\mathcal{P}}_a \oplus \dot{\mathcal{P}}_d \oplus \dot{\mathcal{P}}_g \oplus \dot{\mathcal{P}}_x),$$

where $\dot{\mathcal{P}}_a$, $\dot{\mathcal{P}}_d$, $\dot{\mathcal{P}}_g$, and $\dot{\mathcal{P}}_x$ are as defined in the proof of Lemma 2, and

$$\begin{aligned} \dot{\mathcal{P}}_u = \big\{&(1-g)h(u, a, x, d) : E_0\{h(U, A, X, D) \mid A, X, D\} = 0, \\ &E_0\{Uh(U, A, X, D) \mid A, X, D\} = 0\big\}, \\ \dot{\mathcal{P}}_\mu = \big\{&(1-g)h_u(u, a, x, d)h(x, d) : h(x, d) \in L_2(P_0)\big\}. \end{aligned}$$

A standard result of semiparametric regression gives that

$$\dot{\mathcal{P}}'_u = \{(1-g)uh(a, x, d) : h(a, x, d) \in L_2(P_0)\}$$

is the orthogonal complement of the subspace $\dot{\mathcal{P}}_u \oplus \dot{\mathcal{P}}_a \oplus \dot{\mathcal{P}}_d \oplus \dot{\mathcal{P}}_g \oplus \dot{\mathcal{P}}_x$.

Define $\nu(a, x, d) = \{E(u^2 \mid A = a, X = x, D = d)\}^{-1}$. The orthogonal projection of the subspace $\dot{\mathcal{P}}_\mu$ onto $\dot{\mathcal{P}}'_u$ is

$$\Pi\{\dot{\mathcal{P}}_\mu \mid \dot{\mathcal{P}}'_u\} = \{(1-g)h(x, d)\nu(a, x, d)u : h(x, d) \in L_2(P_0)\}.$$

Now we show that this is true. Take an arbitrary element $J(o) = (1-g)h_u(u, a, x, d)h(x, d) \in \dot{\mathcal{P}}_\mu$. We can verify that $J^*(o) = -(1-g)h(x, d)\nu(a, x, d)u$ is indeed the projection $\Pi\{J \mid \dot{\mathcal{P}}'_u\} = J^*(o)$. This is because the difference

$$J(o) - J^*(o) = (1-g)h(x, d)\{h_u(u, a, x, d) + \nu(a, x, d)u\}$$

is orthogonal to $\dot{\mathcal{P}}'_u$, since for arbitrary $h(a, x, d)$,

$$\begin{aligned} &E_0[(1-G)Uh(A, X, D)h(X, D)\{h_u(U, A, X, D) + \nu(A, X, D)U\}] \\ &\quad = E_0[(1-G)h(A, X, D)h(X, D)E\{Uh_u(U, A, X, D) + \nu(A, X, D)U^2 \mid A, X, D\}] \\ &\quad = E_0[(1-G)h(A, X, D)h(X, D)\{(-1) + 1\}] = 0, \end{aligned}$$

where we have used the equality that $E_0\{Uh_u(U, A, X, D) \mid A, X, D\} = -1$, as a result of differentiating the moment restriction $\int up_{0u}(u \mid a, x, d) = 0$ with respect to u.

The score function of β is

$$\sigma_\beta(o) = -(1-g)h_u(u, a, x, d)a\dot{\delta}_\beta(x).$$

The projection is

$$\Pi\{\sigma_\beta \mid \dot{\mathcal{P}}'_u\} = (1-g)E\{\sigma_\beta(O)U \mid A = a, X = x, D = d\}\nu(a, x, d)u.$$

Let $q(a, x, d) = E_0\{\sigma_\beta(O)U \mid A = a, X = x, D = d\}$. We further project $\Pi\{\sigma_\beta \mid \dot{\mathcal{P}}'_u\}$ onto the space $\Pi\{\dot{\mathcal{P}}_\mu \mid \dot{\mathcal{P}}'_u\}$. Assume the projection takes the form

$$\Pi\big[\Pi\{\sigma_\beta \mid \dot{\mathcal{P}}'_u\} \,\big|\, \Pi(\dot{\mathcal{P}}_\mu \mid \dot{\mathcal{P}}'_u)\big] = (1-g)h^*(x, d)\nu(a, x, d)u.$$

The function $h^*(x, d)$ is the solution to the equation

$$E_0\{\nu(A, X, D)U\{q(A, X, D)\nu(A, X, D)U - h^*(X, D)\nu(A, X, D)U\} \mid X, D\} = 0,$$

which yields

$$h^*(x,d) = \frac{E_0\{q(A,X,D)\nu(A,X,D) \mid X=x, D=d\}}{E_0\{\nu(A,X,D) \mid X=x, D=d\}}$$
$$= \frac{q(1,x,d)\nu(1,x,d)e_0(1 \mid x,d) + q(0,x,d)\nu(0,x,d)e_0(0 \mid x,d)}{\nu(1,x,d)e_0(1 \mid x,d) + \nu(0,x,d)e_0(0 \mid x,d)}.$$

The efficient score of β is the projection

$$\begin{aligned}\varsigma_\beta(o) &= \Pi\{\sigma_\beta \mid \dot{\mathcal{P}}_{\text{sp}}^{\perp}\} \\ &= \Pi\{\sigma_\beta \mid \dot{\mathcal{P}}'_u\} - \Pi\left[\Pi\{\sigma_\beta \mid \dot{\mathcal{P}}'_u\} \,\middle|\, \Pi(\dot{\mathcal{P}}_\mu \mid \dot{\mathcal{P}}'_u)\right] \\ &= (1-g)\{q(a,x,d) - h^*(x,d)\}\nu(a,x,d)u.\end{aligned}$$

Using the generalized information equality [Equation (5) in Newey (1990)] or by direct calculation, $E_0\{\sigma_\beta(O)U \mid A,X,D\} = -E_0\{(\partial/\partial\beta)U \mid A,X,D\}$, so that $q(a,x,d) = a\dot{\delta}_\beta(x)$. Then

$$\varsigma_\beta(o) = (1-g)\dot{\delta}_\beta(x)\left\{a - \frac{\nu(1,x,d)e_0(1 \mid x,d)}{\nu(1,x,d)e_0(1 \mid x,d) + \nu(0,x,d)e_0(0 \mid x,d)}\right\}\nu(a,x,d)u.$$

□

Since the semiparametric model $\mathcal{P}_{\text{sp}}$ is induced by a conditional mean restriction, it follows that any score $\bar{\varsigma}_{\beta_0}(o)$ obtained by replacing V_0 with an arbitrary $\bar{V} \neq 0$ is Neyman orthogonal (Chernozhukov et al., 2018). Making explicit its dependence on the nuisance parameters, a score function of β_0 can be represented as $\tilde{\varsigma}_\beta\{o; \bar{\mu}(0,x,d), \bar{V}(a,x,d), \bar{e}(a,x,d)\}$. It suggests a doubly robust estimating equation in the sense that $P_0\tilde{\varsigma}_\beta(O) = 0$ if either $\bar{\mu}(0,x,d) = \mu_0(0,x,d)$ or $\bar{e}(a \mid x,d) = e_0(a \mid x,d)$. The double robustness is attractive because the propensity scores are generally known in clinical trials. The efficiency theory remains the same if one plugs in the true propensity scores, in which case the empirical mean squared error of the resulting estimator is potentially larger compared to if one estimates the propensity scores from the data.

References

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Springer.

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

Kennedy, E. H. (2024). Semiparametric doubly robust targeted double machine learning: A review. In *Handbook of Statistical Methods for Precision Medicine*. Chapman and Hall/CRC.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135.

Polley, E., LeDell, E., Kennedy, C., Lendle, S., and van der Laan, M. (2024). SuperLearner: Super learner prediction.

van der Vaart, A. W. and Wellner, J. A. (2023). *Weak convergence and empirical processes: With applications to statistics*. Springer Series in Statistics. Springer.

Manuscript III

Improving precision of cumulative incidence estimates in randomized controlled trials with external controls

Zehao Su
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
zehao.su@sund.ku.dk

Helene C. Rytgaard
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
hely@sund.ku.dk

Henrik Ravn
Novo Nordisk A/S,
Vandtårnsvej 108, DK-2860 Søborg, Denmark
hnrv@novonordisk.com

Frank Eriksson
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
eriksson@sund.ku.dk

Abstract

Augmenting the control arm in clinical trials with external data can improve statistical power for demonstrating treatment effects. In many time-to-event outcome trials, participants are subject to truncation by death. Direct application of methods for competing risks analysis on the joint data may introduce bias, for example, due to covariate shifts between the populations. In this work, we consider transportability of the conditional cause-specific hazard of the event of interest under the control treatment. Under this assumption, we derive semiparametric efficiency bounds of causal cumulative incidences. This allows for quantification of the theoretical efficiency gain from incorporating the external controls. We propose triply robust estimators that can achieve the efficiency bounds, where the trial controls and external controls are made comparable through time-specific weights in a martingale integral. We conducted a simulation study to show the precision gain of the proposed fusion estimators compared to their counterparts without utilizing external controls. As a real data application, we used two cardiovascular outcome trials conducted to assess the safety of glucagon-like peptide-1 agonists. Incorporating the external controls from one trial into the other, we observed a decrease in the standard error of the treatment effects on adverse non-fatal cardiovascular events with all-cause death as the competing risk.

Keywords: Data fusion; Transportability; Competing risks; Randomized controlled trial.

1. Introduction

Randomized control trials (RCTs) are the gold standard for evaluation of new treatments. Nonetheless, demonstrating the expected efficacy may require a substantial sample size, thereby requiring long duration of trials and driving up overall costs. Motivated by these issues, recent years have seen a growing interest in the use of historical data in clinical trials. In rare-disease trials, where it may be impractical and unethical to randomize a patient to the standard of care, regulatory bodies have discussed the feasibility of replacing trial controls with an external control arm (Food and Drug Administration, 2023; European Medicines Agency, 2023). Another example is hybrid control designs, in which the control arm in a clinical trial is augmented with external controls. The external controls should match the characteristics of the trial controls to avoid introducing bias, and their transportability should be carefully assessed in the planning phase of trials.

Leveraging external controls in clinical trials is an instance of data fusion. Despite the ubiquity of time-to-event outcomes in clinical trials, current literature on data fusion in causal inference mostly deals with continuous or binary outcomes. In this work, we consider external control augmentation for the estimation of treatment effects on the time-to-event, where an individual is subject to multiple modes of failure. Specifically, we wish to make inference on cumulative incidence functions defined on the counterfactual event time.

In the estimand framework, many transportability studies in survival analysis estimate the risk difference from at-risk indicators at predetermined timepoints (Ramagopalan et al., 2022; Zuo et al., 2022; Dang et al., 2023). Unless the censoring rate is ignorable, risk estimators constructed from dichotomized event times suffer from censoring bias. Lee et al. (2022) and Cao et al. (2024) provide a more formal treatment of the problem in generalizing treatment effects from a clinical trial to its superpopulation. They propose estimators for the target population counterfactual survival curve assuming transportability of the survival time distribution after conditioning on relevant baseline covariates. However, if the data contains competing events, their identification formula directly corresponds to the all-cause survival function, rather than the estimands desired here. Moreover, in our application, we observe the outcome for both trial participants and external controls, hence requiring separate estimation strategies (Colnet et al., 2024).

To accommodate competing risks, we work under the assumption of transportability of the cause-specific hazards, which are natural objects of interest in multi-state models. Although other assumptions can be postulated, they are either unnecessarily strong, such as transportability of the joint distribution of the event time and type, or lacking of interpretability in the data generating process, such as transportability of the subdistribution function (Fine and Gray, 1999). We construct semiparametrically efficient estimators by studying the nonparametric efficient influence functions of the parameters. The resulting estimators show robustness against model misspecification different from existing nonparametric estimators for cumulative incidence functions without data fusion (Rytgaard et al., 2023).

In the absence of competing risks, a related line of work extends dynamic borrowing methods to survival analysis (Kwiatkowski et al., 2024; Tan et al., 2022; Li et al., 2022; Sengupta et al., 2023). These methods control the extent to which external controls are

incorporated into the target population by modifying the data likelihood. They estimate the hazard ratio between the active arm and the control arm, which has been criticized for lacking causal interpretation. In contrast, we directly assume hazard transportability and consider robust estimators for marginal causal parameters with efficiency gain.

2. Identifiability of causal cumulative incidence difference

Without loss of generality, we consider two types of events: the event of interest ($J = 1$) and the competing event ($J = 2$). For the underlying event time T and event type J censored by the censoring time C, we observe the right-censored versions $\tilde{T} = T \wedge C$ and $\tilde{J} = I(T \leqslant C)J$ with a maximum observation period of $(0, \tau]$. Data is collected independently from two populations: the target population ($D = 1$) and the source population ($D = 0$). Besides the outcome tuple $(\tilde{T}, \tilde{J})$, a set of baseline covariates X is also observed in both populations. In our application, the target population is the study population of an RCT with both an active treatment ($A = 1$) and a control treatment ($A = 0$), while the source population contributes only controls. The supports of the baseline covariates in the RCT population and in the external control population are denoted by $\mathcal{X}_1$ and $\mathcal{X}_0$, respectively.

The observed data is sampled in a non-nested fashion, where random sampling is performed separately within the target population and the external control population (Dahabreh et al., 2021). More concretely, we have a probability sample $(\tilde{T}_i, \tilde{J}_i, A_i, X_i)$ from the target population for $i = 1, 2, \ldots, n_1$ and another probability sample $(\tilde{T}_i, \tilde{J}_i, X_i)$ from the external control population for $i = n_1 + 1, n_1 + 2, \ldots, n_1 + n_0$. The total sample size is denoted by n. For the asymptotic arguments that appear later, we need the following condition on the sampling scheme.

Assumption 1 (Stable sampling probability). As $n \to \infty$, $n_1/n \to \alpha \in (0, 1)$.

When the sample size n is large, we may view the joint sample as a random sample from some superpopulation distribution of $O = (\tilde{T}, \tilde{J}, DA, X, D)$ such that $\mathrm{pr}(D = 1) = \alpha$.

We are interested in the causal τ-time cumulative incidence difference in the target population for both event types. Under a specific treatment, the causal τ-time cumulative incidence is defined as the average probability of having an event by time τ, had all subjects in the target population received that treatment. Let the potential outcomes $\{T(a), J(a)\}$ denote time to event and event type under the static intervention $a = 0, 1$. The population-level target parameters defined before can be represented by

$$\theta_j(a) = \mathrm{pr}\{T(a) \leqslant \tau, J(a) = j \mid D = 1\}, \quad \theta_j = \theta_j(1) - \theta_j(0),$$

for event type $j \in \{1, 2\}$.

Assumption 2 (Causal assumptions).

(i) (Consistency) $T_i(a) = T_i$ and $J_i(a) = J_i$ if $A_i = a$ for $a \in \{0, 1\}$;

(ii) (Randomization) $\{T(a), J(a)\} \perp\!\!\!\perp A \mid (X, D = 1)$ and $\mathrm{pr}(A = a \mid X, D = 1) > 0$ for $a \in \{0, 1\}$.

With Assumption 2, the target parameters defined on the counterfactual data distri-

bution are identifiable from the uncensored data distribution. Let $F_{1j}(t \mid a, x) = \mathrm{pr}(T \leqslant t, J = j \mid A = a, X = x, D = 1)$ and $F_{0j}(t \mid x) = \mathrm{pr}(T \leqslant t, J = j \mid X = x, D = 0)$ be the conditional cumulative incidence functions. The causal τ-time cause j cumulative incidence under the intervention a is identified by the g-formula

$$\theta_j(a) = E\{F_{1j}(\tau \mid a, X) \mid D = 1\}.$$

To identify the parameter $\theta_j(a)$ with the observed data, some conditions on the censoring time are needed. Denote the survival functions of the all-cause event time by $S_1(t \mid a, x) = \mathrm{pr}(T > t \mid A = a, X = x, D = 1)$ and $S_0(t \mid x) = \mathrm{pr}(T > t \mid X = x, D = 0)$, and denote the survival functions of the censoring time by $S_1^c(t \mid a, x) = \mathrm{pr}(C > t \mid A = a, X = x, D = 1)$ and $S_0^c(t \mid x) = \mathrm{pr}(C > t \mid X = x, D = 0)$.

Assumption 3 (Censoring).

(i) (Positivity of censoring time) For all $t \in (0, \tau]$,

$$\begin{aligned} &S_1(t \mid a, x) > 0 \Rightarrow S_1^c(t \mid a, x) > 0, && \text{for } a \in \{0, 1\}, x \in \mathcal{X}_1; \\ &S_0(t \mid x) > 0 \Rightarrow S_0^c(t \mid x) > 0, && \text{for } x \in \mathcal{X}_1 \cap \mathcal{X}_0. \end{aligned}$$

(ii) (Independent censoring) $(T, J) \perp\!\!\!\perp C \mid (A, X, D = 1)$; $(T, J) \perp\!\!\!\perp C \mid (X, D = 0)$.

Under Assumption 3, the observed data likelihood at the realization $o = (t, j, a, x, d)$ of $O \sim P$ factorizes as

$$\begin{aligned} \mathrm{d}P(o) = {} & \mathrm{d}P(x)\{\pi(x)e_1(a \mid x)\}^d\{1 - \pi(x)\}^{(1-d)} \\ & \left[\{\mathrm{dA}_{1j}(t \mid a, x)\}^{I(j\neq 0)} S_1(t- \mid a, x)\right]^d \left[\{\mathrm{dA}_{0j}(t \mid x)\}^{I(j\neq 0)} S_0(t- \mid x)\right]^{(1-d)} \\ & \left(\left[\mathrm{dA}_1^c(t \mid a, x)\{1 - \triangle\mathrm{A}_{11}(t \mid a, x) - \triangle\mathrm{A}_{12}(t \mid a, x)\}\right]^{I(j=0)} S_1^c(t- \mid a, x)\right)^d \\ & \left(\left[\mathrm{dA}_0^c(t \mid x)\{1 - \triangle\mathrm{A}_{01}(t \mid x) - \triangle\mathrm{A}_{02}(t \mid x)\}\right]^{I(j=0)} S_0^c(t- \mid x)\right)^{(1-d)} \end{aligned}$$

where $\pi(x) = P(D = 1 \mid X = x)$ is the target population selection score, $e_1(a \mid x) = P(A = a \mid X = x, D = 1)$ is the treatment propensity score in the RCT, and the infinitesimal increment of the conditional cumulative hazards of the events and censoring are

$$\begin{aligned} \mathrm{dA}_{1j}(t \mid a, x) &= \frac{\mathrm{d}F_{1j}(t \mid a, x)}{S_1(t- \mid a, x)}, & \mathrm{dA}_1^c(t \mid a, x) &= -\frac{\mathrm{d}S_1^c(t \mid a, x)}{S_1^c(t- \mid a, x)}, \\ \mathrm{dA}_{0j}(t \mid x) &= \frac{\mathrm{d}F_{0j}(t \mid x)}{S_0(t- \mid x)}, & \mathrm{dA}_0^c(t \mid x) &= -\frac{\mathrm{d}S_0^c(t \mid x)}{S_0^c(t- \mid x)}. \end{aligned}$$

Given Assumptions 2–3, the parameter $\theta_j(a)$ can be identified as a functional of the observed data distribution. In Supplementary Material §S1, we relate the quantities defined in the observed data distribution to those defined in the uncensored data distribution.

3. Semiparametric theory for cumulative incidence

3.1. *Transportability of the cause-specific hazard of the event of interest*

We propose a key transportability assumption under which the RCT controls and the external controls are compatible. In data fusion, we adjust for prognostic variables with shifted distribution between the target population and the source population, so that conditional on these variables, the intervened populations are comparable in a certain respect. The baseline covariates X are obviously sufficient for this purpose if

$$\{N_1(0)(\tau), N_2(0)(\tau)\} \perp\!\!\!\perp D \mid X. \tag{1}$$

This strong condition states that the entire event processes under the control treatment become interchangeable between the populations, once the baseline covariates are controlled for. We will discuss an example where 1 is violated, but a weaker transportability assumption sufficient for our purpose is fulfilled.

We motivate the assumption in the simplified case where the time to event is observed on a discrete grid. Let $\triangle N_j(0)(t) = I\{T(0) = t, J(0) = j\}$ be the counterfactual indicator for an event of type j occurring at time $t \in \{1, \ldots, \tau\}$ under the control treatment and let $N_j(0)(t) = \sum_{s=1}^{t} \triangle N_j(0)(s)$. In addition to the variables in the previous section, we introduce shifted, unobserved prognostic variables U, whose existence may violate (1), because D and $\triangle N_1(0)(t)$ cannot be d-separated without blocking U. Consider a time-discretized data generating process encoded by the single-world intervention graph (Richardson and Robins, 2013) displayed in Figure 1. At any timepoint, the variables U directly affect the competing event $\triangle N_2(0)(t)$ but act only indirectly on the event of interest $\triangle N_1(0)(t)$ through the history of the events $\{N_1(0)(t-1), N_2(0)(t-1)\}$. In this case, it holds that

$$\triangle N_1(0)(t) \perp\!\!\!\perp D \mid \{N_1(0)(t-1), N_2(0)(t-1), X\} \tag{2}$$

for the event of interest without conditioning on the unobserved U. By definition,

$$\begin{aligned}\mathrm{pr}\{\triangle N_1(0)(t) = 1 \mid N_1(0)(t-1) = 0, N_2(0)(t-1) = 0, X, D\} \\ = \mathrm{pr}\{T(0) = t, J(0) = 1 \mid T(0) \geqslant t, X, D\},\end{aligned}$$

so the conditional independence (2) is equivalent to transportability of the cause-specific hazard of the event of interest.

In continuous time, an analogous formulation to (2) is the following.

Assumption 4 (Transportability of conditional cause 1 hazard). $\mathrm{A}_{11}(0)(t \mid x) = \mathrm{A}_{01}(0)(t \mid x)$ for $x \in \mathcal{X}_1 \cap \mathcal{X}_0$.

The interpretation is that for two subjects with the same baseline covariates, one in the RCT and one in the external population, given they have not experienced any event, the probability with which they immediately experience the event of interest are the same. Unobserved variables like U can also be time-varying, as long as they have no direct effect on the event of interest.

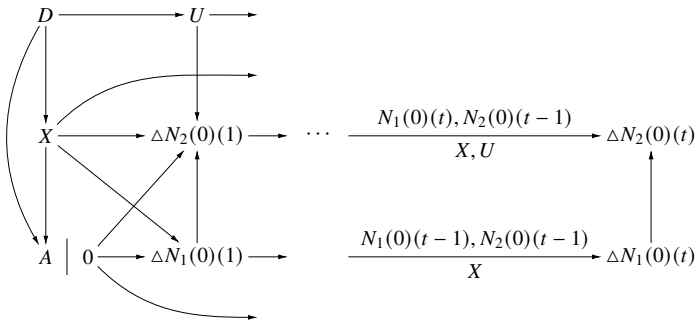

Figure 1. Discrete-time single-world intervention graph of a data generating process satisfying Assumptions 2 and 4.

3.2. *Semiparametric efficiency bound*

If Assumption 2 is satisfied, Assumption 4 further implies that

$$\mathrm{dA}_{11}(t \mid 0, x) = \mathrm{dA}_{01}(t \mid x), \qquad x \in \mathcal{X}_1 \cap \mathcal{X}_0. \tag{3}$$

Consider the model $\mathcal{P}$ of observed data distributions over O such that for any $P \in \mathcal{P}$, the distribution of $A \equiv 0$ is degenerate when $D = 1$, and the conditional cause 1 hazard under the control treatment $A = 0$ is transportable in the sense of (3). Since the hazard increments in (3) are not population-specific, we define $\mathrm{dA}_{\bullet 1}(t \mid 0, x)$ for all $x \in \mathcal{X}_1 \cup \mathcal{X}_0$ such that $\mathrm{dA}_{\bullet 1}(t \mid 0, x) = \mathrm{dA}_{11}(t \mid 0, x)$ if $x \in \mathcal{X}_1$ and $\mathrm{dA}_{\bullet 1}(t \mid 0, x) = \mathrm{dA}_{01}(t \mid x)$ if $x \in \mathcal{X}_0$.

In Proposition S1 of the Supplementary Material, we characterize the semiparametric efficiency bounds of the parameters $\theta_1(0)$ and $\theta_2(0)$ in model $\mathcal{P}$ under general formulations of the event processes with competing risks. To gain insights on how the external controls can be most efficiently integrated under the transportability assumption of the cause 1 hazard, in Lemma 1 below, we state the efficient influence functions under a mild regularity condition on the cumulative hazards. Let $\mathcal{A}$ be the class of functions $\mathrm{A} : (0, \tau] \to [0, \infty)$ which are càdlàg non-decreasing with finite variation and jump sizes no larger than 1. For any $\mathrm{A}, \mathrm{A}^* \in \mathcal{A}$, let $\mathrm{A} \perp\!\!\!\perp_{\triangle} \mathrm{A}^*$ denote that $\triangle\mathrm{A}(t)\triangle\mathrm{A}^*(t) = 0$ for $t \in (0, \tau]$.

Assumption 5 (Disjoint discontinuity points). $\mathrm{A}_{11}(t \mid 0, x) \perp_{\triangle} \mathrm{A}_{12}(t \mid 0, x)$ for $x \in \mathcal{X}_1$ and $\mathrm{A}_{01}(t \mid x) \perp_{\triangle} \mathrm{A}_{02}(t \mid x)$ for $x \in \mathcal{X}_1 \cap \mathcal{X}_0$.

In words, the first part of Assumption 5 states that the conditional cumulative hazards $\mathrm{A}_{11}(t \mid 0, x)$ and $\mathrm{A}_{12}(t \mid 0, x)$ under the control treatment do not share any discontinuity point for any baseline covariates in the target population. When there are jump points in the distribution function of the underlying event time T for a countable set of timepoints, the assumption implies that the probability that events of type 1 and type 2 are observed at the same time is 0. By Assumption 4, this also implies that $\triangle\mathrm{A}_{01}(t \mid x)\triangle\mathrm{A}_{12}(t \mid 0, x) = 0$ for $x \in \mathcal{X}_1 \cap \mathcal{X}_0$. The second part of the assumption can be interpreted analogously. Assumption 5 is certainly satisfied if T has a continuous distribution. The assumption is also satisfied if the conditional cumulative hazard of either cause is continuous. We will revisit Assumption 5 when we construct estimators for the target parameters.

For $a \in \{0,1\}$ and $j,k \in \{1,2\}$, let $N_j(t) = I(\tilde{T} \leqslant t, \tilde{J} = j)$ denote the counting process for the observed event of type j and define

$$\begin{aligned}
H_\bullet(t \mid x) &= \pi(x) e_1(0 \mid x)(S_1 S_1^c)(t \mid 0, x) + \{1 - \pi(x)\}(S_0 S_0^c)(t \mid x), \\
H_1(t \mid a, x) &= e_1(a \mid x)(S_1 S_1^c)(t \mid a, x), \\
W_{kj}(t \mid a, x) &= I(j = k) S_1(t- \mid a, x) - \frac{F_{1j}(\tau \mid a, x) - F_{1j}(t \mid a, x)}{1 - \triangle \mathrm{A}_{1k}(t \mid a, x)}.
\end{aligned}$$

Lemma 1 (Semiparametric efficiency bounds). *Suppose Assumptions 3 and 5 hold. For $a \in \{0,1\}$ and $j \in \{1,2\}$, the efficient influence function of $\theta_j(a)$ at $P \in \mathcal{P}$ is*

$$\begin{aligned}
\varphi_j(a)(O) = \frac{I(A = a)}{\alpha} \int_0^\tau & \left\{ \frac{I(a = 0)\pi(X)}{H_\bullet(t- \mid X)} + \frac{I(a = 1)D}{H_1(t- \mid 1, X)} \right\} W_{1j}(t \mid a, X) \\
& \left\{ \mathrm{d}N_1(t) - I(\tilde{T} \geqslant t) \mathrm{d}\mathrm{A}_{11}(t \mid a, X) \right\} \\
+ \frac{D}{\alpha} \int_0^\tau & \frac{I(A = a)}{H_1(t- \mid a, X)} W_{2j}(t \mid a, X) \left\{ \mathrm{d}N_2(t) - I(\tilde{T} \geqslant t) \mathrm{d}\mathrm{A}_{12}(t \mid a, X) \right\} \\
+ \frac{D}{\alpha} & \{F_{1j}(\tau \mid a, X) - \theta_j(a)\}.
\end{aligned}$$

The semiparametric efficiency bound of $\theta_j(a)$ at $P \in \mathcal{P}$ is $E_P \varphi_j^2(a)$.

The efficient influence functions of the parameters $\theta_j(1)$ are identical to those presented by Eq. (4) in Rytgaard et al. (2023), with the only difference being that they are restricted to the distribution on the RCT population. Since the nuisance parameters in $\varphi_j(1)(O)$ are all variationally independent of the cumulative hazards $\mathrm{A}_{11}(t \mid 0, x)$ and $\mathrm{A}_{01}(t \mid x)$, Assumption 4 does not change the characterization of the efficient estimators of $\theta_j(1)$.

On the other hand, comparing the efficient influence function $\varphi_1(0)(O)$ with the influence function of $\theta_1(0)$ without using the information of external controls, we notice that the inverse weight $1/H_\bullet(t- \mid x)$ is applied for efficient use of data. We can write

$$H_\bullet(t \mid x) = \mathrm{pr}(T > t, C > t, A = 0 \mid X = x) = P(A = 0 \mid X = x) P(\tilde{T} > t \mid A = 0, X = x),$$

which is a product of the probability of receiving the control treatment and the survival function of an event of any type, including censoring, defined on the artificial population conjoining the whole external population and the subset of the target population under the control treatment. It should be noted, however, that the function $W_{11}(t \mid 0, x)$ is identifiable from the target population only. Therefore, the predictable process in the event-of-interest martingale integral from $\varphi_1(0)(O)$ is a combination of pooled and unpooled quantities across populations.

Corollary 1. *Under the same conditions in Lemma 1, the semiparametric efficiency bound of $\theta_1(0)$ under $\mathcal{P}$ is at least as low as that under the model where restriction* (3) *is removed. The reduction is*

$$E\left[\frac{\pi(X)\{1 - \pi(X)\}}{\alpha^2} \int_0^\tau \frac{(S_0 S_0^c)(t- \mid X)}{H_1(t- \mid 0, X) H_\bullet(t- \mid X)} W_{11}^2(t \mid 0, X)\{1 - \triangle \mathrm{A}_{11}(t \mid 0, X)\} \mathrm{d}\mathrm{A}_{11}(t \mid 0, X) \right].$$

In words, incorporating the external controls helps drop the lowest possible variance attainable by a regular estimator of the target parameter $\theta_1(0)$ under the transportability assumption, if two conditions are met. First, there is an overlap in the distributions of the baseline covariates between the populations. Second, in this overlapped population, there is a non-trivial time span in the observation period during which an individual is at risk of experiencing the event of interest.

Corollary 1 shows that the variance reduction is accumulated over time with respect to the cumulative hazard $\mathrm{A}_{11}(t \mid 0, x)$, and the time-varying factors that determine the size of variance reduction cannot be teased apart. We give some intuition on when the use of external controls provides large precision gain. The product integral of any $\mathrm{A} \in \mathcal{A}$ is denoted by $(\Pi\mathrm{A})(t) = \Pi_{s\in(0,t]}\{1 - \mathrm{dA}(t)\}$. Note that

$$\frac{(S_0 S_0^c)(t- \mid x)}{H_\bullet(t- \mid x)} = \left\{\pi(x) e_1(0 \mid x) \frac{(\Pi \mathrm{A}_{12} S_1^c)(t- \mid 0, x)}{(\Pi \mathrm{A}_{02} S_0^c)(t- \mid x)} + \{1 - \pi(x)\}\right\}^{-1} I\big\{(S_0 S_0^c)(t- \mid x) > 0\big\}.$$

All other factors being equal, the reduction is more pronounced when the ratio between the product of product integrals

$$\frac{(\Pi \mathrm{A}_{12} S_1^c)(t- \mid 0, x)}{(\Pi \mathrm{A}_{02} S_0^c)(t- \mid x)}$$

is smaller. In the extreme scenario where the said ratio is simply 0, the variance reduction formula gives

$$E\left[\frac{\pi(X)}{\alpha^2} I\{\pi(X) < 1\} \int_0^\tau I\big\{(S_0 S_0^c)(t- \mid X) > 0\big\} \frac{W_{11}^2(t \mid 0, X)}{H_1(t- \mid 0, X)} \{1 - \triangle \mathrm{A}_{11}(t \mid 0, X)\} \mathrm{dA}_{11}(t \mid 0, X)\right].$$

Effectively, the maximum possible reduction is the portion of asymptotic variance resulting from the martingale $N_1(t) - \int_0^t I(\tilde{T} \geqslant s) \mathrm{dA}_{11}(s \mid 0, X)$ on the region where the indicator $I\big\{\pi(X) < 1, (S_0 S_0^c)(t- \mid X) > 0\big\}$ stays 1. In practical terms, when the hazard of the competing risk is much higher for subjects under the control treatment or when censoring occurs much earlier in the target population, the variance reduction is larger. Intuitively, it is most beneficial to incorporate the external controls on the ground of hazard transportability when the hazard of the event of interest cannot be estimated well from the target population alone otherwise, due to the lack of such events in the observed data.

3.3. *Estimation*

In the following, we discuss the construction of estimators that asymptotically achieve the semiparametric efficiency bounds in Lemma 1. We present results for the parameter $\theta_1(0)$ only. An estimator for $\theta_2(0)$ and its properties can be derived analogously. The estimators of the parameters $\theta_1(1)$ and $\theta_2(1)$ do not involve the external control sample, and thus the estimation strategy for these parameters follows directly from Rytgaard et al. (2023).

Suppose for the nuisance parameters, we have estimators

$$\left\{\hat{\mathrm{A}}_{\bullet 1}(t \mid 0, x), \hat{\mathrm{A}}_{12}(t \mid 0, x), \hat{\mathrm{A}}_{02}(t \mid x), \hat{\mathrm{A}}_1^c(t \mid 0, x), \hat{\mathrm{A}}_0^c(t \mid x)\right\} \subset \mathcal{A}$$

and that $\hat{e}_1(0 \mid x)$ and $\hat{\pi}(x)$ are valid probabilities. The cumulative incidence function of cause 1 in the RCT sample are estimated by

$$\hat{F}_{11}(t \mid 0, x) = \int_0^t \hat{S}_1(s- \mid 0, x) \mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(s \mid 0, x),$$

where the integral is in the Lebesgue–Stieltjes sense, and the conditional survival function of the all-cause event time T is estimated by

$$\hat{S}_1(t \mid 0, x) = (\Pi \hat{\mathrm{A}}_{\bullet 1} \Pi \hat{\mathrm{A}}_{12})(t \mid 0, x).$$

The survival functions of the censoring time are the product integrals $\hat{S}_1^c = \Pi \hat{\mathrm{A}}_1^c$ and $\hat{S}_0^c = \Pi \hat{\mathrm{A}}_0^c$, respectively. Observing the efficient influence function $\varphi_1(0)$ given in Lemma 1, we define the uncentered efficient influence function and its plug-in version as

$$\begin{aligned}
\ell_1(0)(O) &= \varphi_1(0)(O) + \frac{D}{\alpha}\theta_1(0), \\
\hat{\ell}_1(0)(O) &= \frac{1-A}{\hat{\alpha}}\hat{\pi}(X) \int_0^\tau \frac{\hat{W}_{\bullet 1}(t \mid 0, X)}{\hat{H}_\bullet(t- \mid X)} \left\{\mathrm{d}N_1(t) - I(\tilde{T} \geqslant t) \mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(t \mid 0, X)\right\} \\
&\quad + \frac{D(1-A)}{\hat{\alpha}} \int_0^\tau \frac{\hat{W}_{21}(t \mid 0, X)}{\hat{H}_1(t \mid 0, X)} \left\{\mathrm{d}N_2(t) - I(\tilde{T} \geqslant t) \mathrm{d}\hat{\mathrm{A}}_{12}(t \mid 0, X)\right\} \\
&\quad + \frac{D}{\hat{\alpha}} \hat{F}_{11}(\tau \mid 0, X),
\end{aligned}$$

where

$$\begin{aligned}
\hat{S}_0(t \mid x) &= (\Pi \hat{\mathrm{A}}_{\bullet 1})(t \mid 0, x)(\Pi \hat{\mathrm{A}}_{02})(t \mid x), \\
\hat{H}_\bullet(t \mid x) &= \hat{\pi}(x)\hat{e}_1(0 \mid x)(\hat{S}_1 \hat{S}_1^c)(t \mid 0, x) + \{1 - \hat{\pi}(x)\}(\hat{S}_0 \hat{S}_0^c)(t \mid x), \\
\hat{H}_1(t \mid x) &= \hat{e}_1(0 \mid x)(\hat{S}_1 \hat{S}_1^c)(t \mid 0, x), \\
\hat{W}_{\bullet 1}(t \mid 0, x) &= \hat{S}_1(t- \mid 0, x) - \frac{\hat{F}_{11}(\tau \mid 0, x) - \hat{F}_{11}(t \mid 0, x)}{1 - \triangle \hat{\mathrm{A}}_{\bullet 1}(t \mid 0, x)}, \\
\hat{W}_{21}(t \mid 0, x) &= -\frac{\hat{F}_{11}(\tau \mid 0, x) - \hat{F}_{11}(t \mid 0, x)}{1 - \triangle \hat{\mathrm{A}}_{12}(t \mid 0, x)}.
\end{aligned}$$

We propose the estimator

$$\hat{\theta}_1(0) = \frac{1}{n} \sum_{i=1}^n \hat{\ell}_1(0)(O_i)$$

of $\theta_1(0)$.

Assumption 6 (Probability limits). There exist probability limits $0 \leqslant \bar{\pi}(x) \leqslant 1$, $0 \leqslant \bar{e}_1(0 \mid x) \leqslant 1$ such that $\|(\hat{\pi} - \bar{\pi})(X)\|_P = o_P(1)$, $\|(\hat{e}_1 - \bar{e}_1)(0 \mid X)\|_P = o_P(1)$, and

$$\left\{\bar{\mathrm{A}}_{\bullet 1}(t \mid 0, x), \bar{\mathrm{A}}_{12}(t \mid 0, x), \bar{\mathrm{A}}_{02}(t \mid x), \bar{\mathrm{A}}_1^c(t \mid 0, x), \bar{\mathrm{A}}_0^c(t \mid x)\right\} \subset \mathcal{A}$$

such that

$$\left\| I\{\pi(X) > 0\} \sup_{t\in(0,\tau]} |\hat{\mathrm{A}}_{\bullet 1} - \bar{\mathrm{A}}_{\bullet 1}|(t \mid 0, X) \right\|_P = o_P(1),$$

$$\left\| I\{\pi(X) > 0\} \sup_{t\in(0,\tau]} |\hat{\mathrm{A}}_{12} - \bar{\mathrm{A}}_{12}|(t \mid 0, X) \right\|_P = o_P(1),$$

$$\left\| I\{0 < \pi(X) < 1\} \sup_{t\in(0,\tau]} |\hat{\mathrm{A}}_{02} - \bar{\mathrm{A}}_{02}|(t \mid X) \right\|_P = o_P(1),$$

$$\left\| I\{\pi(X) > 0\} \sup_{t\in(0,\tau]} |\hat{\mathrm{A}}^c_1 - \bar{\mathrm{A}}^c_1|(t \mid 0, X) \right\|_P = o_P(1),$$

$$\left\| I\{0 < \pi(X) < 1\} \sup_{t\in(0,\tau]} |\hat{\mathrm{A}}^c_0 - \bar{\mathrm{A}}^c_0|(t \mid X) \right\|_P = o_P(1).$$

Theorem 1 (Asymptotic behavior). *Suppose Assumptions 3 and 6 as well as Assumption S1 in the Supplementary Material hold. Then* $\hat{\theta}_1(0) \xrightarrow{\mathrm{p}} \theta_1(0)$ *if*

(i) $\bar{\mathrm{A}}_{\bullet 1} = \mathrm{A}_{\bullet 1}$ *and* $\bar{\mathrm{A}}_{12} = \mathrm{A}_{12}$*;*

(ii) $\bar{\mathrm{A}}_{\bullet 1} = \mathrm{A}_{\bullet 1}$, $\bar{e}_1 = e_1$, *and* $\bar{\pi} = \pi$*; or*

(iii) $\bar{\mathrm{A}}_{12} = \mathrm{A}_{12}$, $\bar{\mathrm{A}}_{02} = \mathrm{A}_{02}$, $\bar{\mathrm{A}}^c_1 = \mathrm{A}^c_1$, $\bar{\mathrm{A}}^c_0 = \mathrm{A}^c_0$, $\bar{e}_1 = e_1$, *and* $\bar{\pi} = \pi$.

Moreover,

$$\hat{\theta}_1(0) - \theta_1(0) = \frac{1}{n}\sum_{i=1}^{n} \varphi_1(0)(O_i) + o_P(n^{-1/2})$$

if $\bar{\mathrm{A}}_{\bullet 1} = \mathrm{A}_{\bullet 1}$, $\bar{\mathrm{A}}_{12} = \mathrm{A}_{12}$, $\bar{\mathrm{A}}_{02} = \mathrm{A}_{02}$, $\bar{\mathrm{A}}^c_1 = \mathrm{A}^c_1$, $\bar{\mathrm{A}}^c_0 = \mathrm{A}^c_0$, $\bar{e}_1 = e_1$, $\bar{\pi} = \pi$, *and Assumption S2 in the Supplementary Material is satisfied.*

The first part of Theorem 1 shows that the estimator $\hat{\theta}_1(0)$ constructed from the efficient influence function is triply robust against model misspecification. The consistency of $\hat{\theta}_1(0)$ hinges on correct estimation of at least one of the cause-specific hazards, namely $\mathrm{A}_{\bullet 1}(t \mid 0, x)$ or $\mathrm{A}_{12}(t \mid 0, x)$. In particular, if the cause 1 hazard does not converge to the underlying hazard, the cause 2 hazards in both populations need to be modeled correctly. Conditions for the asymptotic linearity of $\hat{\theta}_1(0)$ are given in the second part of Theorem 1. Apart from requiring the consistency of all nuisance models in their respective sense, the von Mises expansion of $\hat{\theta}_1(0)$ around the true parameter $\theta_1(0)$ demands that the two remainder terms in Assumption S2 converge as fast as $o_P(n^{-1/2})$; see Remark S2 for details.

The estimator $\hat{\theta}_1(0)$ attains the semiparametric efficiency bound in the model where the conditional cause 1 hazard under the control treatment is transportable. However, there is no free lunch. Compared to estimators that do not rely on the external controls, the proposed data fusion estimator involve additional nuisance models for the selection score π and the cumulative hazards A_{02} and A^c_0. If these models are not correctly estimated, we have no guarantee that $\hat{\theta}_1(0)$ will be more efficient than estimators based solely on the RCT sample.

A final remark should be made in connection with Assumption 5. When tied event times are observed for event types 1 and 2, the plug-in estimators based on the efficient influence function under Assumption 5 might be unfounded. We can avoid this issue if the event times are continuous by nature, and it is harmless to break the ties by numerical perturbations. Otherwise, we can turn to fully discrete-time methods (Benkeser

et al., 2018) or derive estimators based on Proposition S1 to handle mixed event time distributions.

3.4. *Restricted mean time lost*

Another interpretable parameter in competing risks analysis is the τ-restricted mean time lost to cause j (Andersen, 2013), defined as

$$\gamma_j(a) = E(I\{J(a) = j\}[\tau - \{T(a) \wedge \tau\}] \mid D = 1).$$

We can extend the definition of the parameter $\theta_j(a)$ to the population cumulative incidence of event type j under intervention a at time t, which is $\theta_j(a,t) = \text{pr}\{T(a) \leqslant t, J(a) = j \mid D = 1\}$. Given Assumptions 2–3, the parameter $\gamma_j(a)$ is identifiable as $\gamma_j(a) = \int_0^\tau \theta_j(a,t)\text{d}t$, where $\theta_j(a,t)$ is treated as an observed data parameter. If we view $\theta_j(a,t)$ as a function of time, then $\gamma_j(a)$ is simply the area under the cumulative incidence function capped at τ.

Restricted mean times lost are Hadamard differentiable functionals of the cumulative incidence functions. Hence, their efficient influence functions can be obtained from Lemma 1 by the chain rule.

Corollary 2. *Under the same conditions as in Lemma 1, the efficient influence function of $\gamma_j(a)$ at $P \in \mathcal{P}$ is $\psi_j(a)(O) = \int_0^\tau \varphi_j(a,t)(O)\text{d}t$, where $\varphi_j(a,t)$ is the efficient influence function of $\theta_j(a,t)$.*

Fusion estimators for $\gamma_j(a)$ are straightforward integrals of the fusion estimators $\hat{\theta}_j(a,t)$ for $\theta_j(a,t)$ over time t. The asymptotics of these estimators can be established under conditions similar to those inside Theorem 1. Particularly for asymptotic linearity of $\hat{\gamma}_1(0)$, the rate conditions in Assumption S2 should be modified according to Remark S3 in the Supplementary Material.

4. Simulation study

We investigated the performance of the fusion estimators compared to the RCT-only estimators in a simulation study. The data at baseline (X, D, A) were generated sequentially in the following manner:

$$\begin{aligned} X &\sim 2\Phi[\text{Normal}\{(0,0,0)^{\text{T}}, \Sigma\}] - 1, \\ D \mid X &\sim \text{Bernoulli}\{\pi(X)\}, \\ A \mid (X, D) &\sim \text{Bernoulli}(0.5D), \end{aligned}$$

where $\pi(X) = \text{expit}(-0.2 + 0.4X_1 + 0.2X_2 + 0.3X_3)$, Φ is the distribution function of the standard normal distribution, and the covariance matrix is

$$\Sigma = \begin{pmatrix} 1 & 0.25 & 0.25 \\ 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 1 \end{pmatrix}.$$

The uncensored event times were simulated from distributions with the following multiplicative hazards:

$$\alpha_{11}(t \mid A, X) = \alpha_1(t)\exp(0.5A + 0.2X_1 + 0.7X_3),$$

$$\begin{aligned}\alpha_{12}(t \mid A, X) &= \alpha_2(t) \exp(1 + 0.05A + 0.8X_1 + 0.5X_2),\\ \alpha_{01}(t \mid X) &= \alpha_1(t) \exp(0.2X_1 + 0.7X_3),\\ \alpha_{02}(t \mid X) &= \alpha_2(t) \exp(0.5X_1 + 0.8X_2 - 0.3X_3),\end{aligned}$$

where the baseline hazards $\alpha_1(t)$ and $\alpha_2(t)$ both correspond to the hazard of the Weibull distribution with shape parameter 0.7 and scale parameter 0.2. In other words, $\alpha_d(t) = 0.2 \cdot 0.7 t^{0.7-1}$. The censoring times were simulated from distributions with the following multiplicative hazards:

$$\begin{aligned}\alpha_1^c(t \mid A, X) &= \alpha^c(t) \exp\{0.5 + 0.05(1 - A)X_1 - 0.05X_3\},\\ \alpha_0^c(t \mid X) &= \alpha^c(t) \exp(0.05X_2),\end{aligned}$$

where the baseline hazard $\alpha^c(t)$ is the hazard of the Weibull distribution with shape parameter 0.7 and scale parameter 0.24. Under this data generating mechanism, the proportion of samples from the external control population was around 55%.

The target population selection score $\hat{\pi}(x)$ and the propensity score of treatment in the target population $\hat{e}_1(a \mid x)$ were estimated using logistic regressions. The cause 1 hazard under the control treatment $\mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(t \mid 0, x)$ was fitted with a Cox model combining all samples under the control treatment and the event indicator $I(\tilde{J} = 1)$. The cause 2 hazards under the control treatment $\mathrm{d}\hat{\mathrm{A}}_{d2}(t \mid 0, x)$ were fitted with a Cox model within the respective population using the event indicator $I(\tilde{J} = 2)$. The two cause-specific hazards under the active treatment $\mathrm{d}\hat{\mathrm{A}}_{1j}(t \mid 1, x)$ were obtained with a multi-state Cox model in the RCT population using the state indicator $\tilde{J}$. The hazards of the censoring were fitted with a Cox model for each treatment within the respective population using the event indicator $I(\tilde{J} = 0)$. The nuisance function estimates $\hat{S}_d, \hat{F}_{dj}, \hat{S}_d^c$ were subsequently computed using the hazard estimates.

As an example, we present the nuisance estimators required for the estimator $\hat{\theta}_1(0)$, which included $\hat{S}_1(t \mid 0, x)$, $\hat{F}_{11}(t \mid 0, x)$, $\hat{S}_1^c(t \mid 0, x)$, $\hat{S}_0(t \mid x)$, and $\hat{S}_0^c(t \mid x)$. The cumulative hazard estimates from Cox models are càdlàg step functions. The approximation $\triangle\hat{\mathrm{A}}_{\bullet 1}(s \mid 0, x) \approx 1 - \exp\{-\triangle\hat{\mathrm{A}}_{\bullet 1}(s \mid 0, x)\}$ was applied due to possible jumps whose sizes exceed one, ensuring that it fell between 0 and 1. The survival function of the composite event in the RCT population was approximated by $\hat{S}_1(t \mid 0, x) = \exp\big\{-\hat{\mathrm{A}}_{\bullet 1}(t \mid 0, x) - \hat{\mathrm{A}}_{12}(t \mid 0, x)\big\}$. The cumulative incidence function of the event of interest was computed using the Lebesgue-Stieltjes integral $\hat{F}_{11}(t \mid 0, x) = \int_0^t \hat{S}_1(s- \mid 0, x) \mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(s \mid 0, x)$. Similarly, the survival function of the composite event in the external control population was $\hat{S}_0(t \mid x) = \exp\big\{-\hat{\mathrm{A}}_{\bullet 1}(t \mid 0, x) - \hat{\mathrm{A}}_{02}(t \mid x)\big\}$. The Cox-estimated cumulative hazard $\hat{\mathrm{A}}_{\bullet 1}(t \mid 0, x)$ did not share any discontinuity points with $\hat{\mathrm{A}}_{02}(t \mid x)$, since there were no ties among event times of different types. The survival functions of the censoring time are $\hat{S}_1^c(t \mid 0, x) = \exp\big\{-\hat{\mathrm{A}}_1^c(t \mid 0, x)\big\}$ and $\hat{S}_0^c(t \mid x) = \exp\big\{-\hat{\mathrm{A}}_0^c(t \mid x)\big\}$, respectively.

We simulated data of sample size $n \in \{750, 1500\}$ from the described data generating mechanism. The proposed estimator $\hat{\theta}_1(0, t)$ for the cumulative incidence of the event of interest under control treatment was computed for three time points $t \in \{0.25, 1, 2\}$. The standard error of $\hat{\theta}_1(0, t)$ was estimated by $n^{-1/2}$ times the empirical L_2-norm of the efficient influence function $\varphi_1(0)$. The estimators $\hat{\theta}_2(0, t)$, $\hat{\theta}_1(1, t)$, and $\hat{\theta}_2(1, t)$ for other cumulative incidences and the estimators $\hat{\theta}_1\{t\}$ and $\hat{\theta}_2\{t\}$ for the average treatment effects were also computed using the respective efficient influence functions. To demonstrate the gain in precision, we compared the estimated asymptotic variance of the estimators above to the estimators that would be efficient if only the RCT data was

Table 1. *Simulation results for estimators of cumulative incidences.*

n	Estimand	t	Type	Mean	Bias	RMSE	SE	Coverage	Reduction
750	$\theta_1(0,t)$	0.25	+	0.07	5.73	1.17	1.15	94.5	66.84
			−	0.07	8.20	2.15	2.05	92.7	·
		1	+	0.14	3.66	1.60	1.60	95.1	69.42
			−	0.14	5.50	2.99	2.92	92.3	·
		2	+	0.19	2.03	1.82	1.84	94.8	69.20
			−	0.19	−4.92	3.46	3.35	93.5	·
	$\theta_1(t)$	0.25	+	0.04	−3.03	2.70	2.76	95.1	27.35
			−	0.04	−5.50	3.17	3.25	96.3	·
		1	+	0.08	8.53	3.68	3.77	95.6	29.58
			−	0.08	6.69	4.50	4.50	95.1	·
		2	+	0.09	14.66	4.07	4.23	96.2	30.43
			−	0.09	21.61	5.04	5.08	94.7	·
1500	$\theta_1(0,t)$	0.25	+	0.07	1.79	0.81	0.81	94.3	68.31
			−	0.07	3.51	1.42	1.46	94.9	·
		1	+	0.14	3.60	1.09	1.13	95.5	70.17
			−	0.14	−0.17	2.07	2.07	94.2	·
		2	+	0.19	2.23	1.25	1.30	96.0	70.13
			−	0.19	−5.99	2.38	2.39	94.8	·
	$\theta_1(t)$	0.25	+	0.04	−3.05	2.00	1.95	94.1	27.73
			−	0.04	−4.77	2.28	2.30	95.2	·
		1	+	0.07	−6.90	2.76	2.66	93.6	29.93
			−	0.07	−3.13	3.24	3.18	94.0	·
		2	+	0.09	−12.92	3.06	2.99	94.2	30.98
			−	0.09	−4.70	3.59	3.60	95.1	·

Type: fusion estimator (+) or RCT-only estimator (−); Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-4}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %; Reduction: average of percentage reduction in squared standard error estimates, %.

available. The exact expressions of all other estimators can be found in Supplementary Materials §S3. As an alternative effect measure, we also considered the treatment effect as the difference between restricted mean times lost capped at $t \in \{0.25, 1, 2\}$. The calculations were repeated 1000 times for each sample size.

Summary statistics of selected estimators from the simulation study are reported in Tables 1–2. Results for the remaining estimators are deferred to Tables S1–S4 in the Supplementary Material. All estimators have small empirical bias. The averages of the plug-in standard errors align with the empirical root mean squared errors. The coverage of the 95%-confidence intervals constructed from these plug-in standard errors appears largely correct. The percentage reduction in variance is higher for the control parameters $\theta_1(0,t)$ and $\gamma_1(0,t)$ than it is for the treatment effects $\theta_1\{t\}$ and $\gamma_1\{t\}$. This should be expected since the parameters $\theta_1(1,t)$ and $\gamma_1(1,t)$, and thus their estimators, do not use information from external controls.

5. Real data example

In this data example, we use data from the clinical trials SUSTAIN-6 (ClinicalTrials.gov ID NCT01720446, Marso et al., 2016a) and LEADER (ClinicalTrials.gov ID

Table 2. *Simulation results for estimators of restricted mean times lost.*

n	Estimand	t	Type	Mean	Bias	RMSE	SE	Coverage	Reduction
750	$\gamma_1(0,t)$	0.25	+	0.01	−0.20	0.20	0.20	94.6	64.49
			−	0.01	0.66	0.36	0.35	92.1	·
		1	+	0.10	−4.03	1.17	1.16	94.6	67.61
			−	0.10	0.12	2.12	2.07	93.0	·
		2	+	0.26	−18.36	2.78	2.74	93.4	68.04
			−	0.26	−19.14	5.06	4.89	92.4	·
	$\gamma_1(t)$	0.25	+	0.01	−0.36	0.46	0.47	94.9	26.45
			−	0.01	−1.21	0.55	0.56	95.9	·
		1	+	0.05	−0.03	2.62	2.73	96.0	28.21
			−	0.05	−4.18	3.14	3.22	94.3	·
		2	+	0.14	2.81	6.04	6.33	96.5	29.05
			−	0.14	3.58	7.37	7.52	95.3	·
1500	$\gamma_1(0,t)$	0.25	+	0.01	−0.08	0.14	0.14	94.2	66.55
			−	0.01	0.15	0.25	0.25	93.7	·
		1	+	0.10	−2.50	0.81	0.82	95.0	68.40
			−	0.10	−3.70	1.46	1.47	94.7	·
		2	+	0.26	−9.76	1.88	1.94	95.2	68.73
			−	0.26	−16.13	3.48	3.49	95.2	·
	$\gamma_1(t)$	0.25	+	0.01	−0.62	0.34	0.34	94.6	26.69
			−	0.01	−0.85	0.39	0.40	94.9	·
		1	+	0.05	−4.73	2.01	1.94	93.5	28.43
			−	0.05	−3.54	2.31	2.29	94.8	·
		2	+	0.13	−17.48	4.67	4.49	93.0	29.33
			−	0.14	−11.11	5.39	5.35	94.6	·

Type: fusion estimator (+) or RCT-only estimator (−); Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-4}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %; Reduction: average of percentage reduction in squared standard error estimates, %.

Table 3. *Numbers of randomized subjects and events by arm in SUSTAIN-6 and LEADER.*

	SUSTAIN-6 (once-weekly)				LEADER (once-daily)	
	Semaglutide		Placebo		Liraglutide	Placebo
	1.0 mg	0.5 mg	1.0 mg	0.5 mg	1.8 mg	1.8 mg
Total	822	826	825	824	4668	4672
Non-fatal cardiovascular event	29	38	48	53	242	271
All-cause death	23	24	21	27	135	155

Total: total number of randomized subjects at baseline; the other numbers count the non-fatal cardiovascular events and all-cause deaths on or before day 728.

NCT01179048, Marso et al., 2016b). The overall objective is to incorporate the controls collected in LEADER ($D = 0$) in the statistical analysis on the study population of SUSTAIN-6 ($D = 1$) to boost the precision of estimates. The number of subjects randomized to placebo is 1649 in SUSTAIN-6 and 4672 in LEADER. The placebos are both subcutaneous injections matched to their corresponding active treatment. The frequency of injection is once daily in LEADER but once weekly in SUSTAIN-6. We proceed by regarding the three placebos as the same intervention.

We define the event of interest as the composite event of nonfatal myocardial infarction or nonfatal stroke ($J = 1$), which we refer to as the non-fatal cardiovascular event. We treat death from all causes as the competing event ($J = 0$). Time-zero in the analysis is the time of treatment or placebo randomization. The first set of parameters we considered were the cumulative incidences $\theta_j(a, t)$ for both events at week 26, week 52, week 78, and week 104 in the study population of SUSTAIN-6 of once-weekly semaglutide, 1.0 mg ($A = 1$), as well as the average treatment effects $\theta_j\{t\} = \theta_j(1, t) - \theta_j(0, t)$. We set the limit of the time span as the end of the follow-up period in SUSTAIN-6. We also considered the restricted mean times lost to the events $\gamma_j(a)$ capped at week 26, week 52, week 78, and week 104 and the corresponding effect $\gamma_j\{t\} = \gamma_j(1, t) - \gamma_j(0, t)$. See Table 3 for a breakdown of sample sizes by randomization arm and the numbers of events observed until week 104.

The main analysis was carried out under the transportability assumption that the cause-specific hazard of the event of interest under placebo, conditioning on relevant baseline covariates, is the same in the study population of SUSTAIN-6 and that of LEADER. The baseline covariates X to adjust for included age, sex, weight, duration of type-2 diabetes, glycated hemoglobin level, systolic and diastolic blood pressure, level of low-density lipoprotein cholesterol, smoking status, as well as history of ischemic heart disease, myocardial infarction, heart failure, ischemic stroke, and hypertension. The inclusion-exclusion criteria for the studies are highly comparable. Therefore, the transportability assumption implies that any difference in the marginal hazard of the event of interest can be attributed to the differences in the rates of death, in the rates of censoring, and/or in the baseline characteristics induced by sampling. The causal and transportability assumptions are compatible with the local independence graph (Didelez, 2008) without right-censoring in Figure 2. The cause-specific hazards and hazards of censoring were estimated with the Cox proprotional hazards model. The hazards in the RCT sample were fitted separately within the treatment arms to ensure full treatment-covariate interaction and per-stratum baseline hazards.

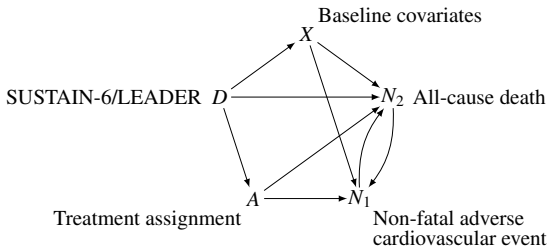

Figure 2. Hypothesized local independence graph of the variables used in the data example. The nodes $N_j(t)$ are uncensored versions of the counting processes.

The results are reported in Tables 4–5. We highlight the results at week 104. The fusion estimate of $\theta_1\{104\}$ demonstrates a decrease of 2.72 percentage points [95%-confidence interval: $(-4.33, -1.12)$] in the cumulative incidence of non-fatal cardiovascular event by semaglutide. There appears to be no evidence for semaglutide's effect on the cumulative incidence of all-cause death $\theta_2\{104\}$. Semaglutide does not seem to lower the risk of non-fatal cardiovascular event because of an increased risk of cardiovascular death. The restricted mean time lost to non-fatal cardiovascular event $\gamma_1\{104\}$ reduces by 1.15 week [95%-confidence interval: $(-2.24, -0.07)$] with semaglutide. Again, semaglutide does not appear to change the restricted mean time lost to all-cause death. The estimates and confidence intervals for time points before week 104 do not hint at any treatment effect on non-fatal cardiovascular event, except for $\theta_1\{78\}$. The results for treatment-specific parameters are displayed in Tables S6–S7 of the Supplementary Material.

For treatment effects of semaglutide in terms of $\theta_1\{t\}$ and $\gamma_1\{t\}$, the fusion point estimate and the RCT-only estimates are rather comparable. On the other hand, the length of confidence intervals is shortened by approximately 9% for all treatment effects. Despite the inclusion of external controls amounting to nearly three times the controls in SUSTAIN-6, the precision gain is hardly impressive. We believe the information bottleneck is the lack of subjects receiving the active treatment, since the size of the placebo group in SUSTAIN-6 was already twice as large as that of the semaglutide 1.0 mg group. This is supported by the observation that the percentage reduction in standard errors is above 20% for the under-placebo parameters $\theta_1(0,t)$ and $\gamma_1(0,t)$.

To mimic the setup where the size of the control arm is much smaller than the size of the treatment arm, we randomly discarded 75% of the controls from SUSTAIN-6 and repeated the analysis. The resulting fusion estimators for the treatment effects exhibited some deviation from the RCT-only estimators, but a large precision gain at approximately 45–50% was observed; see Tables S10 and S11. Finally, to evaluate the impact of omitted variable bias, we performed a sensitivity analysis by removing history of cardiovascular diseases from the set of baseline covariates. While the reduction in standard errors was twice as large compared to the main analysis, there was also a more substantial difference between point estimates of $\theta_1\{104\}$ and $\gamma_1\{t\}$; see Tables S8 and S9. It is thus unclear whether the transportability assumption holds at all with this restricted set of baseline covariates. Further details on the data example and results from the additional analysis are available in Supplementary Material §S4.

Table 4. *Cumulative incidence differences in the real data example.*

Estimand	t (weeks)	Type	Estimate (%)	95%-CI (%)	Reduction
$\theta_1\{t\}$	26	+	−0.26	(−1.27, 0.75)	8.31
		−	−0.39	(−1.49, 0.71)	.
	52	+	−0.61	(−1.87, 0.64)	8.65
		−	−0.73	(−2.10, 0.65)	.
	78	+	−1.99	(−3.40, −0.57)	9.18
		−	−1.78	(−3.34, −0.22)	.
	104	+	−2.72	(−4.33, −1.12)	9.76
		−	−2.56	(−4.33, −0.78)	.
$\theta_2\{t\}$	26	+	−0.49	(−1.05, 0.07)	−0.00
		−	−0.49	(−1.05, 0.07)	.
	52	+	−0.38	(−1.26, 0.50)	−0.00
		−	−0.38	(−1.26, 0.50)	.
	78	+	0.05	(−1.17, 1.28)	−0.00
		−	0.05	(−1.17, 1.28)	.
	104	+	−0.21	(−1.71, 1.28)	−0.00
		−	−0.21	(−1.71, 1.28)	.

Type: fusion estimator (+) or RCT-only estimator (−); CI: confidence interval; Reduction: percentage reduction CI length, %.

Table 5. *Restricted mean time lost differences in the real data example.*

Estimand	t (weeks)	Type	Estimate (weeks)	95%-CI (weeks)	Reduction
$\gamma_1\{t\}$	26	+	−0.10	(−0.23, 0.03)	10.32
		−	−0.11	(−0.25, 0.04)	.
	52	+	−0.22	(−0.62, 0.18)	9.02
		−	−0.26	(−0.70, 0.18)	.
	78	+	−0.59	(−1.31, 0.13)	8.87
		−	−0.62	(−1.41, 0.18)	.
	104	+	−1.15	(−2.24, −0.07)	9.00
		−	−1.14	(−2.33, 0.04)	.
$\gamma_2\{t\}$	26	+	−0.05	(−0.13, 0.03)	−0.00
		−	−0.05	(−0.13, 0.03)	.
	52	+	−0.12	(−0.37, 0.13)	−0.00
		−	−0.12	(−0.37, 0.13)	.
	78	+	−0.21	(−0.68, 0.27)	−0.00
		−	−0.21	(−0.68, 0.27)	.
	104	+	−0.24	(−1.02, 0.53)	−0.00
		−	−0.24	(−1.02, 0.53)	.

Type: fusion estimator (+) or RCT-only estimator (−); CI: confidence interval; Reduction: percentage reduction CI length, %.

6. Discussion

In this work, we assume transportability of the conditional cause-specific hazard of the event of interest between the RCT population and the external control population. We have considered estimation of the cumulative incidence functions and restricted mean times lost with external controls. In fact, this assumption also allows us to derive estimators with improved precision for other estimands in competing risks analysis, including the average hazard with survival weights (Uno and Horiguchi, 2023) and separable effects (Stensrud et al., 2022). We comment in Supplementary Material §S5 that weaker transportability assumptions for competing risks analysis can be difficult to interpret.

In practice, the risk of introducing bias to RCT data when incorporating external controls should be evaluated. One approach is to carry out the analysis with the data fusion estimator for precision gain. Then, post-hoc model diagnostics such as likelihood ratio tests or other omnibus tests may be performed to assess the validity of the transportability assumption. A more principled approach may be to integrate the estimated bias to make an informed decision of whether the RCT-only estimator should be retained. For instance, following Yang et al. (2023), a test-then-pool estimator for the cumulative incidence $\theta_1(0)$ can be constructed from the estimators $\hat{\theta}_1(0)$ with and without external controls via a score test. This is left for future work.

We focus on treatment policy estimands, which ignore treatment trajectories after randomization. Consequently, we do not adjust for post-baseline variables. However, by omitting these variables, we may fail to establish transportability of the cause-specific hazard. In SUSTAIN-6 and LEADER, when subjects experienced non-fatal adverse events, they could receive drop-in medications. If these decisions were based on different treatment guidelines and policies between the two populations, subjects with similar baseline characteristics might have rather different event rates. This is a particular concern for using historical controls in RCTs. Future research can focus on fusion estimates that allow for history beyond baseline.

References

Andersen, P. K. (2013). Decomposition of number of life years lost according to causes of death. *Statistics in Medicine*, 32(30):5278–5285.

Benkeser, D., Carone, M., and Gilbert, P. B. (2018). Improved estimation of the cumulative incidence of rare outcomes. *Statistics in Medicine*, 37(2):280–293.

Cao, Z., Cho, Y., and Li, F. (2024). Transporting randomized trial results to estimate counterfactual survival functions in target populations. *Pharmaceutical Statistics*, 23(4):442–465.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. (2024). Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1):165–191.

Dahabreh, I. J., Haneuse, S. J.-P. A., Robins, J. M., Robertson, S. E., Buchanan, A. L., Stuart, E. A., and Hernán, M. A. (2021). Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology*, 190(8):1632–1642.

Dang, L. E., Fong, E., Tarp, J. M., Clemmensen, K. K. B., Ravn, H., Kvist, K., Buse, J. B., van der Laan, M., and Petersen, M. (2023). Case study of semaglutide and cardiovascular outcomes: An application of the Causal Roadmap to a hybrid design for augmenting an RCT control arm with real-world data. *Journal of Clinical and Translational Science*, 7(1):e231.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):245–264.

European Medicines Agency (2023). Reflection paper on establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation: Considerations on evidence from single-arm trials. https://www.ema.europa.eu/en/documents/scientific-guideline/draft-reflection-paper-establishing-efficacy-based-single-arm-trials-submitted-pivotal-evidence-marketing-authorisation_en.pdf.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley Series in Probability and Statistics. Wiley.

Food and Drug Administration (2023). Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products.

Kwiatkowski, E., Zhu, J., Li, X., Pang, H., Lieberman, G., and Psioda, M. A. (2024). Case weighted power priors for hybrid control analyses with time-to-event data. *Biometrics*, 80(2):ujae019.

Lee, D., Yang, S., and Wang, X. (2022). Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *Journal of Causal Inference*, 10(1):415–440.

Li, H., Tiwari, R., and Li, Q. H. (2022). Conditional borrowing external data to establish a hybrid control arm in randomized clinical trials. *Journal of Biopharmaceutical Statistics*, 32(6):954–968.

Marso, S. P., Bain, S. C., Consoli, A., Eliaschewitz, F. G., Jódar, E., Leiter, L. A., Lingvay, I., Rosenstock, J., Seufert, J., Warren, M. L., Woo, V., Hansen, O., Holst, A. G., Pettersson, J., and Vilsbøll, T. (2016a). Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine*, 375(19):1834–1844.

Marso, S. P., Daniels, G. H., Brown-Frandsen, K., Kristensen, P., Mann, J. F., Nauck, M. A., Nissen, S. E., Pocock, S., Poulter, N. R., Ravn, L. S., Steinberg, W. M., Stockner, M., Zinman, B., Bergenstal, R. M., and Buse, J. B. (2016b). Liraglutide and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, 375(4):311–322.

Ramagopalan, S. V., Popat, S., Gupta, A., Boyne, D. J., Lockhart, A., Hsu, G., O'Sullivan, D. E., Inskip, J., Ray, J., Cheung, W. Y., Griesinger, F., and Subbiah, V. (2022). Transportability of overall survival estimates from US to Canadian patients with advanced non–small cell lung cancer with implications for regulatory and health technology assessment. *JAMA Network Open*, 5(11):e2239874.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. Working Paper 128, Center for Statistics and the Social Sciences, University of Washington.

Rytgaard, H. C. W., Eriksson, F., and van der Laan, M. J. (2023). Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*, 79(4):3038–3049.

Sengupta, S., Ntambwe, I., Tan, K., Liang, Q., Paulucci, D., Castellanos, E., Fiore, J., Lane, S., Micsinai Balan, M., Viraswami-Apanna, K., Sethuraman, V., Samant, M., and Tiwari, R. (2023). Emulating randomized controlled trials with hybrid control arms in oncology: A case study. *Clinical Pharmacology & Therapeutics*, 113(4):867–877.

Stensrud, M. J., Young, J. G., Didelez, V., Robins, J. M., and Hernán, M. A. (2022). Separable effects for causal inference in the presence of competing events. *Journal of the American Statistical Association*, 117(537):175–183.

Tan, W. K., Segal, B. D., Curtis, M. D., Baxi, S. S., Capra, W. B., Garrett-Mayer, E., Hobbs, B. P., Hong, D. S., Hubbard, R. A., Zhu, J., Sarkar, S., and Samant, M. (2022). Augmenting control

arms with real-world data for cancer trials: Hybrid control arm methods and considerations. *Contemporary Clinical Trials Communications*, 30:101000.

Uno, H. and Horiguchi, M. (2023). Ratio and difference of average hazard with survival weight: New measures to quantify survival benefit of new therapy. *Statistics in Medicine*, 42(7):936–952.

Westling, T., Luedtke, A., Gilbert, P. B., and Carone, M. (2024). Inference for Treatment-Specific Survival Curves Using Machine Learning. *Journal of the American Statistical Association*, 119(546):1541–1553.

Yang, S., Gao, C., Zeng, D., and Wang, X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):575–596.

Zuo, S., Josey, K. P., Raghavan, S., Yang, F., Juaréz-Colunga, E., and Ghosh, D. (2022). Transportability methods for time-to-event outcomes: Application in adjuvant colon cancer trials. *JCO Clinical Cancer Informatics*, 6:1–14.

Supplementary material for "Improving precision of cumulative incidence estimates in randomized controlled trials with external controls"

Zehao Su
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
zehao.su@sund.ku.dk

Helene C. Rytgaard
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
hely@sund.ku.dk

Henrik Ravn
Novo Nordisk A/S,
Vandtårnsvej 108, DK-2860 Søborg, Denmark
hnrv@novonordisk.com

Frank Eriksson
Section of Biostatistics, University of Copenhagen,
Øster Farimagsgade 5, DK-1353 Copenhagen, Denmark
eriksson@sund.ku.dk

S1. Notations on the observed data distribution

In the main text, we make use of three models: the counterfactual data distribution, the uncensored data distribution, and the observed data distribution. All three models encompass the population indicator D, the baseline covariates X, and the treatment DA, whereas the counterfactual data distribution contains the potential outcomes $\{T(1), T(0), J(1), J(0)\}$, the uncensored data distribution contains the uncensored event time and event type plus the censoring time (T, J, C), and the observed data distribution contains the censored event time and event type $(\tilde{T}, \tilde{J})$. With Assumption 2, we identify the causal parameters in the uncensored data distribution. We now connect the observed data quantities to their uncensored counterparts using Assumption 3.

Recall the (observed) event counting process $N_j(t) = I(\tilde{T} \leqslant t, \tilde{J} = j)$ for $j = 1, 2$. Let $N^c(t) = I(\tilde{T} \leqslant t, \tilde{J} = 0)$ be the censoring counting process. Define $(\mathcal{F}_t)_{t\in(0,\tau]}$ as the filtration in which the σ-algebra $\mathcal{F}_t = \sigma\big[\{N_1(s), N_2(s), N^c(s), DA, X, D; 0 < s \leqslant t\}\big]$ contains the observed information up to time t (inclusive). The event counting process $N_j(t)$ has a compensator such that

$$\tilde{M}_{Dj}(t \mid A, X) = N_j(t) - \tilde{\Lambda}_{Dj}(t \mid A, X), \qquad D = 1$$
$$\tilde{M}_{Dj}(t \mid X) = N_j(t) - \tilde{\Lambda}_{Dj}(t \mid X), \qquad D = 0,$$

is a martingale adapted to $(\mathcal{F}_t)$. Standard results in time-to-event analysis shows that the compensator satisfies a multiplicative hazard structure such that the increment of the compensator factorizes as

$$\mathrm{d}\tilde{\Lambda}_{Dj}(t \mid A, X) = I(\tilde{T} \geqslant t)\mathrm{d}\tilde{\Lambda}_{Dj}(t \mid A, X), \qquad D = 1,$$
$$\mathrm{d}\tilde{\Lambda}_{Dj}(t \mid X) = I(\tilde{T} \geqslant t)\mathrm{d}\tilde{\Lambda}_{Dj}(t \mid X), \qquad D = 0,$$

where

$$\mathrm{d}\tilde{\mathrm{A}}_{1j}(t \mid a, x) = \frac{\mathrm{d}P(\tilde{T} \leqslant t, \tilde{J} = j \mid A = a, X = x, D = 1)}{P(\tilde{T} \geqslant t \mid A = a, X = x, D = 1)},$$
$$\mathrm{d}\tilde{\mathrm{A}}_{0j}(t \mid x) = \frac{\mathrm{d}P(\tilde{T} \leqslant t, \tilde{J} = j \mid X = x, D = 0)}{P(\tilde{T} \geqslant t \mid X = x, D = 0)}.$$

Consider the filtration $(\mathcal{G}_t)$ where $\mathcal{G}_t = \sigma\big[\{N_1(s+), N_2(s+), N^c(s), DA, X, D; 0 < s \leqslant t\}\big]$. The quantity associated with the observed censoring counting process

$$D\tilde{M}^c_D(t \mid A, X) + (1 - D)\tilde{M}^c_D(t \mid X)$$

is a martingale adapted to $(\mathcal{G}_t)$, where $\tilde{M}^c_D(t \mid A, X) = N^c(t) - \tilde{\Lambda}^c_D(t \mid A, X)$, $\tilde{M}^c_D(t \mid X) = N^c(t) - \tilde{\Lambda}^c_D(t \mid X)$,

$$\mathrm{d}\tilde{\Lambda}^c_D(t \mid A, X) = \big\{I(\tilde{T} \geqslant t, \tilde{J} = 0) + I(\tilde{T} > t, \tilde{J} \neq 0)\big\}\mathrm{d}\tilde{\mathrm{A}}^c_D(t \mid A, X)$$

and

$$\mathrm{d}\tilde{\mathrm{A}}^c_1(t \mid a, x) = \frac{\mathrm{d}P(\tilde{T} \leqslant t, \tilde{J} = 0 \mid A = a, X = x, D = 1)}{P[\{\tilde{T} \geqslant t, \tilde{J} = 0\} \cup \{\tilde{T} > t, \tilde{J} \neq 0\} \mid A = a, X = x, D = 1]},$$
$$\mathrm{d}\tilde{\mathrm{A}}^c_0(t \mid x) = \frac{\mathrm{d}P(\tilde{T} \leqslant t, \tilde{J} = 0 \mid X = x, D = 0)}{P[\{\tilde{T} \geqslant t, \tilde{J} = 0\} \cup \{\tilde{T} > t, \tilde{J} \neq 0\} \mid X = x, D = 0]}.$$

Under Assumption 3, the cause-specific hazards defined on the uncensored data distribution are identifiable from the observed data with $\tilde{\mathrm{A}}_{dj}(t \mid a, x) = \mathrm{A}_{dj}(t \mid a, x)$, and so is the censoring hazard with $\tilde{\mathrm{A}}^c_d(t \mid a, x) = \mathrm{A}^c_d(t \mid a, x)$. Therefore, the survival function of the composite event and the cumulative incidence function of event type j is subsequently identifiable in the observed data distribution as the product integral

$$S_d(t \mid a, x) = \{\Pi(\mathrm{A}_{d1} + \mathrm{A}_{d2})\}(t \mid a, x) = \big\{\Pi\big(\tilde{\mathrm{A}}_{d1} + \tilde{\mathrm{A}}_{d2}\big)\big\}(t \mid a, x) = \tilde{S}_d(t \mid a, x)$$

and the Lebesgue-Stieltjes integral

$$F_{dj}(t \mid a, x) = \int_0^t S_d(s- \mid a, x)\mathrm{d}\mathrm{A}_{dj}(s \mid a, x) = \int_0^t \tilde{S}_d(s- \mid a, x)\mathrm{d}\tilde{\mathrm{A}}_{dj}(s \mid a, x) = \tilde{F}_{dj}(t \mid a, x).$$

S2. Proofs

S2.1. *Proof of Lemma 1*

We show the efficient influence function of $\theta_1(0)$ without Assumption 5. Define

$$G_1(t \mid a, x) = P(\tilde{T} > t \mid A = a, X = x, D = 1),$$
$$G_0(t \mid x) = P(\tilde{T} > t \mid X = x, D = 0),$$
$$\tilde{F}_{1j}(t \mid a, x) = \int_0^t \tilde{S}_1(s- \mid a, x)\mathrm{d}\tilde{\mathrm{A}}_{1j}(s \mid a, x).$$

Lemma 1 is a special case of the following result.

Proposition S1. *The efficient influence function of $\theta_1(0)$ at $P \in \mathcal{P}$ is*

$$\varphi_1(0)(O) = \frac{D(1 - A)\pi(X)}{\alpha} \int_0^\tau \frac{w_1(t \mid X)}{G_1(t- \mid A, X)} g_{11}(t \mid A, X)\mathrm{d}\tilde{M}_{11}(t \mid A, X)$$

$$+\frac{(1-D)\pi(X)}{\alpha}\int_0^\tau \frac{w_0(t\,|\,X)}{G_0(t-\,|\,X)}g_{11}(t\,|\,0,X)\mathrm{d}\tilde{M}_{01}(t\,|\,X)$$
$$+\frac{D}{\alpha}\frac{(1-A)}{e_1(0\,|\,X)}\int_0^\tau \frac{1}{G_1(t-\,|\,A,X)}g_{21}(t\,|\,A,X)\mathrm{d}\tilde{M}_{12}(t\,|\,A,X)$$
$$+\frac{D}{\alpha}\{\tilde{F}_{11}(\tau\,|\,0,X)-\theta_1(0)\}, \tag{S1}$$

where

$$g_{k1}(t\,|\,A,X)=I(k=1)\tilde{S}_1(t-\,|\,A,X)-\frac{\tilde{F}_{11}(\tau\,|\,A,X)-\tilde{F}_{11}(t\,|\,A,X)}{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,A,X)-\triangle\tilde{\mathrm{A}}_{12}(t\,|\,A,X)}$$
$$w_\bullet(t\,|\,X)=\{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,X)\}g_{11}(t\,|\,0,X)\left\{\frac{1-\pi(X)}{G_1(t-\,|\,0,X)}+\frac{\pi(X)e_1(0\,|\,X)}{G_0(t-\,|\,X)}\right\}$$
$$w_1(t\,|\,X)=\frac{1}{w_\bullet(t\,|\,X)}\left\{\frac{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,X)}{G_0(t-\,|\,X)}g_{11}(t\,|\,0,X)\right.$$
$$\left.+\frac{\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,X)}{G_1(t-\,|\,0,X)}\frac{1-\pi(X)}{\pi(X)e_1(0\,|\,X)}g_{21}(t\,|\,0,X)\right\}$$
$$w_0(t\,|\,X)=\frac{1}{w_\bullet(t\,|\,X)}\left\{\frac{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,X)}{G_1(t-\,|\,0,X)}g_{11}(t\,|\,0,X)-\frac{\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,X)}{G_1(t-\,|\,0,X)}g_{21}(t\,|\,0,X)\right\}.$$

The efficient influence function of $\theta_2(0)$ at $P\in\mathcal{P}$ is

$$\varphi_2(0)(O)=\frac{D(1-A)\pi(X)}{\alpha}\int_0^\tau \frac{w_1(t\,|\,X)}{G_1(t-\,|\,A,X)}g_{12}(t\,|\,A,X)\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)$$
$$+\frac{(1-D)\pi(X)}{\alpha}\int_0^\tau \frac{w_0(t\,|\,X)}{G_0(t-\,|\,X)}g_{12}(t\,|\,0,X)\mathrm{d}\tilde{M}_{01}(t\,|\,X)$$
$$+\frac{D}{\alpha}\frac{(1-A)}{e_1(0\,|\,X)}\int_0^\tau \frac{1}{G_1(t-\,|\,A,X)}g_{22}(t\,|\,A,X)\mathrm{d}\tilde{M}_{12}(t\,|\,A,X)$$
$$+\frac{D}{\alpha}\{\tilde{F}_{12}(\tau\,|\,0,X)-\theta_2(0)\}. \tag{S2}$$

Proof. We define the observed data distribution for the censoring time as $Q_{10}(t\,|\,a,x)=P(\tilde{T}\leqslant t,\tilde{J}=0\,|\,A=a,X=x,D=1)$ and $Q_{00}(t\,|\,x)=P(\tilde{T}\leqslant t,\tilde{J}=0\,|\,X=x,D=0)$ for $a\in\{0,1\}$. The observed data density can be factorized as

$$\mathrm{d}P(\tilde{T},\tilde{J},A,X,D)=\{\mathrm{d}Q_{1\tilde{J}}(\tilde{T}\,|\,A,X)e_1(A\,|\,X)\pi(X)\}^D[\mathrm{d}Q_{0\tilde{J}}(\tilde{T}\,|\,X)\{1-\pi(X)\}]^{1-D}\mathrm{d}P(X).$$

Consider the parametric submodel for the observed data density for the event time and event type:

$$\mathrm{d}Q_{1k}(t\,|\,a,x;\varepsilon)=\mathrm{d}Q_{1k}(t\,|\,a,x)\{1+\varepsilon h_1(t,k,a,x)\},$$
$$\mathrm{d}Q_{0k}(t\,|\,x;\varepsilon)=\mathrm{d}Q_{0k}(t\,|\,x)\{1+\varepsilon h_0(t,k,x)\},$$

for $a\in\{0,1\}$ and $k\in\{0,1,2\}$, where $h_1(\tilde{T},\tilde{J},A,X)$ and $h_0(\tilde{T},\tilde{J},X)$ are functions with finite variance that satisfy $E\{h_1(\tilde{T},\tilde{J},A,X)\,|\,A,X,D=1\}=0$ and $E\{h_0(\tilde{T},\tilde{J},X)\,|\,X,D=0\}=0$. The submodel must further obey the restriction $\mathrm{d}\tilde{\mathrm{A}}_{11}(t\,|\,0,x;\varepsilon)=\mathrm{d}\tilde{\mathrm{A}}_{01}(t\,|\,x;\varepsilon)$ for $t\in(0,\tau]$, or equivalently,

$$\frac{\mathrm{d}Q_{11}(t\,|\,0,x;\varepsilon)}{G_1(t-\,|\,0,x;\varepsilon)}=\frac{\mathrm{d}Q_{01}(t\,|\,x;\varepsilon)}{G_0(t-\,|\,x;\varepsilon)}. \tag{S3}$$

The Gateaux derivative of the cumulative hazard increment $\mathrm{d}\tilde{\mathrm{A}}_{1k}(t\,|\,a,x;\varepsilon)$ for $k\in\{0,1,2\}$ is

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{d}\tilde{\mathrm{A}}_{1k}(t\,|\,a,x;\varepsilon)\Big|_{\varepsilon=0}$$

$$
\begin{aligned}
&= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\frac{\mathrm{d}Q_{1k}(t \mid a,x;\varepsilon)}{G_1(t- \mid a,x;\varepsilon)}\bigg|_{\varepsilon=0} \\
&= \frac{1}{G_1(t- \mid a,x)}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{d}Q_{1k}(t \mid a,x;\varepsilon)\bigg|_{\varepsilon=0} - \frac{\mathrm{d}Q_{1k}(t \mid a,x)}{G_1^2(t- \mid a,x)}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}G_1(t- \mid a,x;\varepsilon)\bigg|_{\varepsilon=0} \\
&= \frac{1}{G_1(t- \mid a,x)}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{d}Q_{1k}(t \mid a,x;\varepsilon)\bigg|_{\varepsilon=0} \\
&\quad - \frac{\mathrm{d}Q_{1k}(t \mid a,x)}{G_1^2(t- \mid a,x)}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\int_{u\in[t,\infty)}\sum_{j\in\{0,1,2\}}\mathrm{d}Q_{1j}(u \mid a,x;\varepsilon)\bigg|_{\varepsilon=0} \\
&= \mathrm{d}\tilde{\mathrm{A}}_{1k}(t \mid a,x)\Big\{h_1(t,k,a,x) - \sum_{j\in\{0,1,2\}}\int_{u\in[0,\infty)} h_1(u,j,a,x)\frac{\mathrm{d}Q_{1j}(u \mid a,x)}{G_1(t- \mid a,x)}\Big\}.
\end{aligned}
$$

Similarly, for $k \in \{0,1,2\}$, we have

$$
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{d}\tilde{\mathrm{A}}_{0k}(t \mid x;\varepsilon)\bigg|_{\varepsilon=0} = \mathrm{d}\tilde{\mathrm{A}}_{0k}(t \mid x)\Big\{h_0(t,k,x) - \sum_{j\in\{0,1,2\}}\int_{u\in[t,\infty)} h_0(u,j,x)\frac{\mathrm{d}Q_{0j}(u \mid x)}{G_0(t- \mid x)}\Big\}.
$$

Therefore, differentiating both sides of (S3) with respect to ε and evaluating at zero, the restriction on the scores of the hazards is

$$
\begin{aligned}
&\mathrm{d}\tilde{\mathrm{A}}_{11}(t \mid 0,x)\Big\{h_1(t,1,0,x) - \int_{u\in[t,\infty)}\sum_{k\in\{0,1,2\}} h_1(u,k,0,x)\frac{\mathrm{d}Q_{1k}(u \mid 0,x)}{G_1(t- \mid 0,x)}\Big\} \\
&\qquad = \mathrm{d}\tilde{\mathrm{A}}_{01}(t \mid x)\Big\{h_0(t,1,x) - \int_{u\in[t,\infty)}\sum_{k\in\{0,1,2\}} h_0(u,k,x)\frac{\mathrm{d}Q_{0k}(u \mid x)}{G_0(t- \mid x)}\Big\}.
\end{aligned}
$$

With some algebra, the score restriction can be expressed as a conditional expectation restriction:

$$
\begin{aligned}
&E\left[h_1(\tilde{T},\tilde{J},A,X)\frac{\mathrm{d}\tilde{M}_{11}(t \mid A,X)}{G_1(t- \mid A,X)}\,\middle|\, A=0,X,D=1\right] \\
&\qquad\qquad = E\left[h_0(\tilde{T},\tilde{J},X)\frac{\mathrm{d}\tilde{M}_{01}(t \mid X)}{G_0(t- \mid X)}\,\middle|\, X,D=0\right]. \quad \text{(S4)}
\end{aligned}
$$

For the rest of the components in the observed data density, we also choose appropriate perturbation functions, or score functions, such that $\mathrm{d}P(\tilde{T},\tilde{J},A,X,D;\varepsilon)$ equals $\mathrm{d}P(\tilde{T},\tilde{J},A,X,D)$ when $\varepsilon = 0$. The closed linear subspace of all possible choices of these perturbation functions is the tangent space of the model $\mathcal{P}$ at P, which is

$$
\dot{\mathcal{P}} = \dot{\mathcal{P}}_1 \oplus \dot{\mathcal{P}}_2 \oplus \dot{\mathcal{P}}_3 \oplus \dot{\mathcal{P}}_4,
$$

where $\mathcal{V}_1 \oplus \mathcal{V}_2$ denotes the direct sum of vector spaces $\mathcal{V}_1$ and $\mathcal{V}_2$,

$$
\begin{aligned}
\dot{\mathcal{P}}_1 = \big\{&Dh_1(\tilde{T},\tilde{J},A,X) + (1-D)h_0(\tilde{T},\tilde{J},X) : E\{h_1(\tilde{T},\tilde{J},A,X) \mid A,X,D=1\} = 0, \\
&E\{h_0(\tilde{T},\tilde{J},X) \mid X,D=0\} = 0, h_1(\tilde{T},\tilde{J},A,X) \text{ and } h_0(\tilde{T},\tilde{J},X) \text{ satisfy (S4)}\big\}, \\
\dot{\mathcal{P}}_2 = \big\{&Dh_1(A,X) : E\{h_1(A,X) \mid X,D=1\} = 0\big\}, \\
\dot{\mathcal{P}}_3 = \big\{&h(D,X) : E\{h(D,X) \mid X\} = 0\big\}, \\
\dot{\mathcal{P}}_4 = \big\{&h(X) : E\{h(X)\} = 0\big\}.
\end{aligned}
$$

The decomposition of the tangent space follows from the product structure of the observed data likelihood. Differentiating the target parameter along some submodel $\{P_\varepsilon\}$ with score function

$$
h(O) = Dh_1(\tilde{T},\tilde{J},A,X) + (1-D)h_0(\tilde{T},\tilde{J},X) + Dh(A,X) + h(D,X) + h(X),
$$

we have

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_1(0;\varepsilon)\Big|_{\varepsilon=0} \\
&\quad= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\int_{\mathcal{X}}\int_0^\tau \mathrm{d}\tilde{\mathrm{A}}_{11}(t\,|\,0,x;\varepsilon)\tilde{S}_1(t-\,|\,0,x;\varepsilon)\mathrm{d}P(x\,|\,D=1;\varepsilon)\Big|_{\varepsilon=0} \\
&\quad= \int_{\mathcal{X}}\int_0^\tau \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{d}\tilde{\mathrm{A}}_{11}(t\,|\,0,x;\varepsilon)\Big|_{\varepsilon=0}\tilde{S}_1(t-\,|\,0,x)\mathrm{d}P(x\,|\,D=1) \qquad \text{(S5)} \\
&\qquad - \int_{\mathcal{X}}\int_0^\tau\int_0^{t-}\frac{\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{d}\tilde{\mathrm{A}}_{11}(s\,|\,0,x;\varepsilon)\Big|_{\varepsilon=0}+\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{d}\tilde{\mathrm{A}}_{12}(s\,|\,0,x;\varepsilon)\Big|_{\varepsilon=0}}{1-\triangle\tilde{\mathrm{A}}_{11}(s\,|\,0,x)-\triangle\tilde{\mathrm{A}}_{12}(s\,|\,0,x)} \\
&\qquad\qquad \mathrm{d}\tilde{F}_{11}(t\,|\,0,x)\mathrm{d}P(x\,|\,D=1) \qquad \text{(S6)} \\
&\qquad + \int_{\mathcal{X}}\tilde{F}_{11}(\tau\,|\,0,x)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{d}P(x\,|\,D=1;\varepsilon)\Big|_{\varepsilon=0}, \qquad \text{(S7)}
\end{aligned}
$$

where the outermost integral is over the set $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_0$.

We proceed by analyzing the terms separately. First, the term (S5) is

$$
\begin{aligned}
&\int_{\mathcal{X}}\int_0^\tau h_1(t,1,0,x)\mathrm{d}\tilde{\mathrm{A}}_{11}(t\,|\,0,x)\tilde{S}_1(t-\,|\,0,x)\mathrm{d}P(x\,|\,D=1) \\
&\quad - \int_{\mathcal{X}}\int_0^\tau\int_{u\in[t,\infty)}\sum_{k\in\{0,1,2\}}h_1(u,k,0,x)\frac{\mathrm{d}Q_{1k}(u\,|\,0,x)}{G_1(t-\,|\,0,x)} \\
&\qquad\qquad \mathrm{d}\tilde{\mathrm{A}}_{11}(t\,|\,0,x)\tilde{S}_1(t-\,|\,0,x)\mathrm{d}P(x\,|\,D=1) \\
&\quad = \int_{\mathcal{X}} E\left[h_1(\tilde{T},\tilde{J},A,X)\int_0^\tau\frac{\tilde{S}_1(t-\,|\,A,X)}{G_1(t-\,|\,A,X)}\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)\,\middle|\,A=0,X=x,D=1\right] \\
&\qquad\qquad \mathrm{d}P(x\,|\,D=1).
\end{aligned}
$$

The term (S6) is a sum of two terms, which for $k \in \{1,2\}$ can be seen to be

$$
\begin{aligned}
&-\int_{\mathcal{X}}\int_0^\tau\int_{s\in(0,t)}\frac{\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{d}\tilde{\mathrm{A}}_{1k}(s\,|\,0,x;\varepsilon)\Big|_{\varepsilon=0}}{1-\triangle\tilde{\mathrm{A}}_{11}(s\,|\,0,x)-\triangle\tilde{\mathrm{A}}_{12}(s\,|\,0,x)}\mathrm{d}\tilde{F}_{11}(t\,|\,0,x)\mathrm{d}P(x\,|\,D=1) \\
&\quad = -\int_{\mathcal{X}}\int_0^\tau\int_{s\in(0,t)}\frac{h_1(s,k,0,x)\mathrm{d}\tilde{\mathrm{A}}_{1k}(s\,|\,0,x)}{1-\triangle\tilde{\mathrm{A}}_{11}(s\,|\,0,x)-\triangle\tilde{\mathrm{A}}_{12}(s\,|\,0,x)}\mathrm{d}\tilde{F}_{11}(t\,|\,0,x)\mathrm{d}P(x\,|\,D=1) \\
&\qquad + \int_{\mathcal{X}}\int_0^\tau\int_{s\in(0,t)}\frac{\left\{\int_{u\in[s,\infty)}\sum_{j\in\{0,1,2\}}h_1(u,j,0,x)\mathrm{d}Q_{1j}(u\,|\,0,x)\right\}}{G_1(s-\,|\,0,x)\{1-\triangle\tilde{\mathrm{A}}_{11}(s\,|\,0,x)-\triangle\tilde{\mathrm{A}}_{12}(s\,|\,0,x)\}} \\
&\qquad\qquad \mathrm{d}\tilde{\mathrm{A}}_{1k}(s\,|\,0,x)\mathrm{d}\tilde{F}_{11}(t\,|\,0,x)\mathrm{d}P(x\,|\,D=1) \\
&\quad = -\int_{\mathcal{X}}\int_{s\in(0,\tau)}\int_{t\in(s,\tau]}\mathrm{d}\tilde{F}_{11}(t\,|\,0,x)\frac{h_1(s,k,0,x)\mathrm{d}\tilde{\mathrm{A}}_{1k}(s\,|\,0,x)}{1-\triangle\tilde{\mathrm{A}}_{11}(s\,|\,0,x)-\triangle\tilde{\mathrm{A}}_{12}(s\,|\,0,x)}\mathrm{d}P(x\,|\,D=1) \\
&\qquad + \int_{\mathcal{X}}\int_{s\in(0,\tau)}\int_{t\in(s,\tau]}\mathrm{d}\tilde{F}_{11}(t\,|\,0,x) \\
&\qquad\qquad \frac{\left\{\int_{u\in[s,\infty)}\sum_{j\in\{0,1,2\}}h_1(u,j,0,x)\mathrm{d}Q_{1j}(u\,|\,0,x)\right\}}{G_1(s-\,|\,0,x)\{1-\triangle\tilde{\mathrm{A}}_{11}(s\,|\,0,x)-\triangle\tilde{\mathrm{A}}_{12}(s\,|\,0,x)\}} \\
&\qquad\qquad \mathrm{d}\tilde{\mathrm{A}}_{1k}(s\,|\,0,x)\mathrm{d}P(x\,|\,D=1)
\end{aligned}
$$

$$
\begin{aligned}
&= -\int_{\mathcal{X}} E\Bigg[h_1(\tilde{T},\tilde{J},A,X)\int_{s\in(0,\tau)}\frac{\tilde{F}_{11}(\tau\mid A,X)-\tilde{F}_{11}(s\mid A,X)}{1-\triangle\tilde{\mathrm{A}}_{11}(s\mid A,X)-\triangle\tilde{\mathrm{A}}_{12}(s\mid A,X)}\\
&\qquad\qquad\frac{\mathrm{d}N_k(s)}{G_1(s-\mid A,X)}\;\Bigg|\;A=0,X=x,D=1\Bigg]\mathrm{d}P(x\mid D=1)\\
&\quad+\int_{\mathcal{X}} E\Bigg[h_1(\tilde{T},\tilde{J},A,X)\int_{s\in(0,\tau)}\frac{\tilde{F}_{11}(\tau\mid A,X)-\tilde{F}_{11}(s\mid A,X)}{1-\triangle\tilde{\mathrm{A}}_{11}(s\mid A,X)-\triangle\tilde{\mathrm{A}}_{12}(s\mid A,X)}\\
&\qquad\qquad\frac{I(\tilde{T}\geqslant s)\mathrm{d}\tilde{\mathrm{A}}_{1k}(s\mid A,X)}{G_1(s-\mid A,X)}\;\Bigg|\;A=0,X=x,D=1\Bigg]\mathrm{d}P(x\mid D=1)\\
&= -\int_{\mathcal{X}} E\Bigg[h_1(\tilde{T},\tilde{J},A,X)\int_0^\tau\frac{\tilde{F}_{11}(\tau\mid A,X)-\tilde{F}_{11}(t\mid A,X)}{G_1(t-\mid A,X)\{1-\triangle\tilde{\mathrm{A}}_{11}(t\mid A,X)-\triangle\tilde{\mathrm{A}}_{12}(t\mid A,X)\}}\\
&\qquad\qquad\mathrm{d}\tilde{M}_{1k}(t\mid A,X)\;\Bigg|\;A=0,X=x,D=1\Bigg]\mathrm{d}P(x\mid D=1).
\end{aligned}
$$

The last term (S7) is

$$
\int_{\mathcal{X}}\{\tilde{F}_{11}(\tau\mid 0,x)-\theta_1(0)\}\{h(1,x)+h(x)\}\mathrm{d}P(x\mid D=1).
$$

Collecting the terms yields that

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_1(0;\varepsilon)\Big|_{\varepsilon=0}\\
&\quad=\int_{\mathcal{X}} E\Bigg[\int_0^\tau\Bigg\{g_{11}(t\mid A,X)\frac{\mathrm{d}\tilde{M}_{11}(t\mid A,X)}{G_1(t-\mid A,X)}+g_{21}(t\mid A,X)\frac{\mathrm{d}\tilde{M}_{12}(t\mid A,X)}{G_1(t-\mid A,X)}\Bigg\}\\
&\qquad\qquad h_1(\tilde{T},\tilde{J},A,X)\;\Bigg|\;A=0,X=x,D=1\Bigg]\mathrm{d}P(x\mid D=1)\\
&\qquad+\int_{\mathcal{X}}\{\tilde{F}_{11}(\tau\mid 0,x)-\theta_1(0)\}\{h(1,x)+h(x)\}\mathrm{d}P(x\mid D=1),
\end{aligned}
$$

and by replacing integrals with expectations, we have

$$
\begin{aligned}
&= E\Bigg[\frac{D(1-A)}{\alpha e_1(A\mid X)}h_1(\tilde{T},\tilde{J},A,X)\\
&\qquad\qquad\int_0^\tau\Bigg\{g_{11}(t\mid A,X)\frac{\mathrm{d}\tilde{M}_{11}(t\mid A,X)}{G_1(t-\mid A,X)}+g_{12}(t\mid A,X)\frac{\mathrm{d}\tilde{M}_{12}(t\mid A,X)}{G_1(t-\mid A,X)}\Bigg\}\Bigg]\\
&\quad+E\Bigg[\frac{D}{\alpha}\big\{\tilde{F}_{11}(\tau\mid 0,X)-\theta_1(0)\big\}\{h(D,X)+h(X)\}\Bigg].
\end{aligned}
$$

In the following we show that the function $\varphi_1(0)(O)$ displayed in (S1) is indeed a gradient of the parameter $\theta_1(0)$ by verifying that

$$
E\{\varphi_1(0)(O)h(O)\}=\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\theta_1(0;\varepsilon)\Big|_{\varepsilon=0}
$$

for the score $h(O)$ of an arbitrary submodel $\{P_\varepsilon\}\subset\mathcal{P}$. The inner product

$$
\begin{aligned}
&E\{\varphi_1(0)(O)h(O)\}\\
&\quad=E\Bigg\{\frac{(1-A)\pi(X)}{\alpha}\int_0^\tau\frac{Dw_1(t\mid X)}{G_1(t-\mid A,X)}g_{11}(t\mid A,X)\mathrm{d}\tilde{M}_{11}(t\mid A,X)h_1(\tilde{T},\tilde{J},A,X)\Bigg\}\\
&\qquad+E\Bigg\{\frac{\pi(X)}{\alpha}\int_0^\tau\frac{(1-D)w_0(t\mid X)}{G_0(t-\mid X)}g_{11}(t\mid 0,X)\mathrm{d}\tilde{M}_{01}(t\mid X)h_0(\tilde{T},\tilde{J},X)\Bigg\}
\end{aligned}
$$

$$+ E\left\{\frac{(1-A)}{\alpha e_1(A \mid X)} \int_0^\tau \frac{D}{G_1(t- \mid A, X)} g_{21}(t \mid A, X) \mathrm{d}\tilde{M}_{12}(t \mid A, X) h_1(\tilde{T}, \tilde{J}, A, X)\right\}$$
$$+ E\left[\frac{D}{\alpha}\{\tilde{F}_{11}(\tau \mid 0, X) - \theta_1(0)\}\{h(D, X) + h(X)\}\right].$$

The first two terms of the right hand side of the equation can be simplified by (S4), so that they sum up to

$$\begin{aligned}
&E\left\{\frac{D(1-A)\pi(X)}{\alpha} \int_0^\tau w_1(t \mid X) g_{11}(t \mid A, X) \frac{\mathrm{d}\tilde{M}_{11}(t \mid A, X)}{G_1(t- \mid A, X)} h_1(\tilde{T}, \tilde{J}, A, X)\right\} \\
&\quad + E\left[\frac{(1-D)\pi(X)}{\alpha} E\left\{\int_0^\tau w_0(t \mid X) g_{11}(t \mid A, X) \frac{\mathrm{d}\tilde{M}_{11}(t \mid A, X)}{G_1(t- \mid A, X)}\right.\right. \\
&\qquad\qquad \left.\left. h_1(\tilde{T}, \tilde{J}, A, X) \,\middle|\, A = 0, X, D = 1\right\}\right] \\
&= E\left\{\frac{\pi(X)}{\alpha} E\left(\int_0^\tau [\pi(X) e_1(A \mid X) w_1(t \mid X) + \{1 - \pi(X)\} w_0(t \mid X)]\right.\right. \\
&\qquad\qquad \left.\left. g_{11}(t \mid A, X) \frac{\mathrm{d}\tilde{M}_{11}(t \mid A, X)}{G_1(t- \mid A, X)} h_1(\tilde{T}, \tilde{J}, A, X) \,\middle|\, A = 0, X, D = 1\right)\right\} \\
&= E\left\{\frac{D(1-A)}{\alpha e_1(A \mid X)} h_1(\tilde{T}, \tilde{J}, A, X) \int_0^\tau g_{11}(t \mid A, X) \frac{\mathrm{d}\tilde{M}_{11}(t \mid A, X)}{G_1(t- \mid A, X)}\right\},
\end{aligned}$$

where in the last step we used the identity

$$\pi(x) e_1(0 \mid x) w_1(t \mid x) + \{1 - \pi(x)\} w_0(t \mid x) = 1.$$

This can be established by direct calculation:

$$\begin{aligned}
&\pi(x) e_1(0 \mid x) w_1(t \mid x) + \{1 - \pi(x)\} w_0(t \mid x) \\
&\quad = \frac{1}{w_\bullet(t \mid x)}\left[\{1 - \triangle \tilde{\mathrm{A}}_{11}(t \mid 0, x)\} g_{11}(t \mid 0, x)\left\{\frac{1 - \pi(x)}{G_1(t- \mid 0, x)} + \frac{\pi(x) e_1(0 \mid x)}{G_0(t- \mid x)}\right\}\right] \\
&\quad = 1.
\end{aligned}$$

We have established that $\varphi_1(0)(O)$ is indeed a gradient of $\theta_1(0)$.

In order to show that $\varphi_1(0)$ is the efficient influence function, it remains to ascertain $\varphi_1(0)(O)$ itself is a score of the model at P. To proceed further, we decompose $\varphi_1(0)(O)$ into the following functions:

$$\begin{aligned}
h_1^*(\tilde{T}, \tilde{J}, A, X) &= \frac{\pi(X)(1-A)}{\alpha} \int_0^\tau \frac{w_1(t \mid X) g_{11}(t \mid A, X)}{G_1(t- \mid A, X)} \mathrm{d}\tilde{M}_{11}(t \mid A, X) \\
&\quad + \frac{(1-A)}{\alpha e_1(A \mid X)} \int_0^\tau \frac{g_{21}(t \mid A, X)}{G_1(t- \mid A, X)} \mathrm{d}\tilde{M}_{12}(t \mid A, X), \\
h_0^*(\tilde{T}, \tilde{J}, X) &= \frac{\pi(X)}{\alpha} \int_0^\tau \frac{w_0(t \mid X) g_{11}(t \mid 0, X)}{G_0(t- \mid X)} \mathrm{d}\tilde{M}_{01}(t \mid X), \\
h^*(D, X) &= \frac{D - \pi(X)}{\alpha}\{\tilde{F}_{11}(\tau \mid 0, X) - \theta_1(0)\}, \\
h^*(X) &= \frac{\pi(X)}{\alpha}\{\tilde{F}_{11}(\tau \mid 0, X) - \theta_1(0)\},
\end{aligned}$$

such that

$$\varphi_1(0)(O) = D h_1^*(\tilde{T}, \tilde{J}, A, X) + (1 - D) h_0^*(\tilde{T}, \tilde{J}, X) + h^*(D, X) + h^*(X).$$

It is trivial to show that $h^*(D,X)$ and $h^*(X)$ are valid scores by noting that $E\{h^*(D,X)\mid X\}=0$ and $E\{h^*(X)\}=0$, so $h^*(D,X)\in\dot{\mathcal{P}}_3$ and $h^*(X)\in\dot{\mathcal{P}}_4$. Since $h_1^*(\tilde{T},\tilde{J},A,X)$ and $h_0^*(\tilde{T},\tilde{J},X)$ are zero-mean martingales adapted to the filtration $(\mathcal{F}_t)$ at time τ, it is also clear that

$$E\{h_1^*(\tilde{T},\tilde{J},A,X)\mid A,X,D=1\}=0,\quad E\{h_0^*(\tilde{T},\tilde{J},X)\mid X,D=0\}=0.$$

We will now verify that h_0^* and h_1^* fulfill (S4) in the integral form. We do so by computing both sides of (S4) substituting h_1^* for h_1 and h_0^* for h_0. For any $0<t\leqslant\tau$, we calculate the conditional expectation

$$\begin{aligned}
&E\left\{h_1^*(\tilde{T},\tilde{J},A,X)\int_0^t\frac{\mathrm{d}\tilde{M}_{11}(s\mid A,X)}{G_1(s-\mid A,X)}\,\middle|\,A=0,X=x,D=1\right\}\\
&\quad=\frac{1}{\alpha}E\Bigg[\int_0^t\frac{\mathrm{d}\tilde{M}_{11}(s\mid A,X)}{G_1(s-\mid A,X)}\int_0^t\left\{\pi(X)\frac{w_1(s\mid X)g_{11}(s\mid A,X)}{G_1(s-\mid A,X)}\mathrm{d}\tilde{M}_{11}(s\mid A,X)\right.\\
&\qquad\left.+\frac{1}{e_1(A\mid X)}\frac{g_{21}(s\mid A,X)}{G_1(s-\mid A,X)}\mathrm{d}\tilde{M}_{12}(s\mid A,X)\right\}\,\Bigg|\,A=0,X=x,D=1\Bigg],
\end{aligned}$$

and by the property of the martingale product and the corresponding predictable covariation process (Fleming and Harrington, 1991, Theorem 2.3.4), the term above is

$$\begin{aligned}
&=\frac{1}{\alpha}E\Bigg[\left\langle\int_0^t\frac{\mathrm{d}\tilde{M}_{11}(s\mid A,X)}{G_1(s-\mid A,X)},\int_0^t\left\{\pi(X)\frac{w_1(s\mid X)g_{11}(s\mid A,X)}{G_1(s-\mid A,X)}\mathrm{d}\tilde{M}_{11}(s\mid A,X)\right.\right.\\
&\qquad\left.\left.\frac{1}{e_1(A\mid X)}\frac{g_{21}(s\mid A,X)}{G_1(s-\mid A,X)}\mathrm{d}\tilde{M}_{12}(s\mid A,X)\right\}\right\rangle\,\Bigg|\,A=0,X=x,D=1\Bigg]\\
&=\frac{1}{\alpha}E\left\{\int_0^t\frac{w_1(s\mid X)g_{11}(s\mid A,X)}{\{G_1(s-\mid A,X)\}^2}\pi(X)\mathrm{d}\langle\tilde{M}_{11}\rangle(s\mid A,X)\,\middle|\,A=0,X=x,D=1\right\}\\
&\quad+\frac{1}{\alpha}E\left\{\int_0^t\frac{g_{21}(s\mid A,X)}{\{G_1(s-\mid A,X)\}^2}\frac{\mathrm{d}\langle\tilde{M}_{11},\tilde{M}_{12}\rangle(s\mid A,X)}{e_1(A\mid X)}\,\middle|\,A=0,X=x,D=1\right\},
\end{aligned}$$

and evaluating the predictable (co-)variation processes with the help of Theorem 2.6.1 of Fleming and Harrington (1991), finally gives

$$\begin{aligned}
&=\frac{1}{\alpha}\int_0^t\frac{\mathrm{d}\tilde{\mathrm{A}}_{11}(s\mid 0,x)}{G_1(s-\mid 0,x)}\\
&\qquad\left\{\pi(x)w_1(s\mid x)g_{11}(s\mid 0,x)\{1-\triangle\tilde{\mathrm{A}}_{11}(s\mid 0,x)\}-\frac{g_{21}(s\mid 0,x)}{e_1(0\mid x)}\triangle\tilde{\mathrm{A}}_{12}(s\mid 0,x)\right\}.
\end{aligned}\tag{S8}$$

On the other hand,

$$\begin{aligned}
&E\left\{h_0^*(\tilde{T},\tilde{J},X)\int_0^t\frac{\mathrm{d}\tilde{M}_{01}(s\mid X)}{G_0(s-\mid X)}\,\middle|\,X=x,D=0\right\}\\
&\quad=\frac{1}{\alpha}E\left\{\int_0^t\frac{\mathrm{d}\tilde{M}_{01}(s\mid X)}{G_0(s-\mid X)}\int_0^t\pi(X)\frac{w_0(s\mid X)g_{11}(s\mid 0,X)}{G_0(s-\mid X)}\mathrm{d}\tilde{M}_{01}(s\mid X)\,\middle|\,X=x,D=0\right\}\\
&\quad=\frac{1}{\alpha}E\left\{\int_0^t\pi(X)\frac{w_0(s\mid X)g_{11}(s\mid 0,X)}{\{G_0(s-\mid X)\}^2}\mathrm{d}\langle\tilde{M}_{01}\rangle(s\mid X)\,\middle|\,X=x,D=0\right\}\\
&\quad=\frac{1}{\alpha}\int_0^t\frac{\mathrm{d}\tilde{\mathrm{A}}_{11}(s\mid 0,x)}{G_0(s-\mid X)}\pi(x)w_0(s\mid x)g_{11}(s\mid 0,x)\{1-\triangle\tilde{\mathrm{A}}_{11}(s\mid 0,x)\}.
\end{aligned}\tag{S9}$$

The restriction (S4) holds if (S8) and (S9) are equal. Therefore, we only need to show that for $t\in(0,\tau]$,

$$\pi(x)\{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,x)\}g_{11}(t\,|\,0,x)\left\{\frac{w_1(t\,|\,x)}{G_1(t-|\,0,x)}-\frac{w_0(t\,|\,x)}{G_0(t-|\,x)}\right\}$$
$$=\frac{\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,x)}{G_1(t-|\,0,x)e_1(0\,|\,x)}g_{21}(t\,|\,0,x). \quad \text{(S10)}$$

The difference in the braces is

$$\begin{aligned}
&\frac{w_1(t\,|\,x)}{G_1(t-|\,0,x)}-\frac{w_0(t\,|\,x)}{G_0(t-|\,x)}\\
&\quad=\frac{1}{w_\bullet(t\,|\,x)}\frac{\triangle\tilde{\mathrm{A}}_{12}(t\,|\,a,x)}{G_1(t-|\,0,x)}g_{21}(t\,|\,0,x)\left\{\frac{1-\pi(x)}{\pi(x)e_1(0\,|\,x)G_1(t-|\,0,x)}+\frac{1}{G_0(t-|\,x)}\right\}\\
&\quad=\frac{1}{\pi(x)}\frac{1}{w_\bullet(t\,|\,x)}\left\{\frac{1-\pi(x)}{G_1(t-|\,0,x)}+\frac{\pi(x)e_1(0\,|\,x)}{G_0(t-|\,x)}\right\}\\
&\qquad\frac{\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,x)}{G_1(t-|\,0,x)e_1(0\,|\,x)}g_{21}(t\,|\,0,x),
\end{aligned}$$

which by inserting the definition of $w_\bullet(t\,|\,x)$ is simply

$$=\frac{1}{\pi(x)\{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,x)\}g_{11}(t\,|\,0,x)}\frac{\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,x)}{G_1(t-|\,0,x)e_1(0\,|\,x)}g_{21}(t\,|\,0,x),$$

and therefore (S10) holds.

That the terms (S8) and (S9) are equal shows that $Dh_1^*(\tilde{T},\tilde{J},A,X)+(1-D)h_0^*(\tilde{T},\tilde{J},X)\in\dot{\mathcal{P}}_1$. Therefore, $\varphi_1(0)(O)$ belongs to the tangent space $\dot{\mathcal{P}}$ of the model $\mathcal{P}$ at P. Hence, it is the efficient influence function of $\theta_1(0)$.

The proof of $\varphi_2(0)(O)$ being the efficient influence function of $\theta_2(0)$ at $P\in\mathcal{P}$ can be obtained by slightly modifying the derivations above and is thus omitted. □

Remark S1. Inspecting the proof of Proposition S1, we find two more influence functions of the parameter $\theta_1(0)$ in the model $\mathcal{P}$. The first influence function is obtained by replacing both $w_1(t\,|\,x)$ and $w_0(t\,|\,x)$ with $[\{1-\pi(x)\}+\pi(x)e_1(0\,|\,x)]^{-1}$ in $\varphi_1(0)$. The second one is obtained by replacing $w_1(t\,|\,x)$ with $\{\pi(x)e_1(0\,|\,x)\}^{-1}$ and $w_0(t\,|\,x)$ with 0 in $\varphi_1(0)$. In this case, the resulting influence function is identical to the one proposed in Rytgaard et al. (2023) but restricted to the RCT population.

Lemma 1 is a simplification of the expressions of the efficient influence functions in Proposition S1 under Assumptions 3 and 5.

Proof of Lemma 1. Under Assumptions 3 and 5, we rewrite the following quantities from the main text using only the observed data distribution:

$$\begin{aligned}
H_\bullet(t\,|\,x)&=\pi(x)e_1(0\,|\,x)G_1(t\,|\,0,x)+\{1-\pi(x)\}G_0(t\,|\,x),\\
H_1(t\,|\,a,x)&=e_1(a\,|\,x)G_1(t\,|\,a,x),\\
W_{k1}(t\,|\,a,x)&=I(k=1)\tilde{S}_1(t\,|\,a,x)-\frac{\tilde{F}_{11}(\tau\,|\,a,x)-\tilde{F}_{11}(t\,|\,a,x)}{1-\triangle\tilde{\mathrm{A}}_{k1}(t\,|\,0,X)}.
\end{aligned}$$

Let

$$\begin{aligned}
\varphi_1^*(0)(O)&=\frac{(1-A)\pi(X)}{\alpha}\int_0^\tau\frac{g_{11}(t\,|\,A,X)}{H_\bullet(t-|\,X)}\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)\\
&\quad+\frac{D(1-A)}{\alpha}\int_0^\tau\frac{g_{21}(t\,|\,A,X)}{H_1(t-|\,A,X)}\mathrm{d}\tilde{M}_{12}(t\,|\,A,X)+\frac{D}{\alpha}\left\{\tilde{F}_{11}(\tau\,|\,0,X)-\theta_1(0)\right\}\\
\varphi_1^\dagger(0)(O)&=\frac{(1-A)\pi(X)}{\alpha}\int_0^\tau\frac{W_{11}(t\,|\,A,X)}{H_\bullet(t-|\,X)}\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)
\end{aligned}$$

$$+\frac{D(1-A)}{\alpha}\int_0^\tau \frac{W_{21}(t\,|\,A,X)}{H_1(t-\,|\,A,X)}\mathrm{d}\tilde{M}_{12}(t\,|\,A,X)+\frac{D}{\alpha}\big\{\tilde{F}_{11}(\tau\,|\,0,X)-\theta_1(0)\big\}.$$

The function $\varphi_1^\dagger(0)(O)$ is the expression appearing on the right-hand side of the statement of the efficient influence function in the lemma. To show the lemma, we need to check for $\varphi_1(0)(O)$ from Proposition S1 that

$$E\big\{\varphi_1(0)(O)-\varphi_1^*(0)(O)\big\}^2=0,\quad E\big\{\varphi_1^*(0)(O)-\varphi_1^\dagger(0)(O)\big\}^2=0,$$

so that $\varphi_1(0)(O)=\varphi_1^\dagger(0)(O)$ P-almost surely. Noting that $(1-D)\mathrm{d}\tilde{M}_{11}(t\,|\,0,X)=(1-D)\mathrm{d}\tilde{M}_{01}(t\,|\,X)$ and

$$\frac{g_{11}(t\,|\,0,X)}{w_\bullet(t\,|\,X)}\frac{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,X)}{G_1(t-\,|\,0,X)G_0(t-\,|\,X)}=\frac{1}{H_\bullet(t-\,|\,X)},$$

the $L_2(P)$-norm of the difference $E\big\{\varphi_1(0)(O)-\varphi_1^*(0)(O)\big\}^2$ is the expectation of the sum of two squared martingales

$$\begin{aligned}
&E\big\{\varphi_1(0)(O)-\varphi_1^*(0)(O)\big\}^2\\
&\quad=E\Bigg[\frac{D(1-A)}{\alpha}I\{\pi(X)>0\}\int_0^\tau\frac{1}{w_\bullet(t\,|\,X)}\frac{\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,X)}{G_1^2(t-\,|\,0,X)}\frac{1-\pi(X)}{e_1(0\,|\,X)}\\
&\qquad\qquad(g_{11}g_{21})(t\,|\,0,X)\mathrm{d}\tilde{M}_{11}(t\,|\,0,X)\Bigg]^2\\
&\qquad+E\Bigg\{\frac{1-D}{\alpha}\pi(X)\int_0^\tau\frac{1}{w_\bullet(t\,|\,X)}\frac{\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,X)}{G_1(t-\,|\,0,X)G_0(t-\,|\,X)}\\
&\qquad\qquad(g_{11}g_{21})(t\,|\,0,X)\mathrm{d}\tilde{M}_{01}(t\,|\,X)\Bigg\}^2\\
&\quad=E\Bigg[\frac{D(1-A)}{\alpha^2}I\{\pi(X)>0\}\int_0^\tau\frac{1}{w_\bullet^2(t\,|\,X)}\frac{\triangle\tilde{\mathrm{A}}_{12}^2(t\,|\,0,X)}{G_1^4(t-\,|\,0,X)}\frac{\{1-\pi(X)\}^2}{e_1^2(0\,|\,X)}\\
&\qquad\qquad(g_{11}g_{21})^2(t\,|\,0,X)\mathrm{d}\langle\tilde{M}_{11}\rangle(t\,|\,0,X)\Bigg]\\
&\qquad+E\Bigg\{\frac{1-D}{\alpha^2}\pi^2(X)\int_0^\tau\frac{1}{w_\bullet^2(t\,|\,X)}\frac{\triangle\tilde{\mathrm{A}}_{12}^2(t\,|\,0,X)}{G_1^2(t-\,|\,0,X)G_0^2(t-\,|\,X)}\\
&\qquad\qquad(g_{11}g_{21})^2(t\,|\,0,X)\mathrm{d}\langle\tilde{M}_{01}\rangle(t\,|\,X)\Bigg\}\\
&\quad=E\Bigg[\frac{D(1-A)}{\alpha^2}I\{\pi(X)>0\}\int_0^\tau\frac{1}{w_\bullet^2(t\,|\,X)}\frac{\triangle\tilde{\mathrm{A}}_{12}^2(t\,|\,0,X)}{G_1^4(t-\,|\,0,X)}\frac{\{1-\pi(X)\}^2}{e_1^2(0\,|\,X)}\\
&\qquad\qquad(g_{11}g_{21})^2(t\,|\,0,X)I(\tilde{T}\geqslant t)\{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,X)\}\mathrm{d}\tilde{\mathrm{A}}_{11}(t\,|\,0,X)\Bigg]\\
&\qquad+E\Bigg\{\frac{1-D}{\alpha^2}\pi^2(X)\int_0^\tau\frac{1}{w_\bullet^2(t\,|\,X)}\frac{\triangle\tilde{\mathrm{A}}_{12}^2(t\,|\,0,X)}{G_1^2(t-\,|\,0,X)G_0^2(t-\,|\,X)}\\
&\qquad\qquad(g_{11}g_{21})^2(t\,|\,0,X)I(\tilde{T}\geqslant t)\{1-\triangle\tilde{\mathrm{A}}_{01}(t\,|\,X)\}\mathrm{d}\tilde{\mathrm{A}}_{01}(t\,|\,X)\Bigg\}\\
&\quad=0.
\end{aligned}$$

The last equality is a direct consequence of Assumption 5, that is, $\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,x)\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,x)=0$ and $\triangle\tilde{\mathrm{A}}_{01}(t\,|\,x)\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,x)=0$ for any $x\in\mathcal{X}_0\cap\mathcal{X}_1$. Similarly, we have

$$
\begin{aligned}
&E\big\{\varphi_1^*(0)(O)-\varphi_1^\dagger(0)(O)\big\}^2\\
&\quad=E\left\{\frac{(1-A)\pi(X)}{\alpha}\int_0^\tau\frac{(g_{11}-W_{11})(t\,|\,A,X)}{H_\bullet(t-\,|\,X)}\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)\right\}^2\\
&\qquad+E\left\{\frac{D(1-A)}{\alpha}\int_0^\tau\frac{(g_{21}-W_{21})(t\,|\,A,X)}{H_1(t-\,|\,A,X)}\mathrm{d}\tilde{M}_{12}(t\,|\,A,X)\right\}^2\\
&\quad=E\left\{\frac{(1-A)\pi^2(X)}{\alpha^2}\int_0^\tau\frac{(g_{11}-W_{11})^2(t\,|\,A,X)}{H_\bullet^2(t-\,|\,X)}\mathrm{d}\langle\tilde{M}_{11}\rangle(t\,|\,A,X)\right\}\\
&\qquad+E\left\{\frac{D(1-A)}{\alpha^2}\int_0^\tau\frac{(g_{21}-W_{21})^2(t\,|\,A,X)}{H_1^2(t-\,|\,A,X)}\mathrm{d}\langle\tilde{M}_{12}\rangle(t\,|\,A,X)\right\}\\
&\quad=E\Bigg[\frac{(1-A)\pi^2(X)}{\alpha^2}\int_0^\tau\frac{\triangle\tilde{\mathrm{A}}_{12}^2(t\,|\,0,X)}{\big\{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,X)\big\}^2\big\{1-\triangle(\tilde{\mathrm{A}}_{11}+\tilde{\mathrm{A}}_{12})(t\,|\,0,X)\big\}^2}\\
&\qquad\frac{\big\{\tilde{F}_{11}(\tau\,|\,0,X)-\tilde{F}_{11}(t\,|\,0,X)\big\}^2}{H_\bullet^2(t-\,|\,X)}I(\tilde{T}\geqslant t)\big\{1-\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,X)\big\}\mathrm{d}\tilde{\mathrm{A}}_{11}(t\,|\,0,X)\Bigg]\\
&\qquad+E\Bigg[\frac{D(1-A)}{\alpha^2}\int_0^\tau\frac{\triangle\tilde{\mathrm{A}}_{11}^2(t\,|\,0,X)}{\big\{1-\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,X)\big\}^2\big\{1-\triangle(\tilde{\mathrm{A}}_{11}+\tilde{\mathrm{A}}_{12})(t\,|\,0,X)\big\}^2}\\
&\qquad\frac{\big\{\tilde{F}_{11}(\tau\,|\,0,X)-\tilde{F}_{11}(t\,|\,0,X)\big\}^2}{H_1^2(t-\,|\,0,X)}I(\tilde{T}\geqslant t)\big\{1-\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,X)\big\}\mathrm{d}\tilde{\mathrm{A}}_{12}(t\,|\,0,X)\Bigg]\\
&\quad=0.
\end{aligned}
$$

The last equality follows from Assumption 5, $\triangle\tilde{\mathrm{A}}_{11}(t\,|\,0,x)\triangle\tilde{\mathrm{A}}_{12}(t\,|\,0,x)=0$ for any $x\in\mathcal{X}_1$.

The equivalence for the efficient influence function $\varphi_2(0)(O)$ can be argued analogously. □

S2.2. *Proof of Corollary 1*

Let $\tilde{\mathcal{P}}$ be the same as the model $\mathcal{P}$ but the restriction $\mathrm{d}\tilde{\mathrm{A}}_{11}(t\,|\,0,x)=\mathrm{d}\tilde{\mathrm{A}}_{01}(t\,|\,x)$ is removed. The semiparametric efficiency bound of $\theta_1(0)$ under $P\in\tilde{\mathcal{P}}$ can be characterized by the variance of the efficient influence function

$$
\begin{aligned}
\tilde{\varphi}_1(0)(O)=&\frac{D}{\alpha}\frac{1-A}{e_1(A\,|\,X)}\int_0^\tau\frac{g_{11}(t\,|\,A,X)}{G_1(t-\,|\,A,X)}\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)\\
&+\frac{D}{\alpha}\frac{1-A}{e_1(A\,|\,X)}\int_0^\tau\frac{g_{21}(t\,|\,A,X)}{G_1(t-\,|\,A,X)}\mathrm{d}\tilde{M}_{12}(t\,|\,A,X)\\
&+\frac{D}{\alpha}\big\{\tilde{F}_{11}(\tau\,|\,0,X)-\theta_1(0)\big\}.
\end{aligned}
$$

See Remark S1 for the justification of this claim. Then the variance of the difference in the efficient influence functions under these two models with respect to $P\in\mathcal{P}$ is

$$
\begin{aligned}
&E\{\varphi_1(0)(O)-\tilde{\varphi}_1(0)(O)\}^2\\
&\quad=E\Bigg\{\frac{\pi(X)}{\alpha}(1-A)\int_0^\tau\frac{W_{11}(t\,|\,A,X)}{H_\bullet(t-\,|\,X)}\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)\\
&\qquad\qquad-\frac{D}{\alpha}(1-A)\int_0^\tau\frac{W_{11}(t\,|\,A,X)}{H_1(t-\,|\,A,X)}\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)\Bigg\}^2
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\frac{1-A}{\alpha}\int_0^\tau\left\{\frac{\pi(X)}{H_\bullet(t-\,|\,X)}-\frac{D}{H_1(t-\,|\,A,X)}\right\}W_{11}(t\,|\,A,X)\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)\right]^2\\
&= E\left\langle\frac{1-A}{\alpha}\int_0^\tau\left\{\frac{\pi(X)}{H_\bullet(t-\,|\,X)}-\frac{D}{H_1(t-\,|\,A,X)}\right\}W_{11}(t\,|\,A,X)\mathrm{d}\tilde{M}_{11}(t\,|\,A,X)\right\rangle\\
&= E\left[\frac{1-A}{\alpha^2}\int_0^\tau\left\{\frac{\pi(X)}{H_\bullet(t-\,|\,X)}-\frac{D}{H_1(t-\,|\,A,X)}\right\}^2\{W_{11}(t\,|\,A,X)\}^2\mathrm{d}\langle\tilde{M}_{11}\rangle(t\,|\,A,X)\right]\\
&= E\left[\frac{1-A}{\alpha^2}\int_0^\tau\left\{\frac{\pi(X)}{H_\bullet(t-\,|\,X)}-\frac{D}{H_1(t-\,|\,A,X)}\right\}^2\right.\\
&\qquad\left.\{W_{11}(t\,|\,A,X)\}^2I(\tilde{T}\geqslant t)\{1-\triangle\tilde{\mathrm{A}}_{\bullet1}(t\,|\,A,X)\}\mathrm{d}\tilde{\mathrm{A}}_{\bullet1}(t\,|\,A,X)\right]\\
&= E\left[\frac{1}{\alpha^2}\int_0^\tau\left\{\frac{\{\pi(X)\}^2}{H_\bullet(t-\,|\,X)}+\frac{\pi(X)}{H_1(t-\,|\,0,X)}-\frac{2\{\pi(X)\}^2}{H_\bullet(t-\,|\,X)}\right\}\right.\\
&\qquad\left.\{W_{11}(t\,|\,0,X)\}^2\{1-\triangle\tilde{\mathrm{A}}_{\bullet1}(t\,|\,0,X)\}\mathrm{d}\tilde{\mathrm{A}}_{\bullet1}(t\,|\,0,X)\right]\\
&= E\left[\frac{\pi(X)}{\alpha^2}\int_0^\tau\left\{\frac{1}{H_1(t-\,|\,0,X)}-\frac{\pi(X)}{H_\bullet(t-\,|\,X)}\right\}\right.\\
&\qquad\left.\{W_{11}(t\,|\,0,X)\}^2\{1-\triangle\tilde{\mathrm{A}}_{\bullet1}(t\,|\,0,X)\}\mathrm{d}\tilde{\mathrm{A}}_{\bullet1}(t\,|\,0,X)\right]\\
&= E\left[\frac{\pi(X)\{1-\pi(X)\}}{\alpha^2}\int_0^\tau\frac{(S_0S_0^c)(t-\,|\,X)}{H_1(t-\,|\,0,X)H_\bullet(t-\,|\,X)}\right.\\
&\qquad\left.\{W_{11}(t\,|\,0,X)\}^2\{1-\triangle\tilde{\mathrm{A}}_{\bullet1}(t\,|\,0,X)\}\mathrm{d}\tilde{\mathrm{A}}_{\bullet1}(t\,|\,0,X)\right].
\end{aligned}
$$

The expression in the statement of the corollary follows from Assumptions 3 and 5.

S2.3. *Proof of Theorem 1*

We first state the deferred assumptions in the statement of Theorem 1.

Assumption S1 (Regularity conditions).

(i) There exists a universal constant $C > 1$ such that

$$
\begin{aligned}
&\hat{\alpha}\geqslant C^{-1},\hat{e}_1(0\,|\,x)\geqslant C^{-1},\bar{e}_1(0\,|\,x)\geqslant C^{-1},\\
&(\Pi\hat{\mathrm{A}}_{\bullet1})(\tau\,|\,0,x)\geqslant C^{-1},(\Pi\bar{\mathrm{A}}_{\bullet1})(\tau\,|\,0,x)\geqslant C^{-1},\\
&(\Pi\hat{\mathrm{A}}_{12})(\tau\,|\,0,x)\geqslant C^{-1},(\Pi\bar{\mathrm{A}}_{12})(\tau\,|\,0,x)\geqslant C^{-1},
\end{aligned}
$$

wherever $\pi(x) > 0$, and

$$
\begin{aligned}
&(\Pi\hat{\mathrm{A}}_{02})(\tau\,|\,x)\geqslant C^{-1},(\Pi\bar{\mathrm{A}}_{02})(\tau\,|\,x)\geqslant C^{-1},\\
&(\Pi\hat{\mathrm{A}}_0^c)(\tau\,|\,x)\geqslant C^{-1},(\Pi\bar{\mathrm{A}}_0^c)(\tau\,|\,x)\geqslant C^{-1},
\end{aligned}
$$

wherever $\pi(x)\{1-\pi(x)\} > 0$;

(ii) $\{x:\bar{\pi}(x) > 0\}\subset\mathcal{X}_1$;

(iii) For $x\in\mathcal{X}_1$,

$$
\begin{aligned}
&\{\hat{\mathrm{A}}_{\bullet1}(t\,|\,0,x),\mathrm{A}_{\bullet1}(t\,|\,0,x)\}\perp_\triangle\{\hat{\mathrm{A}}_{12}(t\,|\,0,x),\mathrm{A}_{12}(t\,|\,0,x),\hat{\mathrm{A}}_{02}(t\,|\,x),\mathrm{A}_{02}(t\,|\,x)\},\\
&\{\hat{\mathrm{A}}_{\bullet1}(t\,|\,0,x),\bar{\mathrm{A}}_{\bullet1}(t\,|\,0,x)\}\perp_\triangle\{\hat{\mathrm{A}}_{12}(t\,|\,0,x),\bar{\mathrm{A}}_{12}(t\,|\,0,x),\bar{\mathrm{A}}_{02}(t\,|\,x),\\
&\qquad\qquad\bar{\mathrm{A}}_1^c(t\,|\,0,x),\bar{\mathrm{A}}_0^c(t\,|\,x)\},\\
&\{\hat{\mathrm{A}}_{12}(t\,|\,0,x),\bar{\mathrm{A}}_{12}(t\,|\,0,x)\}\perp_\triangle\{\bar{\mathrm{A}}_{\bullet1}(t\,|\,0,x),\bar{\mathrm{A}}_1^c(t\,|\,0,x)\};
\end{aligned}
$$

(iv) $\hat{\ell}_1(0)$ and $\ell_1(0)$ belong to some P-Donsker class.

Assumption S2 (Rate conditions). The following integrals converge sufficiently fast:

$$P\Bigg[\int_0^\tau \left\{\hat{\pi}(X)\frac{H_\bullet^*}{\hat{H}_\bullet^*}(t- \mid X) - \pi(X)\frac{\Pi A_{12}}{\Pi \hat{A}_{12}}(t- \mid 0, X)\right\} \hat{W}_{\bullet 1}(t \mid 0, X)\{1 - \triangle \hat{A}_{\bullet 1}(t \mid 0, X)\}\mathrm{d}\left(\frac{\Pi A_{\bullet 1}}{\Pi \hat{A}_{\bullet 1}}\right)(t \mid 0, X)\Bigg] = o_P(n^{-1/2}), \quad \text{(AS2.1)}$$

$$P\Bigg[\pi(X)\int_0^\tau \left\{\frac{e_1(0 \mid X)S_1^c(t- \mid 0, X)}{\hat{e}_1(0 \mid X)\hat{S}_1^c(t- \mid 0, X)} - 1\right\}\frac{\Pi A_{\bullet 1}}{\Pi \hat{A}_{\bullet 1}}(t- \mid 0, X) \hat{W}_{12}(t \mid 0, X)\{1 - \triangle \hat{A}_{12}(t \mid 0, X)\}\mathrm{d}\left(\frac{\Pi A_{12}}{\Pi \hat{A}_{12}}\right)(t \mid 0, X)\Bigg] = o_P(n^{-1/2}), \quad \text{(AS2.2)}$$

where

$$H_\bullet^*(t \mid X) = \pi(X)e_1(0 \mid X)\{(\Pi A_{12})S_1^c\}(t \mid 0, X) + \{1 - \pi(X)\}\{(\Pi A_{02})S_0^c\}(t \mid X).$$

Remark S2. Since estimators for cumulative hazards often contain jumps, the convergence is stated in terms of the means of stochastic integrals (Westling et al., 2024). When the event time distribution is absolutely continuous with respect to the Lebesgue measure and the conditional cumulative hazards are estimated by continuous functions, the remainder terms will admit a more conventional product structure. This is because the Cauchy-Schwarz inequality can be applied with respect to the product measure of P marginalized to the support $\mathcal{X}_1 \cup \mathcal{X}_0$ of X and the Lebesgue measure over the time interval $(0, \tau]$; refer to Rytgaard et al. (2023) for precise formulations.

Remark S3. To establish asymptotic linearity of the estimator $\hat{\gamma}_1(0)$, the same convergence rates of the remainder terms (AS2.1)–(AS2.2) should hold when swapping $\hat{W}_{\bullet 1}(t \mid 0, X)$ and $\hat{W}_{12}(t \mid 0, X)$ out for

$$\int_t^\tau \left\{\hat{S}_1(t- \mid 0, X) - \frac{\hat{F}_{11}(s \mid 0, X) - \hat{F}_{11}(t \mid 0, X)}{1 - \triangle \hat{A}_{\bullet 1}(t \mid 0, X)}\right\}\mathrm{d}s,$$
$$\int_t^\tau \frac{\hat{F}_{11}(s \mid 0, X) - \hat{F}_{11}(t \mid 0, X)}{1 - \triangle \hat{A}_{12}(t \mid 0, X)}\mathrm{d}s.$$

We will use the following lemmas from the literature.

Lemma S1 (Integration by parts, Fleming and Harrington, 1991, Theorem A.1.2). *Let $F : [0, \infty) \to \mathbb{R}$ and $G : [0, \infty) \to \mathbb{R}$ be càdlàg functions of bounded variation on any finite interval. Then*

$$F(t)G(t) - F(s)G(s) = \int_{(s,t]} F(u-)\mathrm{d}G(u) + \int_{(s,t]} G(u)\mathrm{d}F(u).$$

Lemma S2 (Duhamel and backward equations, Gill and Johansen, 1990). *Let $F : [0, \infty) \to \mathbb{R}$ and $G : [0, \infty) \to \mathbb{R}$ be càdlàg functions of bounded variation on any finite interval. Then*

$$\prod_{u \in (s,t]} \{1 + \mathrm{d}F(u)\} - \prod_{u \in (s,t]} \{1 + \mathrm{d}G(u)\} = \int_{u \in (s,t]} \prod_{v \in (s,u)} \{1 + \mathrm{d}F(v)\}\mathrm{d}(F - G)(u) \prod_{v \in (u,t]} \{1 + \mathrm{d}G(v)\}, \quad \text{(Duhamel)}$$

$$\prod_{u\in(s,t]}\{1+\mathrm{d}F(u)\}-1=\int_{u\in(s,t]}\prod_{v\in(u,t]}\{1+\mathrm{d}F(v)\}\mathrm{d}F(u). \qquad \text{(backward)}$$

To make the notations more compact, we use the symbol S to represent the product integral $\Pi\mathrm{A}$ for $\mathrm{A}\in\mathcal{A}$, and superscripts and subscripts in A are carried over to S. For example, $S_{\bullet 1}(t\,|\,0,x)=(\Pi\mathrm{A}_{\bullet 1})(t\,|\,0,x)$. We use the nuisance parameters with checkmarks as a placeholder for either the probability limits in Assumption 6 (with bars) or the estimated nuisance parameters (with hats). Define

$$\begin{aligned}
\check{q}_1(t\,|\,X)&=\check{S}_1(t\,|\,0,X)-\check{F}_{11}(\tau\,|\,0,X)+\check{F}_{11}(t\,|\,0,X),\\
\check{q}_2(t\,|\,X)&=-\check{F}_{11}(\tau\,|\,0,X)+\check{F}_{11}(t\,|\,0,X),\\
\check{b}_1(t\,|\,X)&=\check{\pi}(x)\check{e}_1(0\,|\,X)(\check{S}_{12}\check{S}_1^c)(t-\,|\,0,X)+\{1-\check{\pi}(X)\}(\check{S}_{02}\check{S}_0^c)(t-\,|\,X),\\
b_1(t\,|\,X)&=\pi(x)e_1(0\,|\,X)(S_{12}S_1^c)(t-\,|\,0,X)+\{1-\pi(X)\}(S_{02}S_0^c)(t-\,|\,X),\\
\check{b}_2(t\,|\,X)&=\check{e}_1(0\,|\,X)(\check{S}_{\bullet 1}\check{S}_1^c)(t-\,|\,0,X),\\
b_2(t\,|\,X)&=e_1(0\,|\,X)(S_{\bullet 1}S_1^c)(t-\,|\,0,X).
\end{aligned}$$

The function obtained by substituting all nuisance parameters in $\ell_1(0)$ by their version with the checkmark can be written in terms of the quantities above as

$$\check{\ell}_1(0)(O)=\sum_{m=1}^{3}\check{\ell}_{1m}(0)(O),$$

where

$$\begin{aligned}
\check{\ell}_{11}(0)(O)&=\frac{1-A}{\check{\alpha}}\int_0^\tau\check{\pi}(X)\frac{\check{q}_1}{\check{b}_1}(t\,|\,X)\frac{\mathrm{d}\check{M}_{\bullet 1}}{\check{S}_{\bullet 1}}(t\,|\,0,X),\\
\check{\ell}_{12}(0)(O)&=\frac{D(1-A)}{\check{\alpha}}\int_0^\tau\frac{\check{q}_2}{\check{b}_2}(t\,|\,X)\frac{\mathrm{d}\check{M}_{12}}{\check{S}_{12}}(t\,|\,0,X),\\
\check{\ell}_{13}(0)(O)&=\frac{D}{\check{\alpha}}\check{F}_{11}(t\,|\,0,X).
\end{aligned}$$

The following lemma will be used in two versions by substituting the nuisance parameters with checkmark with their estimates and the probability limits of their estimators, respectively.

Lemma S3. *Suppose Assumptions 6 and S1 hold. Then*

$$\begin{aligned}
&P\left\{\check{\ell}_1(0)-\frac{\alpha}{\check{\alpha}}\ell_1(0)\right\}\\
&\quad=P\left[\frac{1}{\check{\alpha}}\int_0^\tau\left\{\check{\pi}(X)\frac{b_1}{\check{b}_1}(t\,|\,X)-\pi(X)\frac{S_{12}}{\check{S}_{12}}(t-\,|\,0,X)\right\}\check{q}_1(t\,|\,X)\mathrm{d}\left(1-\frac{S_{\bullet 1}}{\check{S}_{\bullet 1}}\right)(t\,|\,0,X)\right]\\
&\qquad+P\left[\frac{\pi(X)}{\check{\alpha}}\int_0^\tau\left\{\frac{b_2}{\check{b}_2}(t\,|\,X)-\frac{S_{\bullet 1}}{\check{S}_{\bullet 1}}(t-\,|\,0,X)\right\}\check{q}_2(t\,|\,X)\mathrm{d}\left(1-\frac{S_{12}}{\check{S}_{12}}\right)(t\,|\,0,X)\right].
\end{aligned}$$

Proof. We have

$$\begin{aligned}
P\check{\ell}_{11}(0)&=P\left\{\frac{1-A}{\check{\alpha}}\int_0^\tau\check{\pi}(X)\frac{\check{q}_1}{\check{b}_1}(t\,|\,X)\frac{\mathrm{d}\check{M}_{\bullet 1}}{\check{S}_{\bullet 1}}(t\,|\,0,X)\right\}\\
&=P\left\{\frac{1-A}{\check{\alpha}}\int_0^\tau\check{\pi}(X)\frac{\check{q}_1}{\check{b}_1}(t\,|\,X)I(\tilde{T}\geqslant t)\frac{\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})}{\check{S}_{\bullet 1}}(t\,|\,0,X)\right\}\\
&=P\left\{\frac{1}{\check{\alpha}}\int_0^\tau\check{\pi}(X)\frac{\check{q}_1}{\check{b}_1}(t\,|\,X)P(\tilde{T}\geqslant t,A=0\,|\,X)\frac{\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})}{\check{S}_{\bullet 1}}(t\,|\,0,X)\right\}
\end{aligned}$$

$$= P\left\{\frac{1}{\check{\alpha}}\int_0^\tau \check{\pi}(X)\frac{\check{q}_1}{\check{b}_1}(t\,|\,X)\frac{1}{\check{S}_{\bullet 1}(t\,|\,0,X)}b_1(t\,|\,X)S_{\bullet 1}(t-\,|\,0,X)\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})(t\,|\,0,X)\right\}$$
$$= P\left\{\frac{1}{\check{\alpha}}\int_0^\tau \check{\pi}(X)\frac{b_1\check{q}_1}{\check{b}_1}(t\,|\,X)\frac{S_{\bullet 1}}{\check{S}_{\bullet 1}}(t\,|\,0,X)\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})(t\,|\,0,X)\right\}.$$

Also, we have

$$P\check{\ell}_{12}(0) = P\left\{\frac{D(1-A)}{\check{\alpha}}\int_0^\tau \frac{\check{q}_2}{\check{b}_2}(t\,|\,X)\frac{\mathrm{d}\check{M}_{12}}{\check{S}_{12}}(t\,|\,0,X)\right\}$$
$$= P\left[\frac{\pi(X)}{\check{\alpha}}\int_0^\tau \frac{b_2\check{q}_2}{\check{b}_2}(t\,|\,X)\frac{S_{12}(t-\,|\,0,X)}{\check{S}_{12}(t\,|\,0,X)}\mathrm{d}(\mathrm{A}_{12}-\check{\mathrm{A}}_{12})(t\,|\,0,X)\right].$$

Let $\mathrm{A}_1(t\,|\,0,x) = (\mathrm{A}_{\bullet 1}+\mathrm{A}_{12})(t\,|\,0,x)$ and $\check{\mathrm{A}}_1(t\,|\,0,x) = (\check{\mathrm{A}}_{\bullet 1}+\check{\mathrm{A}}_{12})(t\,|\,0,x)$, the all-cause hazard. By the Duhamel equation in Lemma S2,

$$\begin{aligned}
&(\check{F}_{11}-F_{11})(\tau\,|\,0,x)\\
&\quad= \int_0^\tau \check{S}_1(t-\,|\,0,x)\mathrm{d}\check{\mathrm{A}}_{\bullet 1}(t\,|\,0,x) - \int_0^\tau S_1(t-\,|\,0,x)\mathrm{d}\mathrm{A}_{\bullet 1}(t\,|\,0,x)\\
&\quad= \int_0^\tau (\check{S}_1-S_1)(t-\,|\,0,x)\mathrm{d}\check{\mathrm{A}}_{\bullet 1}(t\,|\,0,x) + \int_0^\tau S_1(t-\,|\,0,x)\mathrm{d}(\check{\mathrm{A}}_{\bullet 1}-\mathrm{A}_{\bullet 1})(t\,|\,0,x)\\
&\quad= \int_0^\tau \int_{s\in(0,t)} \check{S}_1(t-\,|\,0,x)\frac{S_1(s-\,|\,0,x)}{\check{S}_1(s\,|\,0,x)}\mathrm{d}(\mathrm{A}_1-\check{\mathrm{A}}_1)(s\,|\,0,x)\mathrm{d}\check{\mathrm{A}}_{\bullet 1}(t\,|\,0,x)\\
&\qquad+ \int_0^\tau S_1(t-\,|\,0,x)\mathrm{d}(\check{\mathrm{A}}_{\bullet 1}-\mathrm{A}_{\bullet 1})(t\,|\,0,x)\\
&\quad= \int_{s\in(0,\tau)}\int_s^\tau \check{S}_1(t-\,|\,0,x)\mathrm{d}\check{\mathrm{A}}_{\bullet 1}(t\,|\,0,x)\frac{S_1(s-\,|\,0,x)}{\check{S}_1(s\,|\,0,x)}\mathrm{d}(\mathrm{A}_1-\check{\mathrm{A}}_1)(s\,|\,0,x)\\
&\qquad+ \int_0^\tau S_1(t-\,|\,0,x)\mathrm{d}(\check{\mathrm{A}}_{\bullet 1}-\mathrm{A}_{\bullet 1})(t\,|\,0,x)\\
&\quad= \int_{s\in(0,\tau)} \frac{\check{F}_{11}(\tau\,|\,0,x)-\check{F}_{11}(s\,|\,0,x)}{1-\triangle\check{\mathrm{A}}_1(s\,|\,0,x)}\frac{S_1}{\check{S}_1}(s-\,|\,0,x)\mathrm{d}(\mathrm{A}_1-\check{\mathrm{A}}_1)(s\,|\,0,x)\\
&\qquad+ \int_0^\tau S_1(t-\,|\,0,x)\mathrm{d}(\check{\mathrm{A}}_{\bullet 1}-\mathrm{A}_{\bullet 1})(t\,|\,0,x)\\
&\quad= -\int_0^\tau \left\{S_1(t-\,|\,0,x) - \frac{\check{F}_{11}(\tau\,|\,0,x)-\check{F}_{11}(t\,|\,0,x)}{1-\triangle\check{\mathrm{A}}_{\bullet 1}(t\,|\,0,x)}\frac{S_1}{\check{S}_1}(t-\,|\,0,x)\right\}\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})(t\,|\,0,x)\\
&\qquad+ \int_0^\tau \frac{\check{F}_{11}(\tau\,|\,0,x)-\check{F}_{11}(t\,|\,0,x)}{1-\triangle\check{\mathrm{A}}_{12}(t\,|\,0,x)}\frac{S_1}{\check{S}_1}(t-\,|\,0,x)\mathrm{d}(\mathrm{A}_{12}-\check{\mathrm{A}}_{12})(t\,|\,0,x)\\
&\quad= -\int_0^\tau \left\{S_{12}(t-\,|\,0,x) - \frac{\check{F}_{11}(\tau\,|\,0,x)-\check{F}_{11}(t\,|\,0,x)}{\check{S}_{\bullet 1}(t\,|\,0,x)}\frac{S_{12}}{\check{S}_{12}}(t-\,|\,0,x)\right\}\\
&\qquad\qquad S_{\bullet 1}(t-\,|\,0,x)\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})(t\,|\,0,x)\\
&\qquad+ \int_0^\tau \frac{\check{F}_{11}(\tau\,|\,0,x)-\check{F}_{11}(t\,|\,0,x)}{\check{S}_{12}(t\,|\,0,x)}\frac{S_{\bullet 1}S_{12}}{\check{S}_{\bullet 1}}(t-\,|\,0,x)\mathrm{d}(\mathrm{A}_{12}-\check{\mathrm{A}}_{12})(t\,|\,0,x)\\
&\quad= -\int_0^\tau \frac{S_{12}}{\check{S}_{12}}(t-\,|\,0,x)\left\{\check{S}_{\bullet 1}(t\,|\,0,x)\check{S}_{12}(t-\,|\,0,x) - \check{F}_{11}(\tau\,|\,0,x)+\check{F}_{11}(t\,|\,0,x)\right\}\\
&\qquad\qquad \frac{S_{\bullet 1}(t-\,|\,0,x)}{\check{S}_{\bullet 1}(t\,|\,0,x)}\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})(t\,|\,0,x)\\
&\qquad+ \int_0^\tau \frac{S_{\bullet 1}}{\check{S}_{\bullet 1}}(t-\,|\,0,x)\left\{\check{F}_{11}(\tau\,|\,0,x)-\check{F}_{11}(t\,|\,0,x)\right\}\frac{S_{12}(t-\,|\,0,x)}{\check{S}_{12}(t\,|\,0,x)}\mathrm{d}(\mathrm{A}_{12}-\check{\mathrm{A}}_{12})(t\,|\,0,x).
\end{aligned}$$

By Assumption S1,

$$\check{S}_{12}(t- \mid 0,x)\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})(t \mid 0,x)=\check{S}_{12}(t \mid 0,x)\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})(t \mid 0,x).$$

Therefore,

$$\begin{aligned}
&P\left\{\check{\ell}_{13}(0)-\frac{\alpha}{\check{\alpha}}\ell_1(0)\right\}\\
&\quad=P\check{\ell}_{13}(0)-\frac{\alpha}{\check{\alpha}}\theta_1(0)\\
&\quad=P\left[\frac{D}{\check{\alpha}}(\check{F}_{11}-F_{11})(\tau \mid 0,X)\right]\\
&\quad=-P\left[\frac{\pi(X)}{\check{\alpha}}\int_0^\tau \frac{S_{12}}{\check{S}_{12}}(t- \mid 0,X)\check{q}_1(t \mid X)\frac{S_{\bullet 1}(t- \mid 0,X)}{\check{S}_{\bullet 1}(t \mid 0,X)}\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})(t \mid 0,X)\right]\\
&\qquad-P\left[\frac{\pi(X)}{\check{\alpha}}\int_0^\tau \frac{S_{\bullet 1}}{\check{S}_{\bullet 1}}(t- \mid 0,X)\check{q}_2(t \mid X)\frac{S_{12}(t- \mid 0,X)}{\check{S}_{12}(t \mid 0,X)}\mathrm{d}(\mathrm{A}_{12}-\check{\mathrm{A}}_{12})(t \mid 0,X)\right].
\end{aligned}$$

Summing up the three terms in the previous displays gives

$$\begin{aligned}
&P\left\{\check{\ell}_1(0)-\frac{\alpha}{\check{\alpha}}\ell_1(0)\right\}\\
&\quad=P\left[\frac{1}{\check{\alpha}}\int_0^\tau\left\{\check{\pi}(X)\frac{b_1}{\check{b}_1}(t \mid X)-\pi(X)\frac{S_{12}}{\check{S}_{12}}(t- \mid 0,X)\right\}\right.\\
&\qquad\qquad\left.\check{q}_1(t \mid X)\frac{S_{\bullet 1}(t- \mid 0,X)}{\check{S}_{\bullet 1}(t \mid 0,X)}\mathrm{d}(\mathrm{A}_{\bullet 1}-\check{\mathrm{A}}_{\bullet 1})(t \mid 0,X)\right]\\
&\qquad+P\left[\frac{\pi(X)}{\check{\alpha}}\int_0^\tau\left\{\frac{b_2}{\check{b}_2}(t \mid X)-\frac{S_{\bullet 1}}{\check{S}_{\bullet 1}}(t- \mid 0,X)\right\}\right.\\
&\qquad\qquad\left.\check{q}_2(t \mid X)\frac{S_{12}(t- \mid 0,X)}{\check{S}_{12}(t \mid 0,X)}\mathrm{d}(\mathrm{A}_{12}-\check{\mathrm{A}}_{12})(t \mid 0,X)\right]\\
&\quad=P\left[\frac{1}{\check{\alpha}}\int_0^\tau\left\{\check{\pi}(X)\frac{b_1}{\check{b}_1}(t \mid X)-\pi(X)\frac{S_{12}}{\check{S}_{12}}(t- \mid 0,X)\right\}\check{q}_1(t \mid X)\mathrm{d}\left(1-\frac{S_{\bullet 1}}{\check{S}_{\bullet 1}}\right)(t \mid 0,X)\right]\\
&\qquad+P\left[\frac{\pi(X)}{\check{\alpha}}\int_0^\tau\left\{\frac{b_2}{\check{b}_2}(t \mid X)-\frac{S_{\bullet 1}}{\check{S}_{\bullet 1}}(t- \mid 0,X)\right\}\check{q}_2(t \mid X)\mathrm{d}\left(1-\frac{S_{12}}{\check{S}_{12}}\right)(t \mid 0,X)\right],
\end{aligned}$$

where the last step is again by the Duhamel equation in Lemma S2. □

Let $\bar{\ell}_1(0)$ be obtained by substituting the probability limits of the nuisance parameters into the function $\check{\ell}_1(0)$.

Lemma S4. *Suppose Assumptions 6 and S1 hold. Then* $P\{\hat{\ell}_1(0)-\bar{\ell}_1(0)\}^2 \xrightarrow{\mathrm{p}} 0$.

Proof. We use the notation $A_n \lesssim B_n$ to denote $A_n \leqslant CB_n$ for some universal constant $C \geqslant 1$. Let

$$\begin{aligned}
\hat{b}_1(t \mid X)&=\hat{\pi}(X)\hat{e}_1(0 \mid X)(\hat{S}_{12}\hat{S}_1^c)(t- \mid 0,X)+\{1-\hat{\pi}(X)\}(\hat{S}_{02}\hat{S}_0^c)(t- \mid X),\\
\bar{b}_1(t \mid X)&=\bar{\pi}(X)\bar{e}_1(0 \mid X)(\bar{S}_{12}\bar{S}_1^c)(t- \mid 0,X)+\{1-\bar{\pi}(X)\}(\bar{S}_{02}\bar{S}_0^c)(t- \mid X),\\
\hat{b}_2(t \mid X)&=\hat{e}_1(0 \mid X)(\hat{S}_{\bullet 1}\hat{S}_1^c)(t- \mid 0,X),\\
\bar{b}_2(t \mid X)&=\bar{e}_1(0 \mid X)(\bar{S}_{\bullet 1}\bar{S}_1^c)(t- \mid 0,X),\\
\hat{r}_1(t \mid X)&=\frac{1}{\hat{S}_{\bullet 1}(\tau \mid 0,X)}\{\hat{S}_1(t \mid 0,X)-\hat{F}_{11}(\tau \mid 0,X)+\hat{F}_{11}(t \mid 0,X)\},
\end{aligned}$$

$$\bar{r}_1(t \mid X) = \frac{1}{\bar{S}_{\bullet 1}(\tau \mid 0, X)}\{\bar{S}_1(t \mid 0, X) - \bar{F}_{11}(\tau \mid 0, X) + \bar{F}_{11}(t \mid 0, X)\},$$
$$\hat{r}_2(t \mid X) = -\frac{1}{\hat{S}_{12}(\tau \mid 0, X)}\{\hat{F}_{11}(\tau \mid 0, X) - \hat{F}_{11}(t \mid 0, X)\},$$
$$\bar{r}_2(t \mid X) = -\frac{1}{\bar{S}_{12}(\tau \mid 0, X)}\{\bar{F}_{11}(\tau \mid 0, X) - \bar{F}_{11}(t \mid 0, X)\}.$$

Then

$$\hat{\ell}_1(0)(O) = \frac{1-A}{\hat{\alpha}}\int_0^\tau \hat{\pi}(X)\frac{\hat{r}_1}{\hat{b}_1}(t \mid X)\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)}\mathrm{d}\hat{M}_{\bullet 1}(t \mid 0, X)$$
$$+ \frac{D(1-A)}{\hat{\alpha}}\int_0^\tau \frac{\hat{r}_2}{\hat{b}_2}(t \mid X)\frac{\hat{S}_{12}(\tau \mid 0, X)}{\hat{S}_{12}(t \mid 0, X)}\mathrm{d}\hat{M}_{12}(t \mid 0, X) + \frac{D}{\hat{\alpha}}\hat{F}_{11}(\tau \mid 0, X),$$
$$\bar{\ell}_1(0)(O) = \frac{1-A}{\bar{\alpha}}\int_0^\tau \bar{\pi}(X)\frac{\bar{r}_1}{\bar{b}_1}(t \mid X)\frac{\bar{S}_{\bullet 1}(\tau \mid 0, X)}{\bar{S}_{\bullet 1}(t \mid 0, X)}\mathrm{d}\bar{M}_{\bullet 1}(t \mid 0, X)$$
$$+ \frac{D(1-A)}{\bar{\alpha}}\int_0^\tau \frac{\bar{r}_2}{\bar{b}_2}(t \mid X)\frac{\bar{S}_{12}(\tau \mid 0, X)}{\bar{S}_{12}(t \mid 0, X)}\mathrm{d}\bar{M}_{12}(t \mid 0, X) + \frac{D}{\bar{\alpha}}\bar{F}_{11}(\tau \mid 0, X).$$

By Assumption S1, uniformly for $t \in (0, \tau]$ and $x \in \mathcal{X}_1$, $\hat{b}_1$, $\bar{b}_1$, $\hat{b}_2$, and $\bar{b}_2$ are bounded away from 0 and from above, $\hat{r}_1$ and $\bar{r}_1$ are positive and bounded from above, while $\hat{r}_2$ and $\bar{r}_2$ negative and bounded from below.

Decompose the difference as

$$\hat{\ell}_1(0) - \bar{\ell}_1(0) = \sum_{m=1}^{10}\delta_m,$$

where

$$\delta_1 = \frac{D}{\hat{\alpha}}\hat{F}_{11}(\tau \mid 0, X) - \frac{D}{\bar{\alpha}}\bar{F}_{11}(\tau \mid 0, X),$$
$$\delta_2 = \frac{1-A}{\hat{\alpha}}\int_0^\tau (\hat{\pi} - \bar{\pi})(X)\frac{\hat{r}_1}{\hat{b}_1}(t \mid X)\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)}\mathrm{d}\hat{M}_{\bullet 1}(t \mid 0, X),$$
$$\delta_3 = \frac{1-A}{\hat{\alpha}}\int_0^\tau \bar{\pi}(X)\frac{\hat{r}_1 - \bar{r}_1}{\hat{b}_1}(t \mid X)\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)}\mathrm{d}\hat{M}_{\bullet 1}(t \mid 0, X),$$
$$\delta_4 = -\frac{1-A}{\hat{\alpha}\bar{\alpha}}\int_0^\tau \bar{\pi}(X)\frac{\bar{r}_1(\hat{\alpha}\hat{b}_1 - \bar{\alpha}\bar{b}_1)}{\hat{b}_1\bar{b}_1}(t \mid X)\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)}\mathrm{d}\hat{M}_{\bullet 1}(t \mid 0, X),$$
$$\delta_5 = \frac{1-A}{\bar{\alpha}}\int_0^\tau \bar{\pi}(X)\frac{\bar{r}_1}{\bar{b}_1}(t \mid X)\left\{\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)} - \frac{\bar{S}_{\bullet 1}(\tau \mid 0, X)}{\bar{S}_{\bullet 1}(t \mid 0, X)}\right\}\mathrm{d}N_1(t),$$
$$\delta_6 = -\frac{1-A}{\bar{\alpha}}\int_0^\tau \bar{\pi}(X)\frac{\bar{r}_1}{\bar{b}_1}(t \mid X)Y(t)\left\{\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)}\mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(t \mid 0, X) - \frac{\bar{S}_{\bullet 1}(\tau \mid 0, X)}{\bar{S}_{\bullet 1}(t \mid 0, X)}\mathrm{d}\bar{\mathrm{A}}_{\bullet 1}(t \mid 0, X)\right\},$$
$$\delta_7 = \frac{D}{\hat{\alpha}}(1-A)\int_0^\tau \frac{\hat{r}_2 - \bar{r}_2}{\hat{b}_2}(t \mid X)\frac{\hat{S}_{12}(\tau \mid 0, X)}{\hat{S}_{12}(t \mid 0, X)}\mathrm{d}\hat{M}_{12}(t \mid 0, X),$$
$$\delta_8 = -\frac{D}{\hat{\alpha}\bar{\alpha}}(1-A)\int_0^\tau \frac{\bar{r}_2(\hat{\alpha}\hat{b}_2 - \bar{\alpha}\bar{b}_2)}{(\hat{b}_2\bar{b}_2)}(t \mid X)\frac{\hat{S}_{12}(\tau \mid 0, X)}{\hat{S}_{12}(t \mid 0, X)}\mathrm{d}\hat{M}_{12}(t \mid 0, X),$$
$$\delta_9 = \frac{D}{\bar{\alpha}}(1-A)\int_0^\tau \frac{\bar{r}_2}{\bar{b}_2}(t \mid X)\left\{\frac{\hat{S}_{12}(\tau \mid 0, X)}{\hat{S}_{12}(t \mid 0, X)} - \frac{\bar{S}_{12}(\tau \mid 0, X)}{\bar{S}_{12}(t \mid 0, X)}\right\}\mathrm{d}N_2(t),$$
$$\delta_{10} = -\frac{D}{\bar{\alpha}}(1-A)\int_0^\tau \frac{\bar{r}_2}{\bar{b}_2}(t \mid X)Y(t)\left\{\frac{\hat{S}_{12}(\tau \mid 0, X)}{\hat{S}_{12}(t \mid 0, X)}\mathrm{d}\hat{\mathrm{A}}_{12}(t \mid 0, X) - \frac{\bar{S}_{12}(\tau \mid 0, X)}{\bar{S}_{12}(t \mid 0, X)}\mathrm{d}\bar{\mathrm{A}}_{12}(t \mid 0, X)\right\},$$

where in the second to last step, we used Lemma S1.

We first relate the difference of cumulative incidences for cause 1 to the survival functions as follows: for $t \in (0, \tau]$,

$$
\begin{aligned}
&|\hat{F}_{11} - \bar{F}_{11}|(t \mid 0, X) \\
&\quad = \left| \int_0^t (\hat{S}_{\bullet 1}\hat{S}_{12})(s- \mid 0, X) \mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(s \mid 0, X) - \int_0^\tau (\bar{S}_{\bullet 1}\bar{S}_{12})(s- \mid 0, X) \mathrm{d}\bar{\mathrm{A}}_{\bullet 1}(s \mid 0, X) \right| \\
&\quad \leqslant \left| \int_0^t \{(\hat{S}_{12} - \bar{S}_{12})\hat{S}_{\bullet 1}\}(s- \mid 0, X) \mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(s \mid 0, X) \right| \\
&\qquad + \left| \int_0^t \bar{S}_{12}(s- \mid 0, X) \{\hat{S}_{\bullet 1}(s- \mid 0, X) \mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(s \mid 0, X) - \bar{S}_{\bullet 1}(s- \mid 0, X) \mathrm{d}\bar{\mathrm{A}}_{\bullet 1}(s \mid 0, X)\} \right| \\
&\quad \leqslant \sup_{s \in (0,\tau]} |\hat{S}_{12} - \bar{S}_{12}|(s \mid 0, X) \hat{S}_{\bullet 1}(\tau \mid 0, X) \\
&\qquad + \left| \{(\bar{S}_{\bullet 1} - \hat{S}_{\bullet 1})\bar{S}_{12}\}(t \mid 0, X) - \int_0^t (\bar{S}_{\bullet 1} - \hat{S}_{\bullet 1})(s \mid 0, X) \mathrm{d}\bar{S}_{12}(s \mid 0, X) \right| \\
&\quad \leqslant \sup_{s \in (0,\tau]} |\hat{S}_{12} - \bar{S}_{12}|(s \mid 0, X) + \sup_{s \in (0,\tau]} |\hat{S}_{\bullet 1} - \bar{S}_{\bullet 1}|(s \mid 0, X).
\end{aligned}
$$

By the triangular inequality, $\|\hat{\ell}_1(0) - \bar{\ell}_1(0)\|_P \leqslant \sum_{m=1}^{10} \{P\delta_m^2\}^{1/2}$. Below we bound each term $P\delta_m^2$. Let P_1 denote the probability measure $P(\cdot \mid D = 1)$.

Term δ_1.

$$
\begin{aligned}
P\delta_1^2 &= P\left\{ \frac{D}{\hat{\alpha}} \hat{F}_{11}(\tau \mid 0, X) - \frac{D}{\bar{\alpha}} \bar{F}_{11}(\tau \mid 0, X) \right\}^2 \\
&\lesssim P\left[\frac{D}{\hat{\alpha}} \{\hat{F}_{11}(\tau \mid 0, X) - \bar{F}_{11}(\tau \mid 0, X)\} \right]^2 + P\left\{ D \frac{\hat{\alpha} - \bar{\alpha}}{\bar{\alpha}\hat{\alpha}} \bar{F}_{11}(\tau \mid 0, X) \right\}^2 \\
&\lesssim P_1(\hat{F}_{11} - \bar{F}_{11})^2(\tau \mid 0, X) + |\hat{\alpha} - \bar{\alpha}|^2 \\
&\leqslant |\hat{\alpha} - \bar{\alpha}|^2 + P_1\left\{ \sup_{t \in (0,\tau]} |\hat{S}_{12} - \bar{S}_{12}|(t \mid 0, X) \right\}^2 + P_1\left\{ \sup_{t \in (0,\tau]} |\hat{S}_{\bullet 1} - \bar{S}_{\bullet 1}|(t \mid 0, X) \right\}^2.
\end{aligned}
$$

Term δ_2.

$$
\begin{aligned}
P\delta_2^2 &= P\left[\frac{1 - A}{\hat{\alpha}} \int_0^\tau (\hat{\pi} - \bar{\pi})(X) \frac{\hat{r}_1}{\hat{b}_1}(t \mid X) \frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)} \mathrm{d}\hat{M}_{\bullet 1}(t \mid 0, X) \right]^2 \\
&\lesssim P\left[(\hat{\pi} - \bar{\pi})(X) \int_0^\tau \frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)} \mathrm{d}\hat{M}_{\bullet 1}(t \mid 0, X) \right]^2 \\
&\leqslant P\left[|\hat{\pi} - \bar{\pi}|(X) \int_0^\tau \frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)} \{\mathrm{d}N_1(t) + Y(t) \mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(t \mid 0, X)\} \right]^2 \\
&\lesssim P|\hat{\pi} - \bar{\pi}|^2(X),
\end{aligned}
$$

where in the last step we used

$$
\int_0^\tau \frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)} \mathrm{d}N_1(t) = I(\tilde{T} \leqslant \tau, \tilde{J} = 1) \frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(\tilde{T} \mid 0, X)} \leqslant 1,
$$

and, by the backward equation in Lemma S2,

$$
\int_0^\tau \frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)} Y(t) \mathrm{d}\hat{\mathrm{A}}_{\bullet 1}(t) = \frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(\tilde{T} \wedge \tau \mid 0, X)} - \hat{S}_{\bullet 1}(\tau \mid 0, X) \leqslant 2.
$$

Term δ_3.

$$
\begin{aligned}
P\delta_3^2 &= P\left[\frac{1-A}{\hat{\alpha}}\int_0^\tau \bar{\pi}(X)\frac{\hat{r}_1-\bar{r}_1}{\hat{b}_1}(t\,|\,X)\frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t\,|\,0,X)}\mathrm{d}\hat{M}_{\bullet 1}(t\,|\,0,X)\right]^2 \\
&\lesssim P\left[\int_0^\tau (\hat{r}_1-\bar{r}_1)(t\,|\,X)\frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t\,|\,0,X)}\mathrm{d}\hat{M}_{\bullet 1}(t\,|\,0,X)\right]^2 \\
&= P\Bigg\{\int_0^\tau \Bigg(\frac{\bar{S}_{\bullet 1}-\hat{S}_{\bullet 1}}{\bar{S}_{\bullet 1}}(\tau\,|\,0,X)\hat{r}_1(t\,|\,X)+\frac{1}{\bar{S}_{\bullet 1}(\tau\,|\,0,X)}\Big[\{(\hat{S}_{\bullet 1}-\bar{S}_{\bullet 1})\}\hat{S}_{12}(t\,|\,0,X) \\
&\qquad\qquad +\{\bar{S}_{\bullet 1}(\hat{S}_{12}-\bar{S}_{12})\}(t\,|\,0,X)-(\hat{F}_{11}-\bar{F}_{11})(\tau\,|\,0,X)+(\hat{F}_{11}-\bar{F}_{11})(t\,|\,0,X)\Big]\Bigg) \\
&\lesssim P\Bigg[\int_0^\tau \big\{|\bar{S}_{\bullet 1}-\hat{S}_{\bullet 1}|(\tau\,|\,0,X)+|\hat{S}_{\bullet 1}-\bar{S}_{\bullet 1}|(t\,|\,0,X) \\
&\qquad\qquad +|\hat{S}_{12}-\bar{S}_{12}|(t\,|\,0,X)|+|\hat{F}_{11}-\bar{F}_{11}|(\tau\,|\,0,X)+|\hat{F}_{11}-\bar{F}_{11}|(t\,|\,0,X)\big\} \\
&\qquad\qquad \frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t\,|\,0,X)}\{\mathrm{d}N_1(t)+Y(t)\mathrm{d}\hat{\Lambda}_{\bullet 1}(t\,|\,0,X)\}\Bigg]^2 \\
&\lesssim P_1\Bigg[\Big\{\sup_{t\in(0,\tau]}|\hat{S}_{\bullet 1}-\bar{S}_{\bullet 1}|(t\,|\,0,X)+\sup_{t\in(0,\tau]}|\hat{S}_{12}-\bar{S}_{12}|(t\,|\,0,X) \\
&\qquad\qquad +\sup_{t\in(0,\tau]}|\hat{F}_{11}-\bar{F}_{11}|(t\,|\,0,X)\Big\}\int_0^\tau \frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t\,|\,0,X)}\{\mathrm{d}N_1(t)+Y(t)\mathrm{d}\hat{\Lambda}_{\bullet 1}(t\,|\,0,X)\}\Bigg] \\
&\lesssim P_1\Big\{\sup_{t\in(0,\tau]}|\hat{S}_{\bullet 1}-\bar{S}_{\bullet 1}|(t\,|\,0,X)\Big\}^2+P_1\Big\{\sup_{t\in(0,\tau]}|\hat{S}_{12}-\bar{S}_{12}|(t\,|\,0,X)\Big\}^2.
\end{aligned}
$$

Term δ_4.

$$
\begin{aligned}
P\delta_4^2 &= P\left\{\frac{1-A}{\hat{\alpha}\bar{\alpha}}\int_0^\tau \bar{\pi}(X)\frac{\bar{r}_1(\hat{\alpha}\hat{b}_1-\bar{\alpha}\bar{b}_1)}{\hat{b}_1\bar{b}_1}(t\,|\,X)\frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t\,|\,0,X)}\mathrm{d}\hat{M}_{\bullet 1}(t\,|\,0,X)\right\}^2 \\
&\lesssim P_1\left\{\int_0^\tau (\hat{\alpha}\hat{b}_1-\bar{\alpha}\bar{b}_1)(t\,|\,X)\frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t\,|\,0,X)}\mathrm{d}\hat{M}_{\bullet 1}(t\,|\,0,X)\right\}^2 \\
&= P_1\Bigg\{\int_0^\tau \Big[(\hat{\alpha}-\bar{\alpha})\hat{b}_1(t\,|\,X)+\bar{\alpha}(\hat{\pi}-\bar{\pi})(X)\hat{e}_1(0\,|\,X)(\hat{S}_{12}\hat{S}_1^c)(t-\,|\,0,X) \\
&\quad +\bar{\alpha}\bar{\pi}(X)(\hat{e}_1-\bar{e}_1)(0\,|\,X)(\hat{S}_{12}\hat{S}_1^c)(t-\,|\,0,X) \\
&\quad +\bar{\alpha}\bar{\pi}(X)\bar{e}_1(0\,|\,X)\{(\hat{S}_{12}-\bar{S}_{12})\hat{S}_1^c+\bar{S}_{12}(\hat{S}_1^c-\bar{S}_1^c)\}(t-\,|\,0,X) \\
&\quad +\bar{\alpha}(\bar{\pi}-\hat{\pi})(X)(\hat{S}_{02}\hat{S}_0^c)(t-\,|\,X) \\
&\quad +\bar{\alpha}\{1-\bar{\pi}(X)\}\{(\hat{S}_{02}-\bar{S}_{02})\hat{S}_0^c+\bar{S}_{02}(\hat{S}_0^c-\bar{S}_0^c)\}(t-\,|\,X)\Big] \\
&\qquad \frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t\,|\,0,X)}\mathrm{d}\hat{M}_{\bullet 1}(t)\Bigg\}^2 \\
&\lesssim P\left[\big\{|\hat{\alpha}-\bar{\alpha}|+|\hat{\pi}-\bar{\pi}|(X)\big\}\int_0^\tau \frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t\,|\,0,X)}\{\mathrm{d}N_1(t)+Y(t)\mathrm{d}\hat{\Lambda}_{\bullet 1}(t\,|\,0,X)\}\right]^2 \\
&\quad +PI\{\pi(X)>0\}\Bigg[\int_0^\tau \big\{|\hat{e}_1-\bar{e}_1|(0\,|\,X)+|\hat{S}_{12}-\bar{S}_{12}|(t-\,|\,0,X)+|\hat{S}_1^c-\bar{S}_1^c|(t-\,|\,0,X)\big\} \\
&\quad \frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t\,|\,0,X)}\{\mathrm{d}N_1(t)+Y(t)\mathrm{d}\hat{\Lambda}_{\bullet 1}(t\,|\,0,X)\}\Bigg]^2 \\
&\quad +PI[\pi(X)\{1-\pi(X)\}>0]\Bigg[\int_0^\tau \big\{|\hat{S}_{02}-\bar{S}_{02}|(t-\,|\,0,X)+|\hat{S}_0^c-\bar{S}_0^c|(t-\,|\,0,X)\big\}
\end{aligned}
$$

$$
\begin{aligned}
&\left.\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)}\{\mathrm{d}N_1(t) + Y(t)\mathrm{d}\hat{\Lambda}_{\bullet 1}(t \mid 0, X)\}\right]^2 \\
&\lesssim |\hat{\alpha} - \bar{\alpha}|^2 + P|\hat{\pi} - \bar{\pi}|^2(X) + PI\{\pi(X) > 0\}|\hat{e}_1 - \bar{e}_1|^2(0 \mid X) \\
&\quad + PI\{\pi(X) > 0\}\left\{\sup_{t \in (0,\tau]} |\hat{S}_{\bullet 1} - \bar{S}_{\bullet 1}|(t \mid 0, X)\right\}^2 \\
&\quad + PI\{\pi(X) > 0\}\left\{\sup_{t \in (0,\tau]} |\hat{S}_{12} - \bar{S}_{12}|(t \mid 0, X)\right\}^2 \\
&\quad + PI[\pi(X)\{1 - \pi(X)\} > 0]\left\{\sup_{t \in (0,\tau]} |\hat{S}_{02} - \bar{S}_{02}|(t \mid X)\right\}^2 \\
&\quad + PI[\pi(X)\{1 - \pi(X)\} > 0]\left\{\sup_{t \in (0,\tau]} |\hat{S}_1^c - \bar{S}_1^c|(t \mid 0, X)\right\}^2 \\
&\quad + PI[\pi(X)\{1 - \pi(X)\} > 0]\left\{\sup_{t \in (0,\tau]} |\hat{S}_0^c - \bar{S}_0^c|(t \mid X)\right\}^2.
\end{aligned}
$$

Term δ_5.

$$
\begin{aligned}
P\delta_5^2 &= P\left\{\frac{1 - A}{\bar{\alpha}} \int_0^\tau \bar{\pi}(X)\frac{\bar{r}_1}{\bar{b}_1}(t \mid X)\left\{\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)} - \frac{\bar{S}_{\bullet 1}(\tau \mid 0, X)}{\bar{S}_{\bullet 1}(t \mid 0, X)}\right\}\mathrm{d}N_1(t)\right\}^2 \\
&\lesssim PI\{\pi(X) > 0\}I\{\tilde{T} \leqslant \tau, \tilde{J} = 1\}\left|\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(\tilde{T} \mid 0, X)} - \frac{\bar{S}_{\bullet 1}(\tau \mid 0, X)}{\bar{S}_{\bullet 1}(\tilde{T} \mid 0, X)}\right|^2 \\
&\lesssim PI\{\pi(X) > 0\}\left\{\sup_{t \in (0,\tau]} |\hat{S}_{\bullet 1} - \bar{S}_{\bullet 1}|(t \mid 0, X)\right\}^2.
\end{aligned}
$$

Term δ_6.

$$
\begin{aligned}
P\delta_6^2 &= P\left[\frac{1 - A}{\bar{\alpha}} \int_0^\tau \bar{\pi}(X)\frac{\bar{r}_1}{\bar{b}_1}(t \mid X)Y(t)\right. \\
&\qquad \left.\left\{\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)}\mathrm{d}\hat{\Lambda}_{\bullet 1}(t \mid 0, X) - \frac{\bar{S}_{\bullet 1}(\tau \mid 0, X)}{\bar{S}_{\bullet 1}(t \mid 0, X)}\mathrm{d}\bar{\Lambda}_{\bullet 1}(t \mid 0, X)\right\}\right]^2 \\
&\lesssim PI\{\pi(X) > 0\}\left[\int_0^\tau \frac{\bar{r}_1}{\bar{b}_1}(t \mid X)Y(t)\mathrm{d}\left\{\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)} - \frac{\bar{S}_{\bullet 1}(\tau \mid 0, X)}{\bar{S}_{\bullet 1}(t \mid 0, X)}\right\}\right]^2.
\end{aligned}
$$

By Assumption S1, $\{\hat{\Lambda}_{\bullet 1}, \bar{\Lambda}_{\bullet 1}\}$ do not share any discontinuity with $\{\bar{\Lambda}_{12}, \bar{\Lambda}_{02}, \bar{\Lambda}_1^c, \bar{\Lambda}_0^c\}$, then we may replace $\bar{b}(t \mid X)$ in the display before with the right-continuous version $\bar{b}(t+ \mid X)$, which equals

$$
\bar{\pi}(X)\bar{e}_1(0 \mid X)(\bar{S}_{12}\bar{S}_1^c)(t \mid 0, X) + \{1 - \bar{\pi}(X)\}(\bar{S}_{02}\bar{S}_0^c)(t \mid X).
$$

Therefore, we can apply integration by parts from Lemma S1. This leads to

$$
\begin{aligned}
P\delta_6^2 &= PI\{\pi(X) > 0\}\left[\frac{\bar{r}_1}{\bar{b}_1}(t \mid X)\left\{\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t \mid 0, X)} - \frac{\bar{S}_{\bullet 1}(\tau \mid 0, X)}{\bar{S}_{\bullet 1}(t \mid 0, X)}\right\}\right|_0^{\tau \wedge \tilde{T}} \\
&\qquad - \int_0^{\tau \wedge \tilde{T}} \left\{\frac{\hat{S}_{\bullet 1}(\tau \mid 0, X)}{\hat{S}_{\bullet 1}(t- \mid 0, X)} - \frac{\bar{S}_{\bullet 1}(\tau \mid 0, X)}{\bar{S}_{\bullet 1}(t- \mid 0, X)}\right\}\mathrm{d}\left(\frac{\bar{r}_1}{\bar{b}_1}\right)(t \mid X)\Bigg]^2 \\
&\lesssim PI\{\pi(X) > 0\}\left\{\sup_{t \in (0,\tau]} |\hat{S}_{\bullet 1} - \bar{S}_{\bullet 1}|(t \mid 0, X)\right\}^2
\end{aligned}
$$

$$+ PI\{\pi(X) > 0\}\left[\left\{\sup_{t\in(0,\tau\wedge\tilde{T}]}\left|\frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t-\,|\,0,X)} - \frac{\bar{S}_{\bullet 1}(\tau\,|\,0,X)}{\bar{S}_{\bullet 1}(t-\,|\,0,X)}\right|\right\}\left\{\frac{-\bar{S}_1(t\,|\,0,X)}{\bar{b}_1(t\,|\,X)}\right\}\Bigg|_0^{\tau\wedge\tilde{T}}\right]^2$$

$$+ PI\{\pi(X) > 0\}\left[\left\{\sup_{t\in(0,\tau\wedge\tilde{T}]}\left|\frac{\hat{S}_{\bullet 1}(\tau\,|\,0,X)}{\hat{S}_{\bullet 1}(t-\,|\,0,X)} - \frac{\bar{S}_{\bullet 1}(\tau\,|\,0,X)}{\bar{S}_{\bullet 1}(t-\,|\,0,X)}\right|\right\}\right.$$
$$\left.\left\{\frac{\bar{F}_{11}(t\,|\,0,X) - \bar{F}_{11}(\tau\,|\,0,X)}{\bar{b}_1(t\,|\,X)}\right\}\Bigg|_0^{\tau\wedge\tilde{T}}\right]^2$$
$$\lesssim PI\{\pi(X) > 0\}\left\{\sup_{t\in(0,\tau]}|\hat{S}_{\bullet 1} - \bar{S}_{\bullet 1}|^2(t\,|\,0,X)\right\}^2.$$

The remaining terms $P\delta_7^2$, $P\delta_8^2$, $P\delta_9^2$, and $P\delta_{10}^2$ can be analogously bounded as $P\delta_3^2$, $P\delta_4^2$, $P\delta_5^2$, and $P\delta_6^2$, respectively. Now, for any $\mathrm{A}_1, \mathrm{A}_2 \in \mathcal{A}$ with product integrals S_1 and S_2,

$$\begin{aligned}\sup_{t\in(0,\tau]}|S_1 - S_2|(t) &= \sup_{t\in(0,\tau]}\left|\int_0^t \frac{S_2(t)}{S_2(s)}S_1(s-)\mathrm{d}(A_2 - A_1)(s)\right| \\ &\leqslant \sup_{t\in(0,\tau]}\int_0^t \mathrm{d}|A_2 - A_1|(s) \\ &\leqslant \sup_{t\in(0,\tau]}|A_2 - A_1|(t).\end{aligned}$$

Therefore, by Assumption 6, the terms $P\delta_m^2$ are all $o_P(1)$, and the lemma follows. □

Proof of Theorem 1. We first show consistency of the estimator $\hat{\theta}_1(0)$. Decompose the bias as

$$\hat{\theta}_1(0) - \theta_1(0) = (\mathbb{P}_n - P)\hat{\ell}_1(0) + P\{\hat{\ell}_1(0) - \bar{\ell}_1(0)\} + P\left\{\bar{\ell}_1(0) - \frac{\alpha}{\bar{\alpha}}\ell_1(0)\right\} - \frac{\hat{\alpha} - \alpha}{\hat{\alpha}}\theta_1(0).$$

The first term is $o_P(1)$ by the uniform law of large numbers because $\hat{\ell}_1(0)$ belongs to a P-Glivenko-Cantelli class. The second term is bounded by Jensen's inequality as $P\{\hat{\ell}_1(0) - \bar{\ell}_1(0)\} \leqslant [P\{\hat{\ell}_1(0) - \bar{\ell}_1(0)\}^2]^{1/2}$, which converges in probability to zero by Lemma S4. The third term is exactly 0 by $\bar{\alpha} = \alpha$, the assumption on the correct specifications of the nuisance estimators, and Lemma S3. The fourth term is trivially $o_P(1)$ from $\hat{\alpha} \xrightarrow{\mathrm{p}} \alpha$ and Slutsky's theorem. Therefore, $\hat{\theta}_1(0) \xrightarrow{\mathrm{p}} \theta_1(0)$.

Then we show asymptotic linearity of $\hat{\theta}_1(0)$. We decompose the bias again as

$$\hat{\theta}_1(0) - \theta_1(0) = (\mathbb{P}_n - P)\varphi_1(0) + (\mathbb{P}_n - P)\{\hat{\ell}_1(0) - \ell_1(0)\} + P\left\{\hat{\ell}_1(0) - \frac{\alpha}{\hat{\alpha}}\ell_1(0)\right\} + \frac{(\hat{\alpha} - \alpha)^2}{\hat{\alpha}\alpha}\theta_1(0).$$

The second term is $o_P(n^{-1/2})$ by Lemma 19.24 in van der Vaart (1998) because $P\{\hat{\ell}_1(0) - \ell_1(0)\}^2 \xrightarrow{\mathrm{p}} 0$ by Lemma S4. The third term is $o_P(n^{-1/2})$ by assumption and Lemma S3. The fourth term is trivially $O_P(n^{-1}) = o_P(n^{-1/2})$ from $n^{1/2}(\hat{\alpha} - \alpha) \xrightarrow{\mathrm{d}} \mathrm{Normal}\{0, \alpha(1-\alpha)\}$, $\hat{\alpha} \xrightarrow{\mathrm{p}} \alpha$, and Slutsky's theorem. Therefore, $\hat{\theta}_1(0) - \theta_1(0) = \mathbb{P}_n\varphi_1(0) + o_P(n^{-1/2})$, since $E_P\varphi_1(0) = 0$.

□

S3. Details on the simulation study

Define

$$\hat{W}_{kj}(t, s\,|\,a, x) = I(j = k)\hat{S}_1^c(s-\,|\,a, x) - \frac{\hat{F}_{1j}(t\,|\,a, x) - \hat{F}_{1j}(s\,|\,a, x)}{1 - \triangle\hat{\mathrm{A}}_{1k}(s\,|\,a, x)}, \quad (k \neq 1, a \neq 0)$$

Table S1. *Simulation results for cumulative incidences* $\theta_1(1,t)$.

n	t	Mean	Bias	RMSE	SE	Coverage
750	0.25	0.12	2.70	2.54	2.52	93.2
	1	0.22	12.19	3.41	3.43	95.2
	2	0.28	16.69	3.77	3.83	95.4
1500	0.25	0.12	−1.26	1.82	1.78	93.0
	1	0.22	−3.30	2.56	2.43	93.4
	2	0.27	−10.69	2.81	2.71	93.0

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-4}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

$$\hat{W}_{\bullet j}(t,s\,|\,0,x) = I(j=1)\hat{S}_1^c(s-\,|\,0,x) - \frac{\hat{F}_{1j}(t\,|\,0,x) - \hat{F}_{1j}(s\,|\,0,x)}{1 - \triangle\hat{\mathrm{A}}_{\bullet 1}(s\,|\,0,x)}.$$

Consider the following functions with plug-in nuisance estimators:

$$\begin{aligned}
\hat{\ell}_j(0,t)(O) &= \frac{1-A}{\hat{\alpha}}\hat{\pi}(X)\int_0^t \frac{\hat{W}_{\bullet j}(t,s\,|\,0,X)}{\hat{H}_{\bullet}(s-\,|\,X)}\mathrm{d}\hat{M}_{\bullet 1}(s\,|\,0,X) \\
&\quad + \frac{D(1-A)}{\hat{\alpha}}\int_0^t \frac{\hat{W}_{2j}(t,s\,|\,0,X)}{\hat{H}_1(s-\,|\,0,X)}\mathrm{d}\hat{M}_{12}(s\,|\,0,X) + \frac{D}{\hat{\alpha}}\hat{F}_{1j}(t\,|\,0,X), \\
\hat{\ell}_j(1,t)(O) &= \sum_{k\in\{1,2\}}\frac{DA}{\hat{\alpha}}\int_0^t \frac{\hat{W}_{kj}(t,s\,|\,1,X)}{\hat{H}_1(s-\,|\,1,X)}\mathrm{d}\hat{M}_{1k}(s\,|\,1,X) + \frac{D}{\hat{\alpha}}\hat{F}_{1j}(t\,|\,1,X), \\
\hat{\ell}_j^{\dagger}(0,t)(O) &= \sum_{k\in\{1,2\}}\frac{D(1-A)}{\hat{\alpha}}\int_0^t \frac{\hat{W}_{kj}(t,s\,|\,0,X)}{\hat{H}_1(s-\,|\,0,X)}\mathrm{d}\hat{M}_{1k}(s\,|\,0,X) + \frac{D}{\hat{\alpha}}\hat{F}_{1j}(t\,|\,0,X).
\end{aligned}$$

Let $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(O_i)$. Then we have the corresponding estimators

$$\begin{aligned}
\hat{\theta}_j(a,t) &= \mathbb{P}_n\hat{\ell}_j(a,t)(O), & \hat{\theta}_j^{\dagger}(0,t) &= \mathbb{P}_n\hat{\ell}_j^{\dagger}(0,t)(O), \\
\hat{\gamma}_j(a,t) &= \mathbb{P}_n\int_0^t \hat{\ell}_j(a,s)(O)\mathrm{d}s, & \hat{\gamma}_j^{\dagger}(a,t) &= \mathbb{P}_n\int_0^t \hat{\ell}_j^{\dagger}(a,s)(O)\mathrm{d}s,
\end{aligned}$$

where $\hat{\theta}_j^{\dagger}(0,t)$ and $\hat{\gamma}_j^{\dagger}(0,t)$ are the RCT-only estimators for the parameters $\theta_j(0,t)$ and $\gamma_j(0,t)$.

Tables S1–S4 display the summary statistics for the estimand $\theta_1(1,t)$, the set of estimands for cause 2 $\{\theta_2(1,t), \theta_2(0,t), \theta_2\{t\}\}$, the estimand $\gamma_1(1,t)$, and the set of estimands for cause 2 $\{\gamma_2(1,t), \gamma_2(0,t), \gamma_2\{t\}\}$, respectively.

S4. Details on the real data example

In both RCTs used in the data example, the rate of severe adverse events was relatively low, and the vast majority of participants were censored by the end of the study. Considering the three-point major adverse cardiovascular event (MACE, a composite event of cardiovascular death, non-fatal myocardial infarction, and non-fatal stroke) as the primary event, only 12 out of 1649 subjects randomized to placebo in SUSTAIN-6 experienced the competing non-cardiovascular death event. The crude hazard estimate of MACE is 0.37 events per 100 person-years, where 1 year counts as 365.25 days. The number of non-cardiovascular deaths was 133 out of 4672 for the placebo group in LEADER, with a corresponding hazard of 0.79 events per 100 person-year.

Table S2. *Simulation results for cumulative incidences $\theta_2(1,t)$, $\theta_2(0,t)$, and $\theta_2(t)$.*

n	Estimand	t	Type	Mean	Bias	RMSE	SE	Coverage	Reduction
750	$\theta_2(0,t)$	0.25	+	0.23	13.18	3.37	3.27	93.8	0.94
			−	0.23	11.83	3.41	3.28	93.4	·
		1	+	0.44	−2.74	3.87	3.94	95.1	5.21
			−	0.43	−6.66	3.97	4.05	95.3	·
		2	+	0.55	3.00	4.10	4.01	93.9	10.84
			−	0.54	1.31	4.33	4.25	94.3	·
	$\theta_2(1,t)$	0.25	−	0.23	−3.32	3.28	3.28	94.7	·
		1	−	0.42	−14.16	4.06	4.02	94.1	·
		2	−	0.51	−10.60	4.36	4.23	93.9	·
	$\theta_2(t)$	0.25	+	0.00	−16.49	4.55	4.54	95.1	0.50
			−	0.00	−15.15	4.58	4.55	95.1	·
		1	+	−0.02	−11.42	5.45	5.48	95.0	2.78
			−	−0.01	−7.50	5.55	5.56	94.9	·
		2	+	−0.03	−13.60	5.71	5.69	94.8	5.72
			−	−0.03	−11.91	5.94	5.86	94.5	·
1500	$\theta_2(0,t)$	0.25	+	0.23	−1.05	2.26	2.31	95.4	0.91
			−	0.23	−1.61	2.27	2.32	95.0	·
		1	+	0.44	6.17	2.81	2.80	94.1	5.13
			−	0.44	5.88	2.86	2.87	94.5	·
		2	+	0.55	1.46	2.86	2.86	94.0	10.70
			−	0.55	2.04	3.02	3.03	94.5	·
	$\theta_2(1,t)$	0.25	−	0.23	−7.69	2.30	2.32	94.7	·
		1	−	0.42	−12.76	2.84	2.84	94.4	·
		2	−	0.51	−12.98	2.97	3.01	95.1	·
	$\theta_2(t)$	0.25	+	0.00	−6.64	3.17	3.21	95.4	0.48
			−	0.00	−6.08	3.17	3.21	95.6	·
		1	+	−0.02	−18.93	3.84	3.88	95.5	2.74
			−	−0.02	−18.64	3.87	3.93	96.1	·
		2	+	−0.03	−14.44	3.92	4.05	95.7	5.65
			−	−0.03	−15.02	4.04	4.17	96.0	·

Type: fusion estimator (+) or RCT-only estimator (−); Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-4}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %; Reduction: average of percentage reduction in squared standard error estimates, %.

Table S3. *Simulation results for restricted mean times lost $\gamma_1(1,t)$.*

n	t	Mean	Bias	RMSE	SE	Coverage
750	0.25	0.02	−0.55	0.44	0.43	94.3
	1	0.15	−4.06	2.45	2.48	94.3
	2	0.40	−15.55	5.65	5.75	94.5
1500	0.25	0.02	−0.70	0.31	0.31	93.5
	1	0.15	−7.23	1.84	1.76	94.1
	2	0.40	−27.24	4.32	4.08	92.4

Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-4}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %.

Table S4. *Simulation results for restricted mean times lost* $\gamma_2(1,t)$, $\gamma_2(0,t)$, *and* $\gamma_2(t)$.

n	Estimand	t	Type	Mean	Bias	RMSE	SE	Coverage	Reduction
750	$\gamma_2(0,t)$	0.25	+	0.04	−1.11	0.58	0.57	94.6	0.50
			−	0.04	−1.32	0.59	0.57	94.2	·
		1	+	0.30	−21.05	2.98	2.96	94.9	2.85
			−	0.30	−23.22	3.04	3.00	94.7	·
		2	+	0.79	−75.58	6.32	6.32	94.7	6.21
			−	0.79	−80.45	6.54	6.53	94.2	·
	$\gamma_2(1,t)$	0.25	−	0.04	−2.76	0.59	0.57	93.5	·
		1	−	0.29	−33.89	3.13	3.01	93.8	·
		2	−	0.76	−94.02	6.93	6.58	92.9	·
	$\gamma_2(t)$	0.25	+	0.00	−1.65	0.81	0.80	94.1	0.26
			−	0.00	−1.44	0.81	0.80	94.3	·
		1	+	−0.01	−12.84	4.16	4.09	93.7	1.53
			−	−0.00	−10.66	4.21	4.12	94.0	·
		2	+	−0.03	−18.44	8.87	8.82	94.7	3.31
			−	−0.03	−13.57	9.09	8.97	94.4	·
1500	$\gamma_2(0,t)$	0.25	+	0.04	−0.98	0.40	0.40	94.1	0.48
			−	0.04	−1.06	0.41	0.41	94.2	·
		1	+	0.30	−12.13	2.09	2.10	94.8	2.80
			−	0.30	−12.59	2.12	2.13	94.3	·
		2	+	0.79	−36.12	4.53	4.50	94.0	6.18
			−	0.79	−36.57	4.65	4.65	94.3	·
	$\gamma_2(1,t)$	0.25	−	0.04	−1.95	0.41	0.41	93.7	·
		1	−	0.29	−19.07	2.14	2.14	94.7	·
		2	−	0.76	−52.09	4.72	4.68	94.5	·
	$\gamma_2(t)$	0.25	+	0.00	−0.96	0.56	0.56	94.7	0.25
			−	0.00	−0.89	0.56	0.57	94.7	·
		1	+	−0.00	−6.94	2.88	2.91	95.9	1.49
			−	−0.00	−6.47	2.89	2.93	95.9	·
		2	+	−0.03	−15.97	6.17	6.28	95.8	3.28
			−	−0.03	−15.51	6.25	6.38	96.0	·

Type: fusion estimator (+) or RCT-only estimator (−); Mean: average of estimates; Bias: Monte-Carlo bias, 10^{-4}; RMSE: root mean squared error, 10^{-2}; SE: average of standard error estimates, 10^{-2}; Coverage: 95% confidence interval coverage, %; Reduction: average of percentage reduction in squared standard error estimates, %.

Table S5. *Subjects with missing baseline covariates in SUSTAIN-6 and LEADER.*

	SUSTAIN-6		LEADER
	Semaglutide 1.0 mg	Placebo	Placebo
N	822	1649	4672
Missing (N)	9	24	96
Missing (%)	1.09	1.46	2.05

While MACE was the primary outcome in the original analyses of both studies, we turned to the composite event of non-fatal myocardial infarction and non-fatal stroke as the event of interest. The main reason is precisely that a greater number of competing events would allow us to better illustrate our method.

The SUSTAIN-6 trial followed participants for a maximum of 104 weeks since randomization with an end-of-trial visit at week 109. The LEADER trial, on the other hand, planned a much longer follow-up period of up to 54 months. Therefore, without assuming transportability of the conditional cause-specific hazards of both non-fatal cardiovascular outcome and all-cause death, none of the parameters considered would be identifiable beyond week 104. In the data example, we chose to estimate parameters at 4 evenly spaced time points up to week 104.

For transportability of the cause-specific hazard of the composite event under placebo, we needed to control for the baseline covariates that are shifted prognostic variables between the RCT population and the external control population. In the data example, we employed the list of baseline characteristics in Table 1 of Marso et al. (2016a). All cause-specific hazards were fitted by the Cox proportional hazards model with a linear combination of the baseline covariates as the logarithm of the multiplicative risk. The concentration of low-density lipoprotein cholesterol was measured in mmol $\cdot$ l^{-1} and subsequently log-transformed to reduce skewness. History of hemorrhagic stroke was removed from the list, as its presence caused extreme numeric instability during the fitting of the Cox model. Patients with missing baseline covariates were removed from the data. Exact numbers of missing subjects per treatment groups are displayed in Table S5.

The data included 11 tied event times between times to non-fatal cardiovascular event in the joint placebo arm and times to all-cause death in the placebo arm of SUSTAIN-6. Since our estimator is presented under Assumption 5, we broke the ties by jittering the observed event times. Specifically, a random sample of noise was drawn from the uniform distribution between 0 and 10^{-5} and then added to the observed event times. The event times used in the analysis were recorded in days as integers. Therefore, such small perturbations should not have meaningful consequences for the results.

To stabilize the estimators, we set a threshold for inverse weights inside the integrals $\hat{\ell}_j(a)(O)$. For sample size n, the inverse weights above the value $n^{1/2}\log(n)/5$ were set to that value.

S5. Implications of weaker transportability assumptions

In the main text, we have showcased how the transportability of a conditional cause-specific hazard improves the precision of estimators for cumulative incidence functions and restricted mean times lost.

One consideration is whether this assumption can be reasonably weakened according to the parameter of interest. To ground ideas, consider the target parameter $\theta_1(0) = E\{F_{11}(\tau \mid 0, X) \mid D = 1\}$. If we view the parameter as the mean of a binary outcome $E\{I\{T(0) \leqslant \tau, J(0) = 1\} \mid D = 1\}$, a straightforward transportability assumption would be

$$\text{pr}\{T(0) \leqslant \tau, J(0) = 1 \mid X = x, D = 1\} = \text{pr}\{T(0) \leqslant \tau, J(0) = 1 \mid X = x, D = 0\},$$

Table S6. *Treatment-specific cumulative incidences in the real data example.*

Estimand	t (weeks)	Type	Estimate (%)	95%-CI (%)	Reduction
$\theta_1(0,t)$	26	+	1.84	(1.31, 2.36)	23.17
		−	1.97	(1.29, 2.65)	.
	52	+	3.10	(2.43, 3.77)	23.20
		−	3.21	(2.34, 4.09)	.
	78	+	4.87	(4.04, 5.70)	21.30
		−	4.66	(3.60, 5.72)	.
	104	+	6.42	(5.46, 7.38)	21.78
		−	6.25	(5.02, 7.48)	.
$\theta_2(0,t)$	26	+	0.74	(0.32, 1.17)	−0.00
		−	0.74	(0.32, 1.17)	.
	52	+	1.23	(0.69, 1.78)	−0.00
		−	1.23	(0.69, 1.78)	.
	78	+	1.85	(1.19, 2.52)	−0.00
		−	1.85	(1.19, 2.52)	.
	104	+	2.92	(2.08, 3.76)	−0.00
		−	2.92	(2.08, 3.76)	.
$\theta_1(1,t)$	26	.	1.58	(0.72, 2.44)	.
	52	.	2.49	(1.42, 3.55)	.
	78	.	2.88	(1.73, 4.03)	.
	104	.	3.69	(2.40, 4.98)	.
$\theta_2(1,t)$	26	.	0.25	(−0.12, 0.62)	.
	52	.	0.85	(0.17, 1.54)	.
	78	.	1.90	(0.88, 2.93)	.
	104	.	2.71	(1.49, 3.93)	.

Type: fusion estimator (+) or RCT-only estimator (−); CI: confidence interval; Reduction: percentage reduction CI length, %.

Table S7. *Treatment-specific restricted mean times lost in the real data example.*

Estimand	t (weeks)	Type	Estimate (weeks)	95%-CI (weeks)	Reduction
$\gamma_1(0,t)$	26	+	0.27	(0.18, 0.35)	20.83
		−	0.27	(0.17, 0.38)	.
	52	+	0.94	(0.71, 1.17)	22.17
		−	0.98	(0.68, 1.27)	.
	78	+	2.03	(1.63, 2.43)	22.43
		−	2.05	(1.54, 2.57)	.
	104	+	3.50	(2.90, 4.11)	22.35
		−	3.49	(2.71, 4.27)	.
$\gamma_2(0,t)$	26	+	0.08	(0.03, 0.14)	−0.00
		−	0.08	(0.03, 0.14)	.
	52	+	0.33	(0.16, 0.49)	−0.00
		−	0.33	(0.16, 0.49)	.
	78	+	0.73	(0.42, 1.03)	−0.00
		−	0.73	(0.42, 1.03)	.
	104	+	1.36	(0.90, 1.83)	−0.00
		−	1.36	(0.90, 1.83)	.
$\gamma_1(1,t)$	26	.	0.16	(0.06, 0.26)	.
	52	.	0.72	(0.39, 1.05)	.
	78	.	1.44	(0.83, 2.04)	.
	104	.	2.35	(1.45, 3.25)	.
$\gamma_2(1,t)$	26	.	0.03	(−0.03, 0.10)	.
	52	.	0.21	(0.02, 0.39)	.
	78	.	0.52	(0.15, 0.89)	.
	104	.	1.12	(0.51, 1.73)	.

Type: fusion estimator (+) or RCT-only estimator (−); CI: confidence interval; Reduction: percentage reduction CI length, %.

Table S8. *Cumulative incidence differences after removing history of cardiovascular diseases from the baseline variables.*

Estimand	t (weeks)	Type	Estimate (%)	95%-CI (%)	Reduction
$\theta_1(t)$	26	+	−0.00	(−0.88, 0.87)	18.11
		−	−0.56	(−1.63, 0.51)	.
	52	+	−0.66	(−1.79, 0.47)	16.79
		−	−0.96	(−2.32, 0.39)	.
	78	+	−1.85	(−3.10, −0.60)	18.39
		−	−2.03	(−3.56, −0.49)	.
	104	+	−2.44	(−3.86, −1.03)	19.02
		−	−2.88	(−4.62, −1.14)	.
$\theta_2(t)$	26	+	−0.53	(−1.06, 0.00)	0.00
		−	−0.52	(−1.06, 0.01)	.
	52	+	−0.40	(−1.25, 0.45)	−0.00
		−	−0.40	(−1.25, 0.45)	.
	78	+	0.07	(−1.13, 1.26)	−0.00
		−	0.07	(−1.13, 1.27)	.
	104	+	−0.18	(−1.64, 1.28)	−0.00
		−	−0.17	(−1.63, 1.29)	.

Type: fusion estimator (+) or RCT-only estimator (−); CI: confidence interval; Reduction: percentage reduction CI length, %.

Table S9. *Restricted mean time lost differences after removing history of cardiovascular diseases from the baseline variables.*

Estimand	t (weeks)	Type	Estimate (weeks)	95%-CI (weeks)	Reduction
$\gamma_1(t)$	26	+	−0.05	(−0.16, 0.06)	24.42
		−	−0.12	(−0.27, 0.02)	.
	52	+	−0.13	(−0.48, 0.22)	19.66
		−	−0.33	(−0.76, 0.10)	.
	78	+	−0.48	(−1.11, 0.16)	18.37
		−	−0.75	(−1.53, 0.03)	.
	104	+	−0.97	(−1.93, −0.02)	18.10
		−	−1.36	(−2.53, −0.19)	.
$\gamma_2(t)$	26	+	−0.05	(−0.13, 0.02)	0.00
		−	−0.05	(−0.13, 0.02)	.
	52	+	−0.13	(−0.37, 0.11)	−0.00
		−	−0.13	(−0.37, 0.11)	.
	78	+	−0.22	(−0.67, 0.24)	−0.00
		−	−0.21	(−0.67, 0.25)	.
	104	+	−0.24	(−0.99, 0.51)	−0.00
		−	−0.24	(−0.99, 0.51)	.

Type: fusion estimator (+) or RCT-only estimator (−); CI: confidence interval; Reduction: percentage reduction CI length, %.

Table S10. *Cumulative incidence differences after removing controls from SUSTAIN-6.*

Estimand	t (weeks)	Type	Estimate (%)	95%-CI (%)	Reduction
$\theta_1(t)$	26	+	−0.21	(−1.16, 0.73)	48.70
		−	−1.22	(−3.07, 0.62)	.
	52	+	−0.97	(−2.20, 0.25)	45.48
		−	−1.74	(−3.99, 0.50)	.
	78	+	−2.35	(−3.73, −0.96)	46.55
		−	−2.74	(−5.32, −0.15)	.
	104	+	−3.10	(−4.67, −1.52)	48.35
		−	−4.16	(−7.21, −1.12)	.
$\theta_2(t)$	26	+	−1.03	(−2.22, 0.15)	0.01
		−	−1.03	(−2.21, 0.16)	.
	52	+	−1.17	(−2.75, 0.41)	0.01
		−	−1.16	(−2.74, 0.42)	.
	78	+	−1.41	(−3.52, 0.70)	0.01
		−	−1.39	(−3.50, 0.72)	.
	104	+	−1.54	(−3.99, 0.92)	0.01
		−	−1.51	(−3.97, 0.95)	.

Type: fusion estimator (+) or RCT-only estimator (−); CI: confidence interval; Reduction: percentage reduction CI length, %.

Table S11. *Restricted mean time lost differences after removing controls from SUSTAIN-6.*

Estimand	t (weeks)	Type	Estimate (weeks)	95%-CI (weeks)	Reduction
$\gamma_1(t)$	26	+	−0.10	(−0.23, 0.03)	59.69
		−	−0.29	(−0.60, 0.03)	.
	52	+	−0.27	(−0.65, 0.12)	51.75
		−	−0.64	(−1.44, 0.15)	.
	78	+	−0.71	(−1.41, −0.01)	49.08
		−	−1.28	(−2.65, 0.09)	.
	104	+	−1.36	(−2.41, −0.31)	47.99
		−	−2.10	(−4.11, −0.08)	.
$\gamma_2(t)$	26	+	−0.16	(−0.34, 0.02)	0.01
		−	−0.16	(−0.34, 0.02)	.
	52	+	−0.41	(−0.93, 0.10)	0.01
		−	−0.41	(−0.92, 0.10)	.
	78	+	−0.78	(−1.70, 0.14)	0.01
		−	−0.77	(−1.69, 0.15)	.
	104	+	−1.24	(−2.68, 0.19)	0.01
		−	−1.23	(−2.66, 0.20)	.

Type: fusion estimator (+) or RCT-only estimator (−); CI: confidence interval; Reduction: percentage reduction CI length, %.

which is

$$\int_0^\tau S_1(0)(t- \mid x)\mathrm{dA}_{11}(0)(t \mid x)\mathrm{d}t = \int_0^\tau S_0(0)(t- \mid x)\mathrm{dA}_{01}(0)(t \mid x)\mathrm{d}t$$

for $x \in \mathcal{X}_1 \cap \mathcal{X}_0$, where $S_d(0)(t \mid x) = [\Pi\{\mathrm{A}_{d1}(0) + \mathrm{A}_{d2}(0)\}](t \mid x)$.

There are two peculiarities to point out. The first is whether it makes sense at all to only restrict the value of conditional cumulative incidence function of cause 1 at the time point τ. It is very unnatural to only assume transportability for a single time point. If this assumption holds, we should also expect the cumulative incidence functions in the time interval around that time point to be quite comparable across populations, especially when the event time distribution is continuous. Moreover, we would not generally expect a substantial decrease in the semiparametric efficiency bound of the parameter, if the compatibility of the two population exists for a mere single time point on a specific scale defined by the parameter.

The second is a result of the cumulative incidence function of cause 1 being a functional of the cause-specific hazards of both event types. Therefore, by making this assumption, we are also putting restrictions on the cause 2 hazards between the two populations. However, reasoning for comparability of the cumulative incidence functions is arguably more difficult than doing so separately for the two event rates. Note that this observation also applies to transportability assumptions on subdistribution hazards (Fine and Gray, 1999), for example, for $t \in (0, \tau]$,

$$\mathrm{pr}\{T(0) \leqslant t, J(0) = 1 \mid X = x, D = 1\} = \mathrm{pr}\{T(0) \leqslant t, J(0) = 1 \mid X = x, D = 0\}.$$

Apart from the transportability of the cumulative incidence function, we may also consider the transportability of the all-cause survival function that $S_1(0)(t \mid x) = S_0(0)(t \mid x)$. However, that the sum of two cause-specific hazards is equal across the populations can result from many combinations of event rates whose interpretations are drastically different. For instance, this assumption holds if the cause 1 hazard under placebo in the RCT population equals the cause 2 hazard in the external control population, while their competing risks are completely eliminated. Since the estimands used in competing risks analysis often seek to separate the treatment effects on different causes, a transportability assumption that does not acknowledge the nature of competing risks may be hard to justify.

References

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley Series in Probability and Statistics. Wiley.

Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18(4):1501–1555.

Marso, S. P., Bain, S. C., Consoli, A., Eliaschewitz, F. G., Jódar, E., Leiter, L. A., Lingvay, I., Rosenstock, J., Seufert, J., Warren, M. L., Woo, V., Hansen, O., Holst, A. G., Pettersson, J., and Vilsbøll, T. (2016). Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine*, 375(19):1834–1844.

Rytgaard, H. C. W., Eriksson, F., and van der Laan, M. J. (2023). Estimation of time-specific intervention effects on continuously distributed time-to-event outcomes by targeted maximum likelihood estimation. *Biometrics*, 79(4):3038–3049.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.