Safer Causal Inference

Theory & Algorithms for Falsification, Trial Augmentation and Policy Evaluation

Rickard Karl Axel Karlsson

Safer Causal Inference

Theory & Algorithms for Falsification, Trial Augmentation and Policy Evaluation

SAFER CAUSAL INFERENCE

THEORY & ALGORITHMS FOR FALSIFICATION, TRIAL AUGMENTATION AND POLICY EVALUATION

Dissertation

for the purpose of obtaining the degree of doctor

at Delft University of Technology

by the authority of the Rector Magnificus, Prof. dr. ir. H. Bijl,

chair of the Board for Doctorates

to be defended publicly on

Friday 24 April 2026 at 10:00

by

Rickard Karl Axel KARLSSON

This dissertation has been approved by the (co)promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology, *promotor*
Dr. ir. J.H. Krijthe	Delft University of Technology, *copromotor*
Independent members:	
Prof. dr. A.W. van der Vaart	Delft University of Technology
Prof. dr. M.M. de Weerdt	Delft University of Technology
Prof. dr. P. Grünwald	Leiden University & CWI
Prof. dr. R. Evans	University of Oxford, United Kingdom
Dr. J. Josse	Inria Saclay, France

Keywords: causal inference, machine learning, treatment effect estimation, falsification, trial augmentation, policy evaluation

Printed by: ProefschriftMaken.nl

Cover by: R.K.A. Karlsson

ISBN 978-94-6534-273-3

An electronic copy of this dissertation is available at https://repository.tudelft.nl/.

Contents

Summary

Estimating the effect of an intervention on an outcome is a central challenge across science and society. In medicine, we may ask whether a drug effectively treats a disease, and in economics, whether a new policy reduces unemployment. Estimating such effects from data, a process known as causal inference, is essential but inherently difficult because it often relies on untestable assumptions to ensure unbiased identification of treatment effects. A key example of such an untestable assumption is the absence of unmeasured confounding, meaning that no hidden variable influences both the treatment and the outcome. When this assumption fails, something which we cannot directly verify, treatment effect estimates may become biased. This ultimately can lead to untrustworthy conclusions and, in the worst case, unsafe decisions, such as prescribing the wrong drug to a patient. The central question of this dissertation is therefore whether we can develop methods for safer causal inference that either detect violations of its underlying assumptions or remain robust when those assumptions are violated.

In Part One, we address the first aspect of detecting violations of causal identification assumptions. We focus on settings with data from multiple sources, such as hospitals or locations, where distributional shifts naturally occur. Under specific independence conditions on the causal mechanisms driving these shifts, we first present a nonparametric test to falsify the assumption of no unmeasured confounding. To obtain these results, we introduce a novel technique utilizing hierarchical causal graphical models. Thereafter, we focus on improving the statistical efficiency of this test, which is achieved by reformulating the independence condition using parameterized linear models. Finally, we extend the hierarchical modeling approach to other identification settings, specifically by testing the validity of mediators and instrumental variables used in two additional common identification strategies.

In Parts Two and Three, we develop methods that instead are robust when causal identification assumptions are violated. We revisit two commonly occurring problem settings when doing causal inference and demonstrate that it is possible to develop methods that either remove the need for, or rely on, weaker and more plausible assumptions than those traditionally made. In the first setting, we study the problem of augmenting randomized trials using external data to improve efficiency in treatment effect estimation. Typically, such approaches rely on a transportability assumption that relate the populations underlying the trial and external data. But when this transportability assumption is violated, integrating external data can introduce substantial bias. To address this, we propose a novel and efficient estimator that incorporates external data and show that this estimator improves inference on the average treatment effect while guaranteeing that it never per-

forms worse, and sometimes performs better, than the estimator that relies solely on trial data. We further adapt this estimator to learn heterogeneous treatment effects within the trial population and show that similar safety guarantees hold for this problem.

In the second setting, we examine the evaluation of treatment allocation strategies using Qini curves. Standard methods for estimating Qini curves assume no interference between treated units, meaning that the treatment of one unit does not affect others. However, when interference is present, these Qini curves can be misleading and lead to incorrect evaluation of treatment allocation strategies. We therefore propose multiple estimators to handle the interference, specifically in settings where units within a cluster may affect one another but not units in other clusters. We identify a bias-variance trade-off in these estimators and, through both theoretical and empirical results, provide practical guidance on how practitioners can choose among them.

The dissertation concludes with a discussion of broader considerations, limitations of the presented research, and potential directions for future work. We find that it is indeed possible to make causal inference safer by detecting assumption violations and reducing reliance on untestable assumptions. Nonetheless, many open and important questions remain, offering promising avenues for further research on this topic.

Samenvatting

Het schatten van het effect van een interventie op een uitkomst is een centrale uitdaging binnen zowel de wetenschap als de maatschappij. In de geneeskunde kunnen we ons afvragen of een geneesmiddel een ziekte effectief behandelt, en in de economie of een nieuw beleid de werkloosheid vermindert. Het schatten van dergelijke effecten op basis van data, een proces dat bekendstaat als causale inferentie, is essentieel, maar ook inherent moeilijk, omdat het vaak steunt op niet-toetsbare aannames om de zuivere identificatie van behandelingseffecten te waarborgen. Een belangrijk voorbeeld van zo'n niet-toetsbare aanname is de afwezigheid van ongemeten confounders, wat betekent dat er geen verborgen variabelen zijn die zowel de behandeling als de uitkomst beïnvloeden. Wanneer deze aanname niet geldt, iets wat men niet direct kan verifiëren, kunnen schattingen van behandelingseffecten vertekend raken. Dit kan uiteindelijk leiden tot onbetrouwbare conclusies en, in het ergste geval, tot onveilige beslissingen, zoals het voorschrijven van het verkeerde geneesmiddel aan een patiënt. De centrale vraag van dit proefschrift is daarom of we methoden kunnen ontwikkelen voor veilige causale inferentie die schendingen van de onderliggende aannames kunnen detecteren, of robuust blijven wanneer die aannames niet geldig zijn.

In Deel Een onderzoeken we hoe schendingen van identificatie-aannames kunnen worden gedetecteerd. We richten ons op situaties met data afkomstig uit meerdere bronnen, zoals verschillende ziekenhuizen of locaties, waar distributieverschuivingen van nature voorkomen. Onder specifieke onafhankelijkheidsvoorwaarden voor de causale mechanismen die deze verschuivingen veroorzaken, presenteren we eerst een non-parametrische toets om de aanname van geen ongemeten confounders te falsifiëren. Om deze resultaten te verkrijgen, introduceren we een nieuwe techniek die gebruikmaakt van hiërarchische causale graafmodellen. Vervolgens richten we ons op het verbeteren van de statistische efficiëntie van deze toets, wat we bereiken door de onafhankelijkheidsvoorwaarde te herformuleren met behulp van geparametriseerde lineaire modellen. Ten slotte breiden we de hiërarchische modelleringsbenadering uit naar andere identificatie-situaties, door de validiteit te testen van mediatoren en instrumentele variabelen die worden gebruikt in twee andere veelvoorkomende identificatiestrategieën.

In Deel Twee en Deel Drie ontwikkelen we methoden die juist robuust zijn wanneer identificatie-aannames niet gelden. We bekijen twee vaak voorkomende problemen binnen causale inferentie en tonen aan dat het mogelijk is methoden te ontwikkelen die de noodzaak van zulke aannames geheel opheffen, of die gebaseerd zijn op zwakkere en aannemelijkere aannames dan traditioneel wordt gedaan. Eerst bestuderen we het probleem van het aanvullen van gerandomiseerde studies met externe data om de effici-

ëntie van de schatting van behandelingseffecten te verbeteren. Dergelijke benaderingen steunen doorgaans op een transporteerbaarheidsaanname die de populaties van de gerandomiseerde studie en de externe data met elkaar relateert. Wanneer deze aanname wordt geschonden, kan het integreren van externe data echter de zuiverheid van de schatting aanzienlijk verminderen. Om dit te verhelpen, stellen wij een nieuwe en efficiënte schatter voor die externe data incorporeert. We tonen aan dat deze schatter de schatting van het gemiddelde behandelingseffect verbetert, met de garantie dat hij nooit slechter presteert, en soms beter, dan de schatter die enkel op data van de gerandomiseerde studie gebaseerd is. We passen deze schatter verder aan om heterogene behandelingseffecten binnen de data van de gerandomiseerde studie te leren, en tonen aan dat vergelijkbare veiligheidsgaranties ook hier gelden.

Vervolgens onderzoeken we de evaluatie van behandelstrategieën met behulp van Qini-curves. Standaardmethoden voor het schatten van Qini-curves gaan uit van de afwezigheid van interferentie tussen behandelde eenheden, wat betekent dat de behandeling van de ene eenheid geen effect heeft op andere eenheden. Wanneer er echter wél interferentie optreedt, kunnen deze Qini-curves misleidend zijn en leiden tot onjuiste evaluaties van behandelstrategieën. Daarom stellen we meerdere schatters voor die met interferentie kunnen omgaan, specifiek voor toepasssingen waarbij eenheden binnen een cluster elkaar kunnen beïnvloeden, maar geen invloed hebben op eenheden in andere clusters. We identificeren een bias-variantie-afweging tussen deze schatters en bieden, op basis van zowel theoretische als empirische resultaten, praktische richtlijnen voor hoe onderzoekers de meest geschikte schatter kunnen kiezen.

Het proefschrift sluit af met een bespreking van bredere overwegingen, beperkingen van het gepresenteerde onderzoek en mogelijke richtingen voor toekomstig werk. We concluderen dat het inderdaad mogelijk is causale inferentie veiliger te maken door schendingen van aannames te detecteren en de afhankelijkheid van niet-toetsbare aannames te verminderen. Toch blijven er vele open en belangrijke vragen bestaan, die veelbelovende kansen bieden voor verder onderzoek op dit gebied.

1

Introduction

Causal reasoning lies at the heart of our society. Many of the questions we care about are framed in terms of cause and effect: Will I be healthier if I take this drug? Will unemployment rates go down if we introduce this policy? Progress in fields like medicine and economics depends on answering causal questions like these. But even when these questions may be straightforward to ask, answering them is not. Fortunately, there is a scientific process for learning the causal relationships between the actions we take and the outcomes we observe, allowing us to answer these kinds of questions. This process of learning causal relationships is what we will refer to as causal inference.

Throughout history, experiments have played a central role in discovering and understanding cause-and-effect relationships. In the 16th and 17th century, Galileo and Newton conducted experiments that explored the nature of gravity and light, advancing our understanding of the physical world (Galilei, 1638; Newton, 1672). A couple of decades later, James Lind carried out one of the earliest controlled experiments, demonstrating that citrus fruits could prevent scurvy among naval sailors (Lind, 1772). Today, randomized controlled experiments are a cornerstone of evidence-based medicine, with new drugs undergoing multiple phases of clinical trials to determine whether they can be causally linked to both safety and efficacy (Devereaux & Yusuf, 2003). Experiments, therefore, not only help us answer scientific questions about cause and effect but ultimately also have been critical for scientific progress in the past.

Many important scientific questions about cause and effect cannot however be easily answered through experiments, often due to ethical or practical constraints. In such cases, we may still wish to perform causal inference even when conducting an experiment is impossible. In the 1950s, for example, epidemiologists observed a strong correlation between smoking and lung cancer, sparking debate over whether smoking actually causes the disease (Wynder, 1997). A randomized controlled experiment, which could be done by randomly assigning people to start smoking or not, would have been unethical given the suspicion that smoking is harmful. And even if such an experiment were ethically acceptable, the possible time frame between the exposure from smoking and lung cancer onset made it logistically impractical to run such an experiment. Despite these challenges,

scientists still had to confront a causal question: does smoking lead to an increased risk of lung cancer?

A central reason the debate persisted, despite the strong observed correlation, was the concern that unmeasured confounders were present which both increased the likelihood of smoking and the risk of lung cancer, thereby explaining away the observed correlation. For instance, the well-known statistician Ronald Fisher, himself a smoker, argued that a genetic factor could confound the relationship and account for the observed correlation between smoking and cancer risk (Fisher, 1958). Since the correlation was oberved in observational non-experimental data rather than a randomized experiment, it was difficult to rule out the presence of such unmeasured confounders. However, Cornfield *et al.* (1959) presented a decisive counterargument: for an unmeasured confounder to fully explain the correlation in the data, it would need to be associated with lung cancer at least ten times more strongly than smoking itself. Most experts considered it implausible that any unknown factor, including genetic ones, could have such a strong effect. This reasoning led many to conclude that a causal link with smoking was the most reasonable explanation (Loeb *et al.*, 1984; Wynder, 1997).

While the above example shows that reasoning about cause and effect is far from easy, especially when experiments are impossible to run, it also highlights how the field of causal inference has advanced. Today, we have a wide range of techniques to answer causal questions even without explicitly running experiments, for instance whether raising the minimum wage reduces unemployment (Card & Krueger, 1994) or postmenopausal hormone replacement therapies are safe for women (Hernán *et al.*, 2008). And while foundational ideas in causal inference date back to the early 20th century (Neyman, 1923; Wright, 1934), interest in the field has grown substantially in recent decades, with a wide range of disciplines contributing to its development. This is evidenced by textbooks targeting areas from public health (Hernan & Robins, 2023) and philosophy (Spirtes *et al.*, 2000) to the social sciences (Imbens & Rubin, 2015) and computer science (Pearl, 2009).

1.1. Moving towards safer causal inference

Despite all methodological advancements, one of the biggest remaining challenges in causal inference is its reliance on assumptions that are often untestable. As illustrated by the smoking example from the 1950s where they only had non-experimental data, the presence of unmeasured confounders can almost never be ruled out with certainty. Yet, if we want to confidently estimate a causal effect from such non-experimental data, we typically have to assume that no such confounders exist. While the issue with untestable assumptions is more prominent when working with non-experimental data, it can also arise in the context of experimental data. For instance, in both experimental and non-experimental data we need the stable unit treatment value assumption, which requires that an individual is unaffected by another individual's treatment. This assumption can be violated in vaccine studies, where immunity among vaccinated individuals may indirectly protect those who remain unvaccinated (VanderWeele *et al.*, 2012). If the assumptions we

make are violated, the techniques we use to perform causal inference may provide biased results, which can lead to wrong causal claims. This underscores the need for methods in causal inference that can account for the uncertainty inherent in making untestable assumptions and ultimately produce more trustworthy conclusions.

This thesis addresses this fundamental issue by proposing novel methodological tools for safer causal inference, starting from the recognition that key assumptions necessary for valid causal conclusions may be uncertain. We aim to design methods that either (i) allow us to detect when our assumptions are violated, or (ii) are robust to them being violated. Our work spans a range of practical problem settings commonly faced by scientific researchers, integrating ideas from both the traditional statistical literature on causal inference and recent developments in computer science, particularly machine learning.

To explore the challenges in improving causal inference, we will begin by introducing background on its methodology and theory, and identify several ways in which we can improve the reliability of these methods.

1.2. A brief primer on causal inference

In this and subsequent sections, we provide a brief introduction to key concepts in causal inference to explain how this thesis contributes towards safer causal inference. The mathematical level is kept to a minimum to prioritize intuition, even if that means sacrificing some precision. Mathematical rigor will be regained in the subsequent chapters.

One formal way to define causal effects is through the framework of counterfactual potential outcomes (Neyman, 1923; Rubin, 1974). Another common approach is the use of structural causal models (Pearl, 2009). In this thesis, we primarily adopt the former framework; however, the core ideas explored here are applicable in both, and the two can, to varying degrees, be translated into one another. We therefore will focus mainly on describing the potential outcomes framework here. Following standard conventions in causal inference, we refer to causal effects as treatment effects, even when the interventions being studied are not restricted to medical treatments in the traditional sense.

Starting with an individual, let A be a binary variable representing, for example, which of two clinical treatments the individual receives. Further, let Y be an outcome of interest observed after the treatment. In this example, we assume Y is a continuous health score ranging from 0 to 100, where higher values correspond to better health. The potential outcome Y^a represents the outcome we would observe for this individual if, possibly contrary to fact, we intervene by setting the treatment A to a. The individual treatment effect is then defined as the difference between the potential outcomes, $Y^1 - Y^0$. However, we can never observe both potential outcomes Y^1 and Y^0 for the same individual; only one is observed, depending on the actual treatment received. For this reason can the individual treatment effect never be observed, regardless of what data we collect, an issue known as the fundamental problem of causal inference (Holland, 1986). Despite this, other measures of treatment effects can luckily still be estimated if the data we collect

fulfill certain conditions. Some of these measures are, for instance, the average treatment effect and the risk ratio:

$$\text{Average Treatment Effect (ATE):} \quad \mathbb{E}[Y^1 - Y^0] \tag{1.1}$$

$$\text{Risk Ratio (RR):} \quad \frac{\mathbb{E}[Y^1]}{\mathbb{E}[Y^0]} \tag{1.2}$$

where the expectations are taken over a population of individuals from which we assume observations have been sampled. Another common effect measure is the conditional average treatment effect, which accounts for a set of pre-treatment covariates X such as demographic variables or contextual factors, including age and other individual-level attributes. This helps understand how the average treatment effect varies across different covariate levels, formalized as follows:

$$\text{Conditional Average Treatment Effect (CATE):} \quad \mathbb{E}[Y^1 - Y^0 \mid X]\,. \tag{1.3}$$

The choice of effect measure depends on the causal question of interest. For example, if a clinician has experimental data where patients were randomized to receive a drug or not, as shown in Figure 1.1, and wants to compare the relative improvement in health between the two groups, a risk ratio is appropriate. Suppose the average health score is 48.3 in the control group and 54.7 in the treated group. The risk ratio would then indicate that patients who take the drug improve by about 13% in health compared to those who do not. In contrast, if the focus is on the absolute change in health, the average treatment effect is more suitable, showing in this case that the drug increases the health score by 6.4 points. Furthermore, if the goal is to understand how patients respond to treatment depending on a characteristic captured by some other covariate, the clinician may instead want to estimate the conditional average treatment effect.

An important distinction in causal inference is whether the data come from a randomized (experimental) or observational (non-experimental) setting. In observational data, we often do not know a priori the mechanism that determines who receives a treatment. This may lead to the presence of confounders between the treatment and outcome. Confounders are factors that influence both the treatment and outcome, and they can make it particularly hard in observational studies to determine whether an observed association between treatment and outcome is due to the treatment itself or the influence of these factors. This does not mean randomized experiments are always better than observational studies. They come with their own challenges, for instance, strict inclusion criteria can make it harder to generalize results from the study population to a wider population. In the next section, we will formalize the conditions under which treatment effects can be estimated, whether using randomized or observational data.

1.2.1. Identification of treatment effects

The challenge when estimating treatment effects, regardless of the effect measures used, is that the treatment effects are defined in terms of the counterfactual potential outcomes,

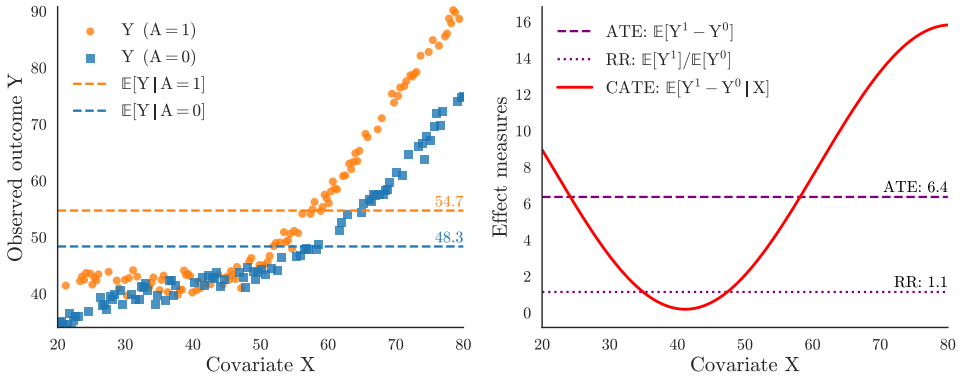


Figure 1.1: Plot of an example dataset from an experiment illustrating different effect measures a practitioner may wish to estimate. The left panel shows the data and quantities observed by the practitioner, while the right panel shows different effect measures of potential interest.

which we never directly observe. As a result, doing causal inference requires translating these counterfactual quantities into estimable quantities based on observed data. Concretely, we aim to translate quantities involving the potential outcomes Y^1 and Y^0, such as the average treatment effect in eq. (1.1), into a quantity expressed only in terms of the observed variables, in this case the covariates X, treatment A, and outcome Y. Once we have done this translation, we can estimate the treatment effect directly from actual data containing these observed variables. This translation process is often called identification, and it is a critical step in any causal analysis. A clear example of identification will be provided later. Whether identification is possible depends first on whether the underlying process from which our data was obtained meets certain conditions.

There are many popular strategies for performing identification in causal inference. Here, we focus on one of the simplest and most widely used strategies, which relies on data that satisfy three key assumptions: First, the stable unit treatment value assumption must hold, which states that the treatment and outcome of one individual does not affect any other, and that each treatment given to an individual has a well-defined and consistent meaning. Second, we assume there are no unmeasured confounders. All variables influencing both treatment assignment and the outcome are accounted for in the covariates which we already have observed. Third, we assume positivity of treatment assignment, meaning that for every combination of covariates, there is a non-zero probability of receiving each treatment option, ensuring that the treatment groups are sufficiently similar in terms of their covariate characteristics.

When data comes from a randomized experiment, the above assumptions are typically well-supported and are expected to hold. For observational data, on the other hand, these assumptions may be controversial and much domain expertise is often needed to motivate their validity. Importantly, these assumptions are in general untestable from

data, hence why we call them assumptions. An exception is the positivity assumption which is, in principle, testable though verifying it can be challenging in practice (D'Amour *et al.*, 2021).

If the above identification assumptions are met, we can express the mean potential outcome $\mathbb{E}[Y^a]$ for $a \in \{0,1\}$ using two possible identification formulas:

$$\mathbb{E}[Y^a] = \mathbb{E}\left[\frac{\mathbf{1}_{A=a}Y}{\Pr(A=a \mid X)}\right] \tag{1.4}$$

$$\mathbb{E}[Y^a] = \mathbb{E}\left[\mathbb{E}[Y \mid X, A=a]\right] \tag{1.5}$$

where $\mathbf{1}_{A=a}$ is the indicator function equal to one when $A=a$ and otherwise zero. These formulas illustrate how identification works: the left-hand side is defined in terms of potential outcomes, whereas the right-hand side is written entirely in terms of observable quantities from the data, allowing us to estimate quantities such as the average treatment effect or risk ratio.

1.2.2. Estimation of treatment effects

Once the identification of a treatment effect is established, one can start estimating this effect from the observed data. An important aspect of causal inference is that to estimate a treatment effect, one must first typically estimate nuisance components that are used to obtain the actual effect. For instance, the Inverse Propensity Weighting (IPW) estimator uses the identification formula in eq. (1.4) and relies on first estimating the probability of treatment given covariates, $\Pr(A=a \mid X)$, also referred to as the propensity score (Rosenbaum & Rubin, 1983b). Meanwhile, the g-computation estimator uses the identification formula in eq. (1.5), and requires us to first estimate the conditional expected outcome as a function of treatment and covariates, $\mathbb{E}[Y \mid X, A]$.

Both the IPW and g-computation estimators are consistent – that is, they converge to their true value as the sample size grows – only if the underlying nuisance components are correctly specified. Other approaches, such as doubly robust estimators, combine both propensity score and outcome modeling and remain consistent as long as at least one of the two nuisance components is correctly specified (Bang & Robins, 2005). Similar types of estimators also exists for other effect measures, such as the conditional average treatment effect (Künzel *et al.*, 2019; Nie & Wager, 2021).

Finally, in recent years, much research has focused on applying machine learning algorithms to estimate the nuisance components in treatment effect estimation (Feuerriegel *et al.*, 2024). While these flexible models can capture complex relationships in the data, they also introduce new challenges, such as overfitting and regularization-induced bias, which can compromise statistical inference. To address these issues, modern causal inference methods often combine machine learning with techniques like cross-fitting and sample splitting, enabling valid statistical inference and unbiased estimates of treatment effects (Chernozhukov *et al.*, 2018; Van Der Laan & Rubin, 2006).

1.2.3. Causal graphical models

One way to reason about identification is to make part of our assumptions more explicit by, quite literally, drawing them out. Directed Acyclic Graphs (DAGs) have become a popular tool to visualize the assumed causal structure of a given problem. Each node in the graph represents a variable of interest, such as the treatment or outcome, and a directed edge from one node to another indicates an assumed causal relationship. We give an example in Figure 1.2, with both X and U being confounders that influence both the treatment A and the outcome Y. The shaded, dashed node U indicates that it is unobserved. In this DAG, the assumption of no unmeasured confounding between treatment A and outcome Y is violated if a variable such as U exists. As a result, if we only observe the variables (X, A, Y), the treatment effect of A on Y cannot be identified from data.

Directed acyclic graphs are valuable practical tools for causal inference. Practitioners can construct DAGs that include all relevant factors they can come up with, both measured and potentially unmeasured, that might influence the treatment and outcome. By encoding their beliefs about the underlying causal structure, they can assess whether the treatment effect they want to study is identifiable from the available data. The presence of unmeasured variables, as in the example above, would signal a possible problem for identification. Importantly, while there exists an entire scientific field dedicated to methods for (partially) learning DAGs from data, verifying whether a given DAG accurately reflects the true data-generating process is generally difficult, if not impossible, in practice (Glymour *et al.*, 2019).

1.2.4. The pitfall of causal inference

To return to our original goal of estimating a treatment effect, we shall continue using the average treatment effect as an example. Assuming that the identification assumptions introduced earlier hold, as would be the case in the DAG in Figure 1.2 when the unmeasured confounder U is absent, the average treatment effect is identified through a formula such as in eq. (1.4) and eq. (1.5). The left-hand side is the causal quantity we want to know, and the right-hand side involves only observed quantities which can actually be estimated from data. However, this equality holds true only if the identification assumptions are valid. Even when these assumptions are violated, such as in the presence of unmeasured confounding, the right-hand side can still be estimated. In such cases, though, it may no longer correspond to the average treatment effect on the left-hand side.

This examples illustrates one of the most common pitfalls in causal inference: the quantity we think we are estimating may not correspond to the *actual* treatment effect. In practice, this disconnect can have serious and possibly unsafe consequences. For example, if we use a biased estimate to inform decisions about an intervention, we risk recommending treatments that are ineffective; or worse, harmful.

The underlying root of this problem is that the identification may fail when untestable assumptions are violated, meaning that there is generally no easy way to detect if we

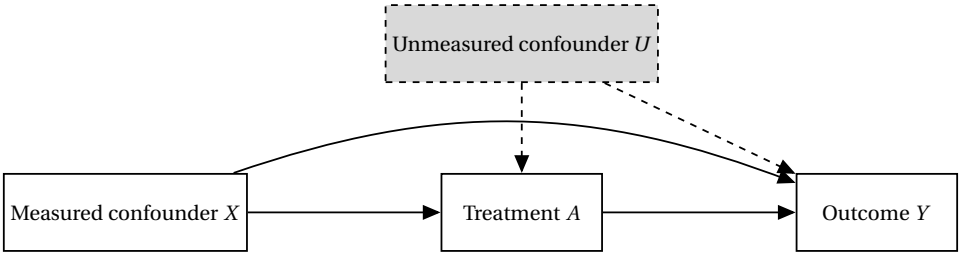


Figure 1.2: An illustration of a directed acyclic graph (DAG) where two covariates, X and U, act as confounders between treatment A and outcome Y. The shaded, dashed node indicates that U is unmeasured. When the dashed edges from U to both A and Y are present, indicating that U is a common cause of both, the treatment effect of A on Y cannot be identified even if X is observed.

have fallen into the pitfall already. This makes causal inference potentially unsafe, since a failure of identification can lead to untrustworthy causal claims – and ultimately to incorrect decisions about whether a new drug or policy is effective. For this reason, we ask the following main question in this thesis: *How can we make causal inference safer against violations of causal identification assumptions?*

In this thesis, we address this question to move towards safer causal inference along two main lines. First, we investigate and propose methods that can be used to detect violations of causal identification assumptions under certain conditions. These methods allow us to explicitly falsify the plausibility of causal identification assumptions, making causal inference safer by helping us recognize when we may already have fallen into the pitfall. Second, we revisit two specific problem settings in causal inference and show that it is possible to develop methods that either remove, or rely on strictly weaker assumptions which are more plausible in practice. This makes causal inference safer by reducing dependence on strong assumptions in the first place, helping us avoid falling into the pitfall altogether.

1.3. Safer causal inference by falsifying assumptions

Understanding the plausibility of the assumptions required to identify and estimate causal quantities from observable data has long been a central focus in the field of causal inference. We began this chapter with the example of epidemiologists in the 1950s asking whether an unmeasured confounder could account for the observed association between smoking and lung cancer. Following the approach of Cornfield *et al.* (1959), one of the earliest examples of what we now refer to as *sensitivity analysis*, it became increasingly clear that it was highly implausible for any single confounder to be strong enough to explain away the observed association. This reasoning contributed to the widespread acceptance that smoking is a cause of lung cancer (Loeb *et al.*, 1984; Wynder, 1997). Sensitivity analysis, therefore, represents the first category of methods for evaluating

violations of causal identification assumptions.

Sensitivity analysis is the investigation of how hypothetical violations of the assumptions required to identify treatment effects could introduce bias (Greenland, 1996). In essence, it provides a framework for assessing how robust our conclusions are if those assumptions are wrong. Sensitivity methods have been applied not only to unmeasured confounding (Tan, 2006; VanderWeele & Ding, 2017), but also to selection bias (Robins *et al.*, 2000), interference (VanderWeele *et al.*, 2015), which relates to the stable unit treatment value assumption discussed earlier, and measurement error (Díaz & van der Laan, 2013). Such analyses often yield bounds on the treatment effect, which is why sensitivity analysis is frequently described as a form of partial identification (Manski, 2003).

Sensitivity analysis is, in principle, a procedure that can always be carried out, since it investigates "what-if" scenarios involving violations of causal identification assumptions. This flexibility of sensitivity analysis is a strength, as it allows researchers to explore the biases of various types of violations. However, the downside of sensitivity analysis is that it does not actually let us refute whether an assumption is valid. Moreover, sensitivity analysis typically requires outside knowledge to determine which "what-if" scenarios are plausible. Thus, an open question in the field of causal inference is when is it also possible to directly refute our assumptions from the data.

This question motivates a second category of methods for evaluating assumptions in causal inference; which is the category we focus on in this thesis. These methods aim to empirically falsify whether assumptions hold, based on testing possible observable implications those assumptions have for the data. For example, when using instrumental variables, which are an important type of variables for certain identification strategies (Angrist *et al.*, 1996), violations of the instrumental variable inequalities proposed by Pearl (1995) indicate that the conditions for valid instruments do not hold. Similarly, when multiple seemingly valid covariate sets are available for adjusting for confounding, examining the stability of treatment effect estimates across these sets can reveal potential misspecification of our initial assumptions (Lu & White, 2014; Su & Henckel, 2022). In both cases, however, these tests cannot definitively confirm the validity of our identification assumptions. While they may indicate that an assumption is invalid, they cannot be taken as evidence for its validity. For this reason, we will refer to these as *falsification methods*.

While falsification methods are appealing because they allows us to directly refute the validity of a causal analysis, unlike sensitivity analysis, constructing such methods is not always possible, since not all assumptions lead to observable implications which can be tested. The successful development of falsification methods therefore relies on first discovering what those implications might be and thereafter developing efficient ways to test them using data.

1.3.1. Falsification by combining information from multiple sources

In this dissertation, we identify a setting where falsification becomes possible, namely when data from multiple sources can be combined. In scientific practice, this combination of data already occurs in contexts such as meta-analyses, where results from randomized trials and observational studies in different populations are compared to strengthen or challenge scientific claims. To understand how falsification may be possible here, one may realize that such comparisons make it possible to falsify existing theories when conflicting results arise from different sources. Building on this idea, we first review an existing falsification method and then introduce the new method developed in this dissertation. In subsequent chapters, we will occasionally refer to sources as environments, reflecting differences in terminology between the statistics and computer science literature on this topic.

Transportability-based falsification

An important concept appearing in much of the scientific literature when combining data from different sources is transportability (Bareinboim & Pearl, 2016; Dahabreh, Robertson *et al.*, 2020). This concept asserts that certain causal quantities, such as the conditional average treatment effect $\mathbb{E}[Y^1 - Y^0 \mid X]$, remain the same across the populations underlying the different studies. Because this condition is defined in terms of counterfactuals, which are only partially observed, it is not directly testable. However, when the transportability condition is assumed together with standard identification assumptions within each population, such as the absence of unmeasured confounding, the joint set of assumptions can yield testable implications in the observed data.

To make the idea of falsification under transportability concrete, we follow Dahabreh, Robertson *et al.* (2020) and consider two datasets examining the same treatment and outcome: one from a randomized study, denoted $S = 1$, and one from an observational study, denoted $S = 0$. If the conditional average treatment effect transports between the two studies, and if the standard identification assumptions hold in both, then we should expect that

$$\begin{aligned} &\mathbb{E}[Y \mid X, A = 1, S = 1] - \mathbb{E}[Y \mid X, A = 0, S = 1] = \\ &\quad = \mathbb{E}[Y \mid X, A = 1, S = 0] - \mathbb{E}[Y \mid X, A = 0, S = 0]. \end{aligned} \tag{1.6}$$

The above equality can be tested from data (Z. Hussain *et al.*, 2023; Luedtke *et al.*, 2019), so if we observe that the equality does not hold, we must conclude that at least one of the underlying assumptions has been violated. As one of the datasets is randomized, we are effectively "benchmarking" the observational study against the randomized study, meaning that an inequality in eq. (1.6) could be explained by the failure of the identification assumptions in the observational study. However, since the discrepancy could also arise because the transportability condition itself is violated, the test cannot tell us whether the failure is due to problems in the observational study, lack of transportability, or both.

The above example of "benchmarking" studies against each other shows that we can falsify causal identification assumptions if we have data from multiple sources. Doing

this successfully will however rely on a key assumption, namely that the transportability condition holds between all sources. This condition can be interpreted as requiring the absence of unmeasured effect modifiers whose distributions are shifted between sources (Hernán & VanderWeele, 2011). In other words, to falsify the presence of one type of unmeasured variable (confounders), we must assume the absence of another (shifted effect modifiers). While it may sometimes be possible to argue in favor of this reasoning, it should be noted that confounders may themselves act as effect modifiers, hence that the very variables we seek to rule out could invalidate the assumption needed for falsification. This challenge motivates the first main question addressed in this thesis: *How can we falsify causal identification assumptions using multiple data sources that are not transportable with one another?*

Mechanism independence-based falsification

Rather than relying on assuming transportability between multiple data sources, in this dissertation we explored another alternative condition. This condition, referred to as the *independence of causal mechanisms* (ICM), has attracted significant attention in the machine learning literature in recent years (Schölkopf *et al.*, 2021), although its origins can be traced back to earlier work by e.g. Haavelmo (1944) and Pearl (2009). Here, a causal mechanism describes how a variable is influenced by its direct causes. For example, in an observational study, the treatment A is typically determined by various factors captured in the covariates X. This process can be represented by the conditional distribution $\Pr(A \mid X)$, which defines a causal mechanism for the treatment. Similarly, the outcome Y may depend on both A and X, and this dependence can be expressed through the conditional distribution of potential outcomes, $\Pr(Y^a \mid X)$, for each treatment level a, which defines the causal mechanism for the outcome. Informally, the ICM condition states that changes in one causal mechanism do not convey information about changes in other mechanisms. In other words, if the assignment mechanism $\Pr(A \mid X)$ changes across different populations, this change is assumed to be uninformative and have no influence of any change in $\Pr(Y^a \mid X)$.

Although variants of ICM in multi-source settings have been studied mainly for causal discovery before, see e.g. B. Huang *et al.* (2020) and Peters *et al.* (2016), it had not yet been leveraged to falsify the assumptions required to identify treatment effects. Thus, to develop a successful falsification method under ICM, we first need to identify a testable implication in the observable data which follows when one assumes some identification assumptions together with ICM, and secondly, we need approaches to efficiently test these implications from the observable data.

In Chapter 2, we address the first step. We show that under ICM, it is possible to falsify the assumption of no unmeasured confounding. A key contribution in this chapter is the use of a new formalism for representing causal graphical models in multi-source settings. This formalism links ICM with nonparametric Bayesian hierarchical modeling and was developed in concurrent work by Guo *et al.* (2023), although in their case it had only been applied to causal systems without unmeasured variables. Together with the hierarchical

graphical models, we use techniques from constraint-based causal discovery (Glymour *et al.*, 2019) to falsify the assumption of no unmeasured confounding in the data. While the literature on hierarchical models in causal inference is scarce, recent work by Weinstein and Blei (2024) examined different identification strategies using hierarchical models. Interestingly, these authors emphasize that addressing subunit-level confounding in hierarchical settings is an important direction for future research – precisely the type of confounding that we are able to detect in this chapter.

Chapter 3 then turns to the second step of falsification by improving the efficiency of the falsification method introduced in Chapter 2. We achieve this by reformulating ICM directly in terms of parameterized linear statistical models. Each parameter in these models corresponds to an underlying causal mechanism. This framing naturally leads to a two-step procedure that proved more data-efficient than the constraint-based approach from Chapter 2. In this procedure, we first estimate the parameters, and second, test for independence among them.

Chapter 4 goes back to discovering more testable implications by extending the ideas from Chapter 2 to other identification strategies that rely on mediators or instrumental variables. Using a similar hierarchical model as before, we derive new testable implications under ICM in these settings as well.

A key advantage of ICM, as shown in part one of this dissertation, is that it permits certain violations of transportability. While transportability in essence constrains causal quantities to remain unchanged across sources, ICM allows for this but restricts how such changes occur. And interestingly, we found that no randomized data is required to allow for this falsification to be possible. Both these aspects suggests that in practice assuming ICM may be less restrictive than assuming transportability when performing falsification. However, like transportability, ICM is not directly testable from observed data and therefore should not be accepted at face value.

The falsification methods we discussed so far are meant to be applied before any treatment effect estimation is performed, during the phase when one is still evaluating which assumptions are plausible with the collected data at hand. The conclusions drawn at this stage help guide the identification strategy, which then later informs how to proceed with estimating treatment effects from the actual data. The next part of this dissertation moves from testing assumptions to estimation. Here, following our goal of achieving safer causal inference, the focus shifts to developing methods that remain robust even when some untestable assumptions fail.

1.4. Safer causal inference by assumption-robust methods

The remainder of this thesis focuses on another aspect of safe causal inference: estimating treatment effects in ways that are robust to violations of untestable assumptions. In general, this is a difficult goal to achieve. If a method truly required no such assumptions, it would not need to account for their potential violation in the first place, but without

assumptions we have no way of connecting the causal quantities of interest to the observed data. Nevertheless, we will examine two problem settings where we relax some of the assumptions that have been presented earlier in this introductory chapter.

1.4.1. Trial augmentation using external data

We first consider the problem of trial augmentation using external data, which has received increasing attention in recent years (Jahanshahi *et al.*, 2021), particularly in medicine. This involves working with data from multiple sources. Specifically, we start with data from a randomized trial, aiming to improve the estimation of the treatment effect for the underlying trial-eligible population by leveraging data from a different, external study population.

The randomized design of trials helps ensure that the assumptions needed to identify the treatment effect in the trial population are satisfied. However, the high cost of trials often limits sample sizes, making statistical inference challenging. To address this, one idea is to use external data, such as from another trial or an observational study, to improve the precision of treatment effect estimates (Colnet *et al.*, 2024).

Using external data to improve the efficiency in trials works if the trial and external populations are sufficiently aligned. In much of the existing literature, this alignment is formalized through a transportability condition (Colnet *et al.*, 2024), which we already described earlier. However, in many cases if the populations are misaligned, for instance if transportability is violated or the external data includes unmeasured confounding, incorporating external data can introduce bias. For instance, X. Li *et al.* (2023) derive an efficient estimator that pools trial and external data under the assumption of transportability, but this estimator can become biased if transportability is violated. Other methods attempt to address this by adaptively selecting external data samples that are aligned with the trial (Yang *et al.*, 2023). However, this selection is itself challenging, and consequently bias may be introduced by still including misaligned external data, leading to bias in the treatment effect estimate.

That we rely on additional assumptions to include external data creates a fundamental problem: data from randomized trials provide unbiased estimates by design, but incorporating external data may again reintroduce untestable assumptions that can undermine the advantage brought by randomized trials. Consequently, one could argue that many existing methods for augmenting trials with external data do not offer a safe solution. Instead, a safe solution would be a method that leverages external data when it may bring benefit but ensures that regardless of whether the external data is used or not, we never do worse than if we had looked at only the trial data. This open challenge lead to the second main question addressed in this thesis: *How can we safely augment treatment effect estimation in randomized trials with external data without assuming the trial and external population to be well-aligned?*

To tackle this problem, in Chapter 5, we propose a novel, efficient estimator that integrates

external data with individuals who specifically received the same control treatment into a new randomized trial. This estimator improves inference on the average treatment effect while ensuring two key asymptotic properties: first, it does not introduce bias when including misaligned external data; and second, it achieves variance lower than or, in the worst case, equal to that of the best estimator using only the trial data.

Next, in Chapter 5, we address a related task to one in the previous chapter. Randomized trials are often underpowered to estimate measures of treatment effect heterogeneity, such as the conditional average treatment effect. We show that incorporating external data can make this estimation feasible by adapting the estimator from Chapter 5 to handle the conditional average treatment effect. Additionally, we expand the framework to allow integration of external data across multiple trial arms, rather than being limited to the control arm.

1.4.2. Evaluating treatment policies

For the final problem setting studied in this dissertation, we consider the problem of evaluating policies which decide on how to allocate a treatment across different individuals. This problem arises in settings such as personalized medicine (Kravitz *et al.*, 2004) and marketing or e-commerce (Goldenberg *et al.*, 2020), and is especially important when treatments are costly, effects vary across individuals, and it is necessary to determine who should be prioritized getting the treatment.

One common approach to evaluate if a treatment allocation strategy manages to identify those who benefit the most from a treatment is through Qini curves (Radcliffe, 2007), which have become especially popular in marketing and e-commerce applications. Estimating a Qini curve allows us to evaluate a treatment allocation strategy by their ability to rank individuals by their predicted treatment effect, from greatest expected benefit to least. To estimate a Qini curve we need a hold-out test dataset, typically collected from a randomized experiment, which are segmented based on the ranking from a given treatment prioritization strategy. This approach can provide insight into the expected treatment effect compared to if no one is treated when treatment is allocated to the top 10%, the top 20%, and so on, according to the strategy.

Figure 1.3 illustrates two Qini curves for different treatment allocation strategies in a dataset with strong treatment effect heterogeneity, where only about 30% of individuals respond positively to treatment. Comparing the curves, it allows us to identify the optimal strategy that outperforms the suboptimal one, achieving nearly all the potential treatment gain after treating roughly 30% of individuals; here it is assumed that higher is better. In contrast, the suboptimal strategy ranks individuals poorly, which results in a less effective allocation.

Even when the hold-out test dataset comes from a randomized experiment, however, an untestable assumption is often made when estimating Qini curves: treating one individual does not affect outcomes for others in the population. This is known as the assumption of

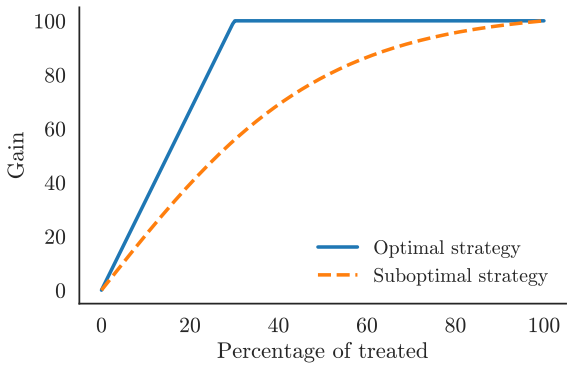


Figure 1.3: Qini curves for two treatment allocation strategies in a population with treatment effect heterogeneity. The gain on the y-axis represents the average improvement in outcome (assuming higher is better) when applying the strategies, compared to a scenario in which no one in the population is treated. The optimal strategy prioritizes individuals who benefit most, achieving nearly all the potential treatment gain after treating about 30% of the population, while the suboptimal strategy ranks individuals poorly and yields a less effective allocation.

no interference. In some applications, this untestable assumption is expected to be violated. For example, in marketing, promoting different products can affect buyers' overall preferences (Bajari *et al.*, 2023), while in vaccine studies, immunity among vaccinated individuals can protect those who remain unvaccinated (VanderWeele *et al.*, 2012).

Violations of interference generally lead to biased treatment effect estimates, and the same applies to Qini curves. Consequently, failing to account for interference in Qini curves can result in incorrect evaluations of the cost-effectiveness of treatment allocation strategies. While previous research on interference has primarily focused on treatment effect estimation, see e.g. Hudgens and Halloran (2008) and Tchetgen and VanderWeele (2012), considerably less attention has been given to its impact on the evaluation of treatment allocation strategies. To address this gap, the third and final question examined in this thesis was: *How can we evaluate treatment allocation strategies in the presence of interference?*

To address this problem, in Chapter 7, we propose a framework for estimating Qini curves under a setting that allows for interference; specifically, when one has clustered network interference. In this framework, individuals are grouped into independent clusters that do not affect each other, restricting interference to occur only within each cluster. Returning to the marketing example on an e-commerce site, this assumption is reasonable if products can be divided into categories: for instance, promoting pet products is unlikely to interfere with sales of baby products. With our proposed framework, we demonstrate how various existing estimators from the literature on estimating treatment effects under

interference can be adapted for estimating Qini curves.

To conclude, both in this chapter on evaluating treatment allocation strategies and in the preceding chapters on robust trial augmentation, we introduce novel methods for causal inference that rely on fewer or weaker assumptions than existing approaches, thereby providing practitioners with additional tools for conducting causal inference more safely.

1.5. Thesis outline

We present a brief overview of papers which are included in this thesis.

In Part One, we consider the problem of falsification from multi-source data.

Chapter 2 Through the novel use of a hierarchical causal graphical model of the data assuming independent causal mechanisms, we propose a falsification strategy that can falsify the assumption of no unmeasured confounding.

Chapter 3 We show that we can improve the method proposed in Chapter 2 by incorporating functional assumptions on causal mechanisms, instead of relying on the graphical model, leading to a simpler and more sample-efficient falsification strategy.

Chapter 4 We further demonstrate that the hierarchical model used in Chapter 2 can sometimes be used to obtain falsifiable implications for other causal identification assumptions, in particular the validity of front-door criteria and instrumental variables.

In Part Two, we study methods for robust trial augmentation leveraging data from external sources outside the trial.

Chapter 5 We propose a novel robust estimator for the average treatment effect, which can improve precision in estimating the average treatment effect in a trial population with the help of external control data from a different population. We prove that the estimator is robust in the sense that it asymptotically never performs worse than an estimator which only uses data from the trial alone, even if the external data source is misaligned with the trial.

Chapter 6 We consider the problem of estimating heterogeneous treatment effects in the same setting, but where we also might have data for both the treatment and control arms in the external data. We use ideas from Chapter 5 to propose a robust conditional average treatment effect learner, proving and illustrating that it can improve predictions of treatment effects while never performing worse than if we had used trial data alone, even if the integrated data is misaligned.

In Part Three, we study a specific problem setting when evaluating treatment allocation strategies using Qini curves when the assumption of no interference is violated.

Chapter 7 We propose a general methodology for evaluating treatment prioritization rules when the assumption of no interference is relaxed, instead allowing a structured form of interference known as clustered network interference. Specifically, we develop a framework for estimating Qini curves in this setting and illustrate how different existing estimators from the literature fit into this framework. We then analyze different choices of estimators both theoretically and empirically and provide recommendations for how to select between them in practice.

In the final part of this dissertation, we summarize the main findings from the preceding chapters and highlight key takeaways. We then outline several interesting directions for future research that could be pursued over the next five to ten years in the causal inference community.

Part One

Falsification of Causal Assumptions

2

Falsifying Unconfoundedness with Causal Hierarchical Graphical Models

A common assumption in causal inference from observational data is that there is no hidden confounding. Yet it is, in general, impossible to verify this assumption from a single dataset. Under the assumption of independent causal mechanisms underlying the data-generating process, we demonstrate a way to detect unobserved confounders when having multiple observational datasets coming from different environments. We present a theory for testable conditional independencies that are only absent when there is hidden confounding and examine cases where we violate its assumptions: degenerate & dependent mechanisms, and faithfulness violations. Additionally, we propose a procedure to test these independencies and study its empirical finite-sample behavior using simulation studies and semi-synthetic data based on a real-world dataset. In most cases, the proposed procedure correctly predicts the presence of hidden confounding, particularly when the confounding bias is large.

2.1. Introduction

Estimating the causal effect of a treatment on an outcome is a fundamental challenge in many areas of science and society. While this is straightforwardly done using data from randomized studies, using observational data for this task is appealing since they are often more feasible to collect while also being more representative of the population of interest (Pearl, 2009). To identify causal effects using such data it is often assumed

This chapter appears as: Karlsson, R., & Krijthe, J. (2023). Detecting hidden confounding in observational data using multiple environments. *Advances in Neural Information Processing Systems, 36*, 44280–44309

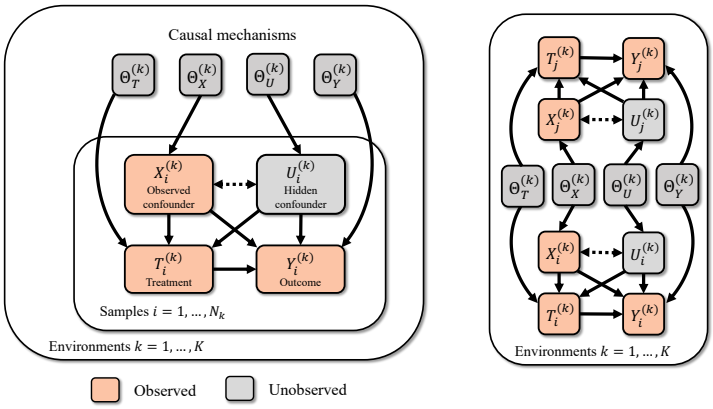
there is no hidden confounding. When this *untestable* assumption is violated we run the risk of confusing causal relationships with spurious correlations. This can have serious consequences such as unknowingly suggesting a non-effective or, even worse, potentially harmful treatment. Therefore, detecting the presence of hidden confounding is an important problem.

Data collected from different sources often tend to be heterogeneous due to e.g. changing circumstances or time shifts. In this work, we show how such heterogeneity can be exploited to make hidden confounding testable *solely* from observational data. We consider a setting where observational data has been collected from different environments E. In each environment, we observe the same treatment T and outcome Y, as well as covariates X that are known confounders between T and Y. Further, we assume that the data is heterogeneous across these environments under the principle of *Independent Causal Mechanisms* (Peters *et al.*, 2017; Schölkopf *et al.*, 2021), which states that a causal system consists of autonomous modules that do not inform or influence each other. The question we ask is whether there exists further hidden confounding between T and Y after having adjusted for X. If that is the case, the causal effect of T on Y is not identifiable in general. Perhaps surprisingly, we demonstrate a way to decide if the causal effect would be identifiable by deriving testable implications for whether hidden confounding is present or not. We achieve this by exploiting the hierarchical structure of the problem, shown in Figure 2.1.

As an illustration of a setting where this might be applied, there are many existing multi-level studies in which individuals are nested in clusters and non-randomly assigned to a treatment/control on an individual level. For instance, we can have pre-defined clusters in a multi-level observational study that investigate a specific treatment and outcome from multiple hospitals (Goldstein *et al.*, 2002) or schools (Leite *et al.*, 2015) that care for patients/pupils from different demographics. Here the clusters constitute different environments. Often we might suspect the existence of potential individual-level confounders such as socio-economic status. Now, if these confounding factors have different distributions at each cluster our work proposes a way to statistically test the presence of confounding between the treatment and outcome – even when we do not observe the confounding factors directly.

Another example where our method is suitable is when there are multiple observational studies where no randomized control trials are available, a problem area where systematic procedures are still lacking (Mueller *et al.*, 2018). In particular, individual participant data meta-analyses are a type of analysis that uses all individual-level data from multiple studies instead of aggregating summary statistics (Di Angelantonio *et al.*, 2016; Riley *et al.*, 2010). With this information and given that we observe the same treatment and outcome across all studies, our proposed algorithm can also be used to detect if there is common hidden confounding among the studies.

Contributions We prove that there exists, under the principle of independent causal mechanisms, testable independencies that are only violated in the presence of unob-



(a) (b)

Figure 2.1: We have multiple observations $i = 1, \ldots, N_k$ of treatment $T_i^{(k)}$, outcome $Y_i^{(k)}$ and confounder $X_i^{(k)}$ in different environments $E^{(k)}$. The dashed bi-directed edge between $X_i^{(k)}$ and $U_i^{(k)}$ allows for any causal relationship, or lack thereof, between the observed and hidden confounder. **(a)**: The hierarchical structure of the multi-environment data; the causal mechanisms are unobserved but we know the indicator $E^{(k)}$ for what environment observations belong to. **(b)**: By unrolling the graph in (a) we see that dependencies exist between any pairs of observations (i, j) from the same environment (when not conditioning on the mechanisms). This can be exploited to detect the presence of the hidden confounder.

served confounding between treatment and outcome (Sec. 2.4, Theorem 2.1). Further, we explore the effect of changes and violations of our assumptions – while most assumptions are necessary, we find that some can be relaxed (Sec. 2.4.1). We then introduce a statistical testing procedure that uses any suitable conditional independence test to detect the presence of hidden confounding in observational datasets from multiple environments (Sec. 2.4.2). Lastly, we perform an empirical finite-sample analysis of it using both synthetic and semi-synthetic data generated with real-world covariates from the Twins dataset (Almond *et al.*, 2005; Louizos *et al.*, 2017). We observe that our proposed procedure correctly predicts the presence of hidden confounding in most cases, particularly when the confounding bias is large (Sec. 2.5).

2.2. Problem setting

We start with some preliminaries of the causal terminology used in this paper.

Definition 2.1 (Causal Graphical Model (CGM))**.** *A causal graphical model* $M = (\mathcal{G}, P)$

over d random variables $\mathbf{V} = (V_1, V_2, \ldots, V_d)$ *comprises (i) a directed acyclic graph (DAG)* $\mathcal{G}$ *with vertices* $\mathbf{V}$ *and edges* $V_i \to V_j$ *iff* V_i *is a direct cause of* V_j*, and (ii) a joint distribution* P *such that it has the following Markov or causal factorization over* $\mathcal{G}$*:*

$$P(V_1, V_2, \ldots, V_d) = \prod_{i=1}^{d} P(V_i \mid \text{Pa}(V_i)) \tag{2.1}$$

where $\text{Pa}(V_i)$ *denotes the parents (direct causes) of* V_i *in* $\mathcal{G}$ *and* $P(V_i \mid \text{Pa}(V_i))$ *is the causal mechanism of* V_i*.*

The DAG $\mathcal{G}$ encodes various conditional independencies between the variables – also known as d-separations in the DAG, see Pearl (1988, Chapter 3.3) – which we write as $\mathbf{A} \perp\!\!\!\perp_d \mathbf{B} \mid \mathbf{C}$ over some disjoint sets of variables $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. We shall assume that conditional independencies in $\mathcal{G}$ imply the same conditional independencies in P, and vice versa:

Assumption 2.1 (Faithfulness & Causal Markov Property)**.** *For* P *and* $\mathcal{G}$ *we have (i) the faithfulness property that* $\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \Rightarrow \mathbf{A} \perp\!\!\!\perp_d \mathbf{B} \mid \mathbf{C}$*, and (ii) the causal Markov property that* $\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \Leftarrow \mathbf{A} \perp\!\!\!\perp_d \mathbf{B} \mid \mathbf{C}$*.*

We consider a setting with the following variables in our causal graphical model: a one-dimensional treatment $T \in \mathcal{T}$ and outcome $Y \in \mathcal{Y}$, in addition to some observed covariates $X \in \mathcal{X}$ and unobserved covariates $U \in \mathcal{U}$. We do not restrict the dimensionality of X and U. Additionally, in this setting, the environment E has a direct effect on all other variables, making it a root node in $\mathcal{G}$. We say that a variable is a confounder between T and Y if it is a cause of both T and Y in $\mathcal{G}$. We assume that X is a known confounder between T and Y, while the relationship between U and the other variables is unknown. Hence, U could be an unobserved hidden confounder (as illustrated in Figure 2.2) or, for instance, completely unrelated to the other variables.

Expressed in the framework of Pearl (2009), the goal of causal inference is to estimate the probability $P(Y \mid do(T = t))$ where $do(T = t)$ represents an intervention on the treatment. Without any further assumptions, $P(Y \mid do(T = t))$ is not identifiable from an observational dataset; that is data where we have observed the choice of treatment without influencing it (Pearl, 2009). In particular, in the setting we consider here, the interventional effect remains unidentifiable if the unobserved U is a confounder between T and Y. [1] Unfortunately, there is no way to check whether such unobserved confounders are present in a single dataset. We will show, however, that things are different when we have access to observational datasets from multiple environments. In this setting, we present a way to detect confounding even if it is not observed, hence demonstrating a novel and valuable approach for verifying an essential prerequisite for causal inference from observational data. In the rest of this section, we present the main assumptions that enable us to do this.

[1] There exist other procedures that could circumvent this issue, but these alternatives seldom avoid the unconfoundedness assumption completely. As an example, instrumental variable estimation is applicable when there is unobserved confounding between T and Y, but only when the relationship between the instrumental variable and T is unconfounded (Angrist *et al.*, 1996).

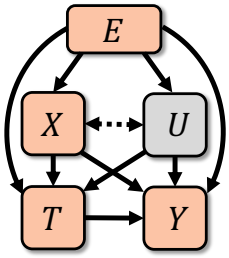


Figure 2.2: The setting where we want to detect the presence of a hidden confounder U in $\mathcal{G}$.

First, we have data from multiple environments $E \in \mathcal{E}$ with different joint distributions $P(T, Y, X, U \mid E)$. We shall use $P_E(\cdot)$ to denote $P(\cdot \mid E)$, and use small letters for the random variables whenever they take particular values. In our setting, we have datasets $D_k = \{t_i^{(k)}, y_i^{(k)}, x_i^{(k)}\}_{i=1}^{N_k}$ from multiple environments $e^{(1)}, e^{(2)}, \ldots, e^{(K)}$; each has N_k observations which are assumed to be i.i.d. within the environment. N_k is fixed but can be different for each environment. The environments are related to each other through the following assumption.

Assumption 2.2 (Shared Causal Graph)**.** *All environments share the same underlying causal DAG $\mathcal{G}$.*

Next, we specify how changes in $P_E(T, Y, X, U)$ arise between the different environments. We shall assume that the conditional probabilities in eq. (2.1) – which we refer to as causal mechanisms – vary independently per environment. This is known as the independent causal mechanism principle.

Assumption 2.3 (Independent Causal Mechanism (ICM) Principle (Peters *et al.*, 2017))**.** *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its parents in the causal graph) does not inform or influence the other mechanisms.*

The above assumption covers two aspects: one concerning *informing* and the other about *influencing*. That the mechanisms do not *inform* each other can be interpreted as that knowing the conditional probability of one variable does not tell us anything about the conditional probabilities of other variables (Guo *et al.*, 2023; Janzing & Schölkopf, 2010). Further, we assume that changing (or performing an intervention upon) one mechanism has no *influence* on other mechanisms (Schölkopf *et al.*, 2012). While this notion of independence between mechanisms can be described through a non-stochastic, algorithmic mutual information (Janzing & Schölkopf, 2010), we focus in this work explicitly on statistically independent mechanisms.

To model changes between environments with independent causal mechanisms, we parameterize each causal mechanism with $\Theta_V \in \mathcal{O}_V$ for $V \in \{T, Y, X, U\}$. In each environment, these parameters are fixed and determine the distribution

$$P_E(T, Y, X, U) = \prod_{V \in \{T,Y,X,U\}} P_{\Theta_V}(V \mid \text{Pa}(V)) \,. \tag{2.2}$$

Note that while we need to know which observations come from which environment, we do not assume to know the particular values of the individual parameters $(\Theta_T, \Theta_Y, \Theta_X, \Theta_U)$ in any environment. We shall assume that environments are randomly sampled from a *distribution over mechanisms* by defining non-degenerate probability measures for each causal mechanism.

Assumption 2.4 (Non-degenerate Probabilistic Independent Causal Mechanisms)**.** *The independent causal mechanisms are non-degenerate random variables with probability measures $P(\Theta_V)$ for all $V \in \{T, Y, X, U\}$ such that Θ_T, Θ_Y, Θ_X and Θ_U are pairwise independent random variables.*

With the above assumption, when we now say *independent* causal mechanisms, we refer to statistical independence between them. We argue that the above assumption is not particularly strong if we already have Assumption 2.3; the mechanisms are now allowed to change across environments in a probabilistic manner. Guo *et al.* (2023) proved the existence of such probability measures for the causal mechanisms when the data comprises an infinitely exchangeable sequence of random variables, drawing parallels to de Finetti's theorem (de Finetti, 1937).

As a final note on the assumptions we have made: these assumptions should not be taken for granted and it is crucial to also understand how violations of them will influence our theory. For this reason, we will cover this topic in Section 2.4.1.

Hierarchical model of the environments Using these assumptions, we can now express the distribution of the datasets $\{D_k\}_{k=1}^K$ as a hierarchical model (Gelman *et al.*, 2013, Chapter 5), wherein we first sample the mechanisms i.i.d. $\Theta_V^{(k)} \sim P(\Theta_V)$ for $k = 1, \ldots, K$ and $V \in \{T, Y, X, U\}$ and then, for each environment k, obtain $(T_i^{(k)}, Y_i^{(k)}, X_i^{(k)}, U_i^{(k)})$ by repeatedly sampling N_k times according to eq. (2.2). Using plate notation, we can compactly represent this hierarchical model in the augmented DAG $\mathcal{G}^*$ shown in Figure 2.1a. The edges in $\mathcal{G}^*$ between $(T_i^{(k)}, Y_i^{(k)}, X_i^{(k)}, U_i^{(k)})$ are the same as those between (T, Y, X, U) in $\mathcal{G}$ and $\Theta_V^{(k)} \in \text{Pa}(V_i^{(k)})$ where $V \in \{T, Y, X, U\}$, for all i and k.

In the next parts of the paper, we will prove how the structure of $\mathcal{G}^*$ implies novel observable constraints in the multi-environment data distribution that can be exploited to statistically test the presence of hidden confounding between T and Y after having adjusted for X. But first, we discuss the main literature related to our work.

2.3. Related work

This paper contributes to the growing body of research based on the principle of *Independent Causal Mechanisms* (Peters *et al.*, 2017) which has inspired further research on integrating machine learning and causality (Kügelgen *et al.*, 2020; Peters *et al.*, 2016; Schölkopf *et al.*, 2012, 2021). Multiple works have demonstrated how the independent causal mechanism principle could improve causal structure learning when data comes from heterogeneous environments that share the same causal model (Ghassami *et al.*, 2018; Guo *et al.*, 2023; K. Zhang *et al.*, 2017). In particular, Guo *et al.* (2023) demonstrated how independent causal mechanisms imply independence constraints similar to ours when the data is exchangeable – but they assume there exist no unobserved latent variables in contrast to our work where we detect the presence of such variables.

Detecting hidden confounding is hard, and often we can only reason about the plausibility of having unmeasured confounders using some sort of sensitivity analysis (Cinelli *et al.*, 2019; Rosenbaum & Rubin, 1983a; VanderWeele & Ding, 2017). Other approaches check whether a treatment effect estimate is robust to changes in our assumptions by varying the adjustment set (Lu & White, 2014; Oster, 2019; Su & Henckel, 2022). However, the guarantees are elusive for whether this type of robustness implies unconfoundedness. Similarly, one could test for heterogeneity of the treatment effect estimates from multiple environments and conclude that if they are different, then it is due to unobserved confounding; this idea bears resemblance to the pseudo-treatment approach discussed by Imbens and Rubin (2015) for assessing unconfoundedness. But testing heterogeneity to detect confounding only works if the treatment effect is assumed to be fixed across all environments, which excludes many real-world settings. Lastly, Janzing and Schölkopf (2018) proposed a method to detect hidden confounding which is restricted to settings with linear models.

In the setting with data from multiple environments, various approaches have been proposed to deal with hidden confounding, typically by combining both experimental and observational data (Athey *et al.*, 2020; Bareinboim & Pearl, 2016; Hatt *et al.*, 2022; Ilse *et al.*, 2022; Imbens *et al.*, 2022; Kallus *et al.*, 2018). In contrast, we consider a setting combining *only* observational data from multiple environments. Some works make parametric assumptions in this case, such as B. Huang *et al.* (2020), assuming linearity with non-Gaussian noise. Since we want to avoid strong parametric assumptions, we consider approaches that avoid these assumptions. The principled *Joint Causal Inference* (JCI) framework (Mooij *et al.*, 2020) is one such approach. It demonstrates how to apply traditional constraint-based methods for causal discovery (Glymour *et al.*, 2019) with multi-environment data. In the simpler setting with observed variables (T, Y, E) excluding X, the JCI framework informs us that $Y \perp\!\!\!\perp_P E \mid T$ is violated in the presence of a hidden confounder U if E is an instrumental variable. But this means, once again, that the treatment effect is fixed across environments as we assume E has no direct effect on Y. Variants of this type of test have also been mentioned by others, for instance Athey *et al.* (2020, Lemma 3) and Dahabreh, Robins and Hernán (2020). We demonstrate the limitations of using this approach in our experiments, and provide a more in-depth explanation using graph-based arguments in Appendix 2.C. Our contribution is a more general non-

parametric test that works even if E is an invalid instrument that can influence any of the other variables.

2.4. Detecting hidden confounding in multi-environment data

Our goal is to detect the presence of hidden confounding between treatment T and outcome Y after having adjusted for some observed confounders X. Graphically, this corresponds to detecting the existence of both edges $U \to T$ and $U \to Y$ in the causal DAG $\mathcal{G}$. In this section, we demonstrate testable conditional independencies between the observed variables that are *only* violated when both those edges exist – hence providing testable implications for hidden confounding.

While we do not assume to know the complete causal DAG $\mathcal{G}$ between our variables, we put two restrictions on it: (i) that Y is not an ancestor of T and (ii) that X is a confounder to both T and Y in contrast to, for instance, being a mediator or only a cause to either one of them. These restrictions are relatively weak as (i) holds in all practical causal inference settings as a treatment T happens before outcome Y in time and (ii) can sometimes be verified by checking that both T and Y depend on X. Under this setting, we prove the following.

Theorem 2.1. *Let* $\mathbf{T}^{(k)} = (T_1^{(k)}, \dots, T_{N_k}^{(k)})$ *be the vector of all observed treatments in environments* $E^{(k)}$*; define* $\mathbf{Y}^{(k)}$, $\mathbf{X}^{(k)}$, *and* $\mathbf{U}^{(k)}$ *similarly. We consider the data distribution* $P(\mathbf{T}^{(k)}, \mathbf{Y}^{(k)}, \mathbf{X}^{(k)}, \mathbf{U}^{(k)})$ *with* $N_k \geq 2$ *under assumption 2.1,2.2, 2.3 and 2.4. Furthermore, assume an underlying causal DAG* $\mathcal{G}$ *where* Y *is not an ancestor of* T*, and that* X *is a known common cause to* T *and* Y*. Then, for any* $k = 1, \dots, K$*, there exists hidden confounding between* T *and* Y *in* $\mathcal{G}$ *if and only if*

$$T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)} \;\; \forall i, j = 1, \dots, N_k : i \neq j \,. \tag{2.3}$$

Proof sketch. To prove the statement, we look at d-separations in the extended causal graphical model $\mathcal{G}^*$ and show that eq. (2.3) only is true for corresponding graphs $\mathcal{G}$ where the unobserved U is a confounder between T and Y. Figure 2.1b illustrates how open paths may exist between pairs of observations (i, j) going through $\Theta_T^{(k)}, \Theta_Y^{(k)}, \Theta_X^{(k)}$ or $\Theta_U^{(k)}$ by unrolling the augmented graph $\mathcal{G}^*$. These paths are open because of Assumption 2.4. This technique resembles the twin network method used for counterfactual inference (Balke & Pearl, 1994) but the results we obtain from using this approach are distinctly different. The complete proof can be found in the Appendix. □

The variables $T_j^{(k)}$ and $Y_i^{(k)}$ are the treatment and outcome of two different observations in the same environment. Intuitively, the theorem states that after having adjusted for $(T_i^{(k)}, X_i^{(k)}, X_j^{(k)})$, we would expect under the ICM principle that $T_j^{(k)}$ to not provide any

information about how $Y_i^{(k)}$ behaves. Thus, if it still does, then this can only be due to unobserved confounding. Testing this independence hence provides us with a testable implication in our observed data distribution on whether the unobserved U is a confounder or not.

Two-variable case without observed confounders We can drop the observed confounder X in Theorem 2.1 and, interestingly, in that case, obtain even stronger results for detecting the presence of a hidden confounder. This setting is interesting as even the two-variable case is notoriously difficult in causal discovery (Peters *et al.*, 2017; Reichenbach, 1956). Unlike in the more general setting, we no longer need to know the direction of the causal relationship between T and Y.

Theorem 2.2. *Let* $\mathbf{T}^{(k)} = (T_1^{(k)}, \dots, T_{N_k}^{(k)})$ *be the vector of all observed treatments in environments* $E^{(k)}$*; define* $\mathbf{Y}^{(k)}$ *and* $\mathbf{U}^{(k)}$ *similarly. We consider the data distribution* $P(\mathbf{T}^{(k)}, \mathbf{Y}^{(k)}, \mathbf{U}^{(k)})$ *without any observed confounders and* $N_k \geq 2$ *under assumption 2.1,2.2, 2.3 and 2.4. Then, for any* $k = 1, \dots, K$*, there exists hidden confounding between* T *and* Y *in* $\mathcal{G}$ *if and only if*

$$(i)\ T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)} \quad \text{and} \quad (ii)\ T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid Y_j^{(k)} \ \ \forall i, j = 1, \dots, N_k : i \neq j\ . \tag{2.4}$$

Guo *et al.* (2023) studied a similar setting to Theorem 2.2 and demonstrated how to decide the causal direction between T and Y in this case when there is no latent variable. Our results extend theirs as we now also show how to exclude the possibility of a latent common cause in this setting. The proof is similar to that of Theorem 2.1, but the conditional independencies are different. Firstly, we have $T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}$ which is the conditional independence in Theorem 2.1 without conditioning on $X_i^{(k)}$ and $X_j^{(k)}$. Secondly, we have $T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid Y_j^{(k)}$. This one is necessary as we no longer assume anything about the ancestral relationship between treatment and outcome. If we had assumed that T could not be a descendant of Y, we can show that only condition (i) in the theorem is necessary. Similarly, condition (ii) is only necessary when Y could not be a descendant of T.

2.4.1. Influence of the assumptions

Our theory shows how to test for hidden confounding, but it now relies on other untestable assumptions: namely non-degenerate independent causal mechanisms and the faithfulness & causal Markov property. Due to this, we investigate the necessity of these assumptions and identify various failure cases when they are violated. On a more positive note, we also demonstrate that the assumption of non-degenerate mechanisms can be weakened. We present here the main conclusions regarding violations on two of the assumptions while more elaborate explanations can be found in Appendix 2.D, together with a demonstration of how our procedure can fail due to faithfulness violations as well as a discussion on assumptions about positivity and selection bias.

Violation of Assumption 2.3: dependent causal mechanisms What happens if any of the pair-wise independencies between $\Theta_T, \Theta_Y, \Theta_X$ or Θ_U are violated? To investigate this, we go through the same procedure for proving Theorem 2.1 where we allow any of these mechanisms to be dependent. We find that $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$ can be violated even when there is no confounding in all but one case with dependent mechanisms, meaning that it no longer works for detecting hidden confounding. The only case where our theory still works is when $\Theta_X \not\perp\!\!\!\perp_P \Theta_U$ – i.e. the mechanisms of the observed and unobserved confounders are allowed to co-vary across environments.

Violation of Assumption 2.4: degenerate causal mechanisms What happens if one or more of the distributions $P(\Theta_T)$, $P(\Theta_Y)$, $P(\Theta_X)$ and $P(\Theta_U)$ are degenerate, meaning that some mechanisms are fixed across all environments? In the most extreme case, if all mechanisms are fixed then the distribution P_E would be identical in each environment. We investigate these scenarios by first adding Θ_T, Θ_Y, Θ_X and/or Θ_U to the conditioning set of the independence in Theorem 2.1. Then, we check whether this independence still is violated in the presence of hidden confounding using the same procedure used for proving the theorem. We find that the theorem fails only when we condition on both Θ_T and Θ_U. In other words, it is only strictly necessary for our theory that changes in $P_{\Theta_T}(T \mid \text{Pa}(T))$ or $P_{\Theta_U}(U \mid \text{Pa}(U))$ occur between environments.

Remark 2.1. *We may now identify a more conservative interpretation of our proposed procedure. First, one can verify the assumption of non-degenerate causal mechanisms by checking from data whether $P_E(T \mid X)$ varies across environments; if it does, then that is likely because Θ_T and/or – through potential downstream effects – Θ_U are non-degenerate. Next, we would run our proposed procedure. Now if the null is rejected then we can be conservative by concluding that this is either because we have hidden confounding and/or dependent mechanisms. But in the case of no rejection, it can only be interpreted as having no hidden confounders present. This is because having dependent causal mechanisms (violation of assumption 2.3) can only cause false positives.*

2.4.2. Testing the independence

Here, we explain how test the conditional independence $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$ from Theorem 2.1 using multi-environment data; the full procedure is summarized in Algorithm 1 where we have defined $t_{2i-1}^{\pi_{2i}} := \{t_{2i-1}^{(k)}\}_{k \in \pi_{2i}}$ and similarly for $y_{2i}^{\pi_{2i}}, t_{2i}^{\pi_{2i}}, x_{2i-1}^{\pi_{2i}}$ and $x_{2i}^{\pi_{2i}}$.

To test $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$, we need to simulate sampling from the joint distribution $P(T_i^{(k)}, Y_i^{(k)}, X_i^{(k)}, T_j^{(k)}, Y_j^{(k)}, X_j^{(k)})$. Note here that we do not condition on $E^{(k)}$. The idea is as follows: we select two different observations i and j from all environments such that we get a vector of observed treatments $t_i = (t_i^{(1)}, t_i^{(2)}, \ldots, t_i^{(K)})$; outcomes $y_i = (y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(K)})$; and so on for x_i, t_j and x_j. Then, we can use any suitable method

Algorithm 1: Algorithm for statistically testing the presence of hidden confounding

Input: Datasets $D_k = \{(t_i^{(k)}, y_i^{(k)}, x_i^{(k)})\}_{i=1}^{N_k}$ from environments $k = 1, \dots, K$; significance level α; minimum number of environments required for hypothesis test $K_{\min}$

$L_{\max} \leftarrow \text{ceiling}(\max_k N_k/2)$ The maximum number of possible hypothesis tests

for $i = 1, \dots, L_{max}$ **do**

  $\pi_{2i} \leftarrow \{k \in [K] : N_k \leq 2i\}$ Retrieve all environments with at least $2i$ samples

  if $size(\pi_{2i}) < K_{min}$ **then**

    $L_{\max} \leftarrow i - 1$ Stop for-loop if we run out of a sufficient number of environments

    break

  $p_i \leftarrow \text{cond\_indep\_test}\,(t_{2i-1}^{\pi_{2i}} \perp\!\!\!\perp y_{2i}^{\pi_{2i}} \mid t_{2i}^{\pi_{2i}}, x_{2i-1}^{\pi_{2i}}, x_{2i}^{\pi_{2i}})$ Get p-value from independence test

$z \leftarrow -2\sum_{i=1}^{L_{\max}} \log(p_i)$ Aggregate p-values with Fisher's method

$p \leftarrow 1 - \text{cdf}_{\chi^2_{2L_{\max}}}(z)$ Compute global p-value

return $p \leq \alpha$

for conditional independence testing with t_i, y_i, x_i, t_j and x_j. Note that the choice of observations within each environment is arbitrary as long as we do not pick the same observation for i and j, this is a consequence of observations being i.i.d. within each environment.

Increasing power of test with Fisher's method In essence, we perform a conditional independence test where the "sample size" of the test is the number of available environments. Thus, for a small number of environments, our test might have low power (probability of detecting hidden confounding when it is present). To alleviate this issue, we recognize that we can perform this test multiple times if we have many samples N_k per environment. Then, we select new observations from every environment for each hypothesis test until all observations have been used up. It is important to note that we only select from environments where there still are observations that have not yet been used for the hypothesis testing. Since each hypothesis test is independent and has the same null, we can aggregate the p-values from all tests using Fisher's method to obtain a global hypothesis test (Fisher, 1925). As we show in our experiments, using Fisher's method drastically improves the power of our method, thus reducing the number of environments needed to detect the presence of hidden confounding. Having a different number of samples per environment N_k also necessitates specifying the hyperparameter $K_{\min}$, which determines the minimum observations required in each hypothesis test. This parameter should be chosen to ensure that the used independent testing method works properly if it is provided with at least $K_{\min}$ samples.

2.5. Experiments

To evaluate and investigate the theory for testing hidden confounding in multi-environment data, we perform a series of simulation studies with synthetic data in addition to experiments with semi-synthetic data generated using the Twins dataset (Almond *et al.*, 2005; Louizos *et al.*, 2017).[2] As we want to evaluate our method's ability to detect confounding, we use data where the ground-truth causal graph is known. Unless otherwise stated, each experiment is repeated 50 times where we use a significance level $\alpha = 0.05$. Depending on the variable types in the experiment, we state what suitable conditional independence testing method is used by our algorithm.

2.5.1. Synthetic data

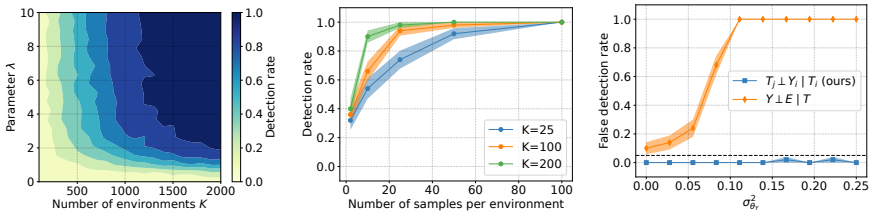
For the synthetic data experiments, we generate data as follows: we have the confounder $U_i^{(k)} \sim \text{Normal}(\Theta_U^{(k)}, 1)$; treatment $T_i^{(k)} \sim \text{Ber}(\text{Sigm}(U_i^{(k)} + \Theta_T^{(k)}))$; and outcome $Y_i^{(k)} \sim \text{Ber}(\text{Sigm}(\lambda U_i^{(k)} + T_i^{(k)} + \Theta_Y^{(k)}))$. Note $\text{Sigm}(x) = 1/(1+e^{-x})$ is the logistic function and $\Theta_V^{(k)} \sim \text{Normal}(0, \sigma^2_{\Theta_V})$ for $V \in \{T, Y, U\}$. Unless otherwise stated, we use $\sigma_{\Theta_T} = \sigma_{\Theta_U} = \sigma_{\Theta_Y} = 1$. We control the strength of confounding by varying λ, where $\lambda = 0$ corresponds to no confounding.

Larger confounder effect sizes increase the probability of detection We investigate how the effect size of the confounding variable influences our proposed testing procedure. We vary λ between 0 (no confounding) and 10 while also varying the number of environments. We perform this experiment with $N_k = 2$ for all k and use the G-test for conditional independence testing (McDonald, 2014). The results are shown in Figure 2.3a. We note two things: the probability of detection grows for larger confounder effect size and it also grows when the number of environments is increased.

The growth rate in detection depends on the number of environments We investigate the probability of detecting confounding when varying both the number of environments and the number of samples per environment for a fixed confounding strength $\lambda = 5$. We use a permutation-based method for the conditional independence test (Tsamardinos & Borboudakis, 2010) as we do not want to rely only on asymptotic validity (such as in the G-test) due to the limited number of environments. The results show that the performance of the testing procedure is highly dependent on the number of environments K, see 2.3b. The probability of detection grows as we increase the number of samples. Noticeably, the rate of growth increases with the number of environments K.

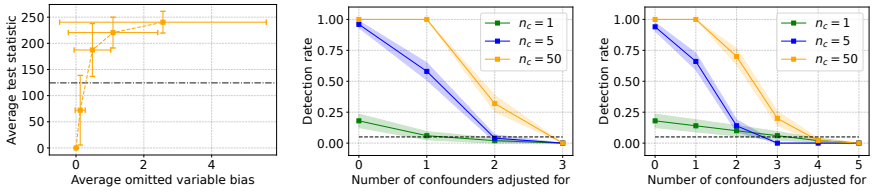
Robustness to environmental changes We compare our proposed procedure to the alternative approach of testing $Y \perp\!\!\!\perp_P E \mid T$ to detect hidden confounding, the latter being

[2] Code available at github.com/RickardKarl/detect-hidden-confounding.



(a) Varying λ. (b) $\lambda = 5$ (c) $\lambda = 0$ (no confounding)

Figure 2.3: **Synthetic data** – **(a)**: Detecting confounding with $N_k = 2$ across a range of confounder effect sizes and numbers of environments K. (500 repetitions) **(b)**: Simulations with fixed confounding strength $\lambda = 5$ for $N_k > 2$ with a small number of environments K. **(c)**: Comparing the proposed procedure and an alternative testing procedure by varying the standard deviation of Θ_Y in the absence of confounding. The black dashed line corresponds to the desired type 1 error control $\alpha = 0.05$. The shaded area shows the standard error from 50 repetitions.



(a) (b) $p = 3$ (c) $p = 5$

Figure 2.4: **Twins dataset** – **(a)**: Effect of bias from omitting confounders on the test statistic of our hypothesis test. **(b), (c)**: Performance when adjusting for observed confounders with either a total of 3 or 5 confounders in the data. The different curves correspond to combining numbers of hypothesis tests. The black dashed line corresponds to the rejection threshold / desired type 1 error $\alpha = 0.05$ in all figures, and the error bars / shaded area shows standard deviation (figure a) or standard error (figures b and c) from 50 repetitions.

valid when E is an instrumental variable (Mooij *et al.*, 2020). Here we test the sensitivity to violating one of its conditions, namely that $P_E(Y \mid T)$ is fixed under the null. We vary σ_{Θ_Y} between 0 and $\frac{1}{4}$ when there is no confounding by setting $\lambda = 0$ with $N = 100$ and $K = 500$, and we use the G-test for conditional independence testing (McDonald, 2014). As shown in Figure 2.3c, the probability of false detection using $Y \perp\!\!\!\perp_P E \mid T$ increases when σ_{Θ_Y} starts to increase. Meanwhile, the false detection rate (type 1 error) remains bounded by $\alpha = 0.05$ for our procedure as desired. In Appendix 2.F, we also include the same comparison when confounding is present to confirm that our method is able to detect confounding in this case.

2.5.2. Twins dataset

We use data from twin births in the USA between 1989-1991 (Almond *et al.*, 2005; Louizos *et al.*, 2017) to construct an observational dataset with continuous treatment/outcome and non-linear relationships. Here the environments are different states, and a notable element of our dataset is that all variations between environments stem solely from the real-world distribution shifts of the covariates between birth states. The strength of confounding is controlled by a parameter λ, where $\lambda = 0$ corresponds to no confounding. The full procedure for data generation is described in Appendix 2.E. For the following experiments, we use the Kernel Conditional Independence Test (K. Zhang *et al.*, 2011) in our algorithm due to having continuous variables and, unless otherwise stated, combine 50 hypothesis tests using Fisher's method.

Detection rate increases with bias from unobserved confounding We perform an experiment having $p = 5$ unobserved confounders, where we vary confounding strength λ between 0 and 5. We compute the bias from omitting the unobserved confounders when estimating the average treatment effect of T on Y in each environment. We then compare the average bias to the test statistic computed by our algorithm averaged over multiple iterations. As observed in Figure 2.4a, the test statistic increases together with the bias. The black dashed line in the figure represents the rejection threshold at $\alpha = 0.05$, hence we can see that for sufficient bias the method will detect it.

Adjusting for observed confounders In the last experiments, we attempt to detect hidden confounding while also adjusting for observed confounders. We go from observing none to all confounders while having a confounding strength of $\lambda = 5$. We do this for the case with either a total of $p = 3$ or $p = 5$ confounders, shown in Figure 2.4b and 2.4c, respectively. In addition, we investigate the influence of combining multiple hypothesis tests (n_c denotes the number of tests) using Fisher's method. We observe first that adjusting for more confounders leads to a decrease in detection rate, and that our desired type 1 error of $\alpha = 0.05$ is controlled when we have adjusted for all confounders. Secondly, the performance deteriorates when the total number of confounders increases, as indicated by the detection rate, which is lower when adjusting for 4 confounders when $p = 5$ than adjusting for 2 confounders when $p = 3$. This is likely because the conditional independence test loses power as the conditioning set becomes larger (K. Zhang *et al.*, 2011). Thirdly, we see that the combination of multiple hypothesis tests using Fisher's method does improve the power of our algorithm. We did, however, not see any significant benefit in combining more than 50 hypothesis tests in these experiments.

2.6. Discussion

In this work, we studied a setting where observational data has been collected from different heterogeneous environments in which the same treatment T, outcome Y, and

covariates X have been observed. We showed that assuming independent causal mechanisms, there exist testable conditional independencies that are violated in the presence of hidden confounders, for which we also proposed a statistical procedure to test these independencies from observed data. In many cases, with a sufficient number of environments, we show that we are able to detect confounding when it is present. While our main goal was to derive testable implications of hidden confounding, open questions remain on how to improve sample efficiency and tackle loss of power when adjusting for many observed confounders. Addressing these can lead to better tools for researchers to validate their causal assumptions and move towards making safer causal inferences.

Appendices

2.A. Future work

We have identified a set of open questions deriving from our work. Firstly, our theory applies to the common setting in causal inference with treatment T, outcome Y, and a possibly high-dimensional confounder X, in which we want to detect the presence of additional hidden confounding U (that also can be high-dimensional). While this is arguably the most typical setting in causal inference, it is of interest to consider other scenarios with more variables and interactions between them. We conjecture that other testable implications exist for confounding in these settings that could be found with similar arguments as we use. A particularly interesting setting is when we observe a proxy to a hidden confounder, which can be used for adjustment instead (Kuroki & Pearl, 2014; Miao *et al.*, 2018). In this case, it is no longer straightforward to say whether there could be a hidden confounder that is unrelated to the proxy. We also believe that our techniques might be applicable to scenarios with instrumental variables (IV) (Angrist *et al.*, 1996; Martens *et al.*, 2006), to test whether there exists any confounding between the IV and treatment which is a requirement for valid IV estimation.

Secondly, our theorem fundamentally relies on a set of untestable assumptions: independent & non-degenerate causal mechanisms and the faithfulness & causal Markov property. Although we investigated various violations, these results raised new questions. In particular, the effect of faithfulness violations, perhaps surprisingly, had a large influence on our procedure. Therefore, it is important to understand whether similar observations can be made in more realistic settings.

Thirdly, an interesting direction for future research would be to investigate how our approach can be used to estimate confounding strength, to be used in well-studied approaches in sensitivity analysis (Cinelli *et al.*, 2019; Rosenbaum & Rubin, 1983a).

Lastly, while our main goal was to derive testable implications of hidden confounding, there are opportunities to improve the way we test these from data. We observed that our proposed procedure sometimes requires a large number of environments. While it is unclear whether this is a property of the theory or the lack of efficiency in the test procedure we used, we note that combining multiple hypothesis tests using Fisher's method helped with performance. A possible research direction could be to investigate how to refine this approach further. Further, we observed performance deteriorating as we adjusted for more observed confounders, likely due to the curse of dimensionality (K. Zhang *et al.*, 2011). A promising solution here could be to use popular dimensionality-reduction techniques from causal inference such as the propensity score (Rosenbaum & Rubin, 1983b).

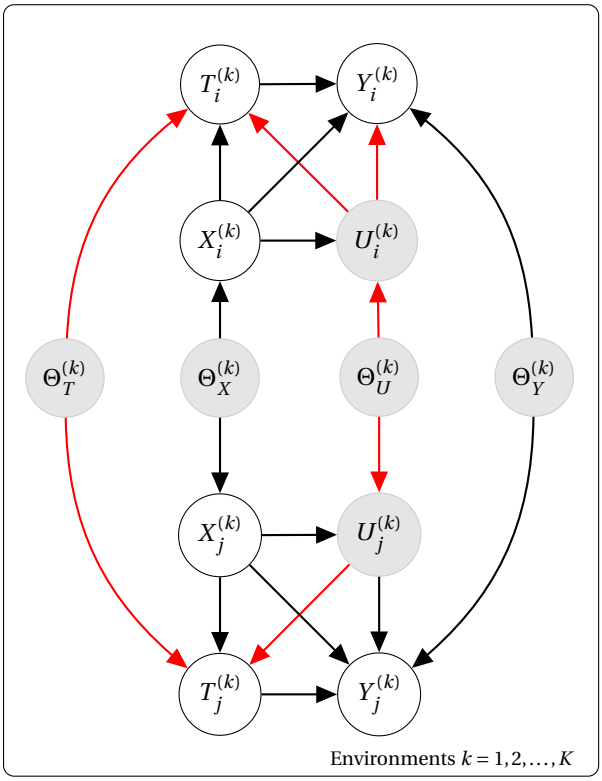


Figure 2.5: Example of unrolling the augmented causal DAG $\mathcal{G}^*$ for some pair of observations (i, j). The samples are generally not independent due to the shared mechanisms $(\Theta_T, \Theta_X, \Theta_U, \Theta_Y)$, unless we condition on them. Confounding is present, and the red edges mark open paths such that $T_j^{(k)} \not\perp\!\!\!\perp_d Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$ in this graph. Note that these paths go through the outgoing edges from $U_i^{(k)}$ to $T_i^{(k)}$ and $Y_i^{(k)}$ (and similarly for j), and that the paths are closed if either of the edges is removed.

2.B. Proofs

In this section, we present the proof for Theorem 2.1 and 2.2. Let $\mathbf{T}^{(k)} = (T_1^{(k)}, \ldots, T_{N_k}^{(k)})$ be the vector of all observed treatments in environments $E^{(k)}$. Define $\mathbf{Y}^{(k)}$, $\mathbf{X}^{(k)}$, and $\mathbf{U}^{(k)}$ similarly.

Theorem 1. *We consider the data distribution $P(\mathbf{T}^{(k)}, \mathbf{Y}^{(k)}, \mathbf{X}^{(k)}, \mathbf{U}^{(k)})$ with $N_k \geq 2$ under assumption 2.1,2.2, 2.3 and 2.4. Furthermore, assume an underlying causal DAG $\mathcal{G}$ where Y is not an ancestor of T, and that X is a known common cause to T and Y. Then, for any*

$k = 1, \ldots, K$, there exists hidden confounding between T and Y in $\mathcal{G}$ if and only if

$$T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)} \;\; \forall i, j = 1, \ldots, N_k : i \neq j \,. \tag{2.5}$$

Proof. We constrain ourselves to DAGs $\mathcal{G}$ with variables (T, Y, XU) where X is a known common cause to both T and Y, and Y is not an ancestor of T in $\mathcal{G}$. Under the assumption of non-degenerate, independent causal mechanisms (Assumption 2.3 and 2.4), we can introduce the mechanisms Θ_V for each variable $V \in \{T, Y, X, U\}$. Further, we can augment $\mathcal{G}$ with a hierarchical structure (Gelman *et al.*, 2013, Chapter 5), wherein we first sample the mechanisms i.i.d. $\Theta_V^{(k)} \sim P(\Theta_V)$ for $k = 1, \ldots, K$ and $V \in \{T, Y, X, U\}$ and then, for each environment k, obtain $(T_i^{(k)}, Y_i^{(k)}, X_i^{(k)}, U_i^{(k)})$ by repeatedly sampling N_k times conditioned on the mechanisms. We denote this augmented graph with the hierachical structure as $\mathcal{G}^*$, where edges between $(T_i^{(k)}, Y_i^{(k)}, X_i^{(k)}, U_i^{(k)})$ are the same as for (T, Y, X, U) and $\Theta_V^{(k)} \in \text{Pa}(V_i^{(k)})$ where $V \in \{T, Y, X, U\}$, for all i and k. An example of such an augmented graph $\mathcal{G}^*$ is shown in Figure 2.5. Notably, this augmentation can be done for all k because we assume that all environments share the same causal graph $\mathcal{G}$ (Asssumption 2.2).

Now, given the constraints that we have defined, we consider every combination of the edges between (T, Y, X, U) that are DAGs. In total, there are 40 different DAGs that encompass all these combinations of edges. We say that U is a confounder in one of these DAGs if both the edges $U \to T$ and $U \to Y$ exist. For each of these graphs $\mathcal{G}$, we shall investigate the d-separations in its augmented version $\mathcal{G}^*$. Notably, due to the assumption of non-degenerate mechanisms (Assumption 2.4), we allow open paths in $\mathcal{G}^*$ that go through $\Theta_T^{(k)}, \Theta_Y^{(k)}, \Theta_X^{(k)}$, or $\Theta_U^{(k)}$. This means that two different observations $\left(T_i^{(k)}, Y_i^{(k)}, X_i^{(k)}, U_i^{(k)}\right)$ and $\left(T_j^{(k)}, Y_j^{(k)}, X_j^{(k)}, U_j^{(k)}\right)$ can be dependent when $i \neq j$ for $i, j = 1, \ldots, N_k$. These paths are best illustrated by unrolling the augmented graph $\mathcal{G}^*$ as in the example in Figure 2.5. However, such dependencies can only happen if we do not condition on the mechanisms (that is, the environment) as we know that the observations are sampled i.i.d. within each environment. This demonstrates the need for multiple environments, as the randomness from sampling $\left(\Theta_T^{(k)}, \Theta_Y^{(k)}, \Theta_X^{(k)}, \Theta_U^{(k)}\right)$ allows us to treat them as ordinary random variables. To capture this randomness, we need to observe multiple environments though. Note that we also need $N_k \geq 2$ for $i \neq j$ to hold.

We will pay attention to a particular d-separation in $\mathcal{G}^*$, namely

$$T_j^{(k)} \perp\!\!\!\perp_d Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)} \,.$$

We automatically iterate over our list of DAGs using the *dagitty* package in R (Textor *et al.*, 2017) and check whether this d-separation holds; the results are displayed in Table 2.1. We note that the shaded rows in the table are the cases where U is a confounder, and these are the only cases where $T_j^{(k)} \perp\!\!\!\perp_d Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$ is violated in $\mathcal{G}^*$. In other words, checking this d-separation is sufficient to determine whether U is a confounder in $\mathcal{G}$. Assuming the faithfulness and causal Markov property (Assumption 2.1), we have that:

$$T_j^{(k)} \not\perp\!\!\!\perp_d Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)} \iff T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)} \,.$$

Consequently, it follows that

$$T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)} \text{ for } i \neq j \iff \text{U is a confounder to } T \text{ and } Y \,.$$

This result holds for any k since we assume that all environments share the same causal DAG (Assumption 2.2).

□

Remark 2.2. *$X_j^{(k)}$ can be removed from the conditioning set in the independence of Theorem 2.1 when we assume that the observed and unobserved confounders are independent of each other. In practice, however, we most likely would not like to make this assumption which is why we recommend to condition on both $X_i^{(k)}$ and $X_j^{(k)}$.*

Theorem 2. *We consider the data distribution $P(\mathbf{T}^{(k)}, \mathbf{Y}^{(k)}, \mathbf{U}^{(k)})$ without any observed confounders and $N_k \geq 2$ under assumption 2.1, 2.2, 2.3 and 2.4. Then, for any $k = 1, \ldots, K$, there exists hidden confounding between T and Y in $\mathcal{G}$ if and only if*

$$(i)\ T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)} \quad \textit{and} \quad (ii)\ T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid Y_j^{(k)} \ \forall i, j = 1, \ldots, N_k : i \neq j \,. \tag{2.6}$$

Proof. Using the same arguments as in the proof of Theorem 2.1, we can show that $T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}$ and $T_j^{(k)} \not\perp\!\!\!\perp_P Y_i^{(k)} \mid Y_j^{(k)}$ exclusively holds for those where there exists no common cause between T and Y when we only have the variables (T, Y, U), excluding any observed confounder X. The corresponding table to check this is demonstrated in Table 2.2. □

2.C. Alternative test for detecting hidden confounding in two-variable case

In this section, we explain when (and when not) testing $Y \perp\!\!\!\perp_P E \mid T$ is a valid strategy for detecting hidden confounding between T and Y. We require that $Y \perp\!\!\!\perp_P E \mid T, U$ when there is no confounding meaning that there can not exist any direct causal relationships $E \to Y$ or $E \to U$; if this does not hold then testing $Y \perp\!\!\!\perp_P E \mid T$ would not be informative. This is demonstrated with the dashed edges in Figure 2.6a. Meanwhile, in Figure 2.6b, we display the case where a violation of $Y \perp\!\!\!\perp_P E \mid T$ correctly detects hidden confounding being present. The conclusion is that this alternative test, which is easy to test, only works if the changes between environments happen for the conditional distribution $P_E(T \mid \text{Pa}(T))$, as only T may be directly caused by E if this test shall work. This corresponds to E being an instrumental variable (Angrist *et al.*, 1996), meaning that $T \not\perp\!\!\!\perp_P E$, $U \perp\!\!\!\perp_P E$, and the environment E can not have any direct effect on the outcome Y under the null (that is no confounding).

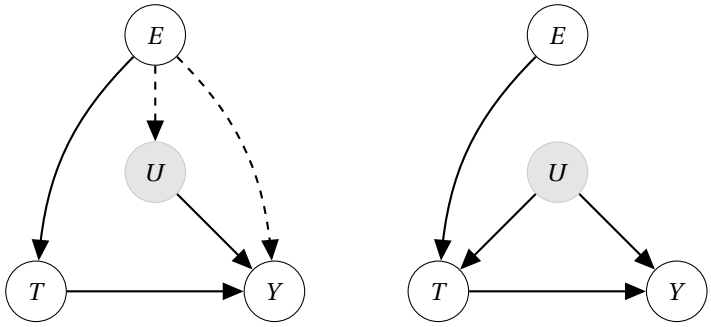


Figure 2.6: Examples of using $Y \perp\!\!\!\perp_P E \mid T$ to detect confounding. Gray corresponds to unobserved variables. **(a)** Violations of $Y \perp\!\!\!\perp_P E \mid T$ despite there being no confounding between T and Y. Removing the two dashed edges would make $Y \perp\!\!\!\perp_P E \mid T$ hold. **(b)** Case with correct violation of $Y \perp\!\!\!\perp_P E \mid T$ when unobserved confounding is present.

2.D. Further analysis on the influence of the assumptions

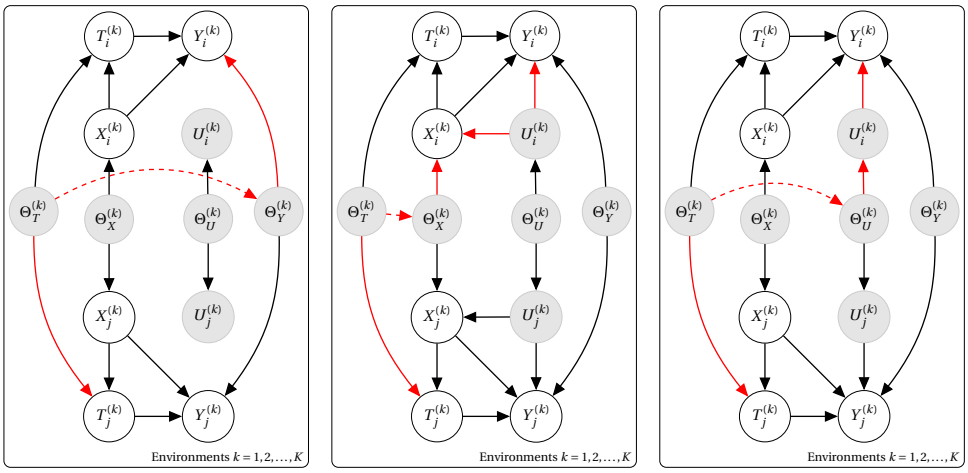
In this section, we present more elaboration on the examples of violating the assumptions behind our theory, as discussed in Section 2.4.1. First, we provide a counter-example that shows how Theorem 2.1 fails when we have dependent causal mechanisms. Secondly, we demonstrate how we got to our conclusions on having degenerate mechanisms. Thirdly, we provide insights into when faithfulness violations could occur in an example with a linear-Gaussian structural causal model. Finally, we discuss the analogy to the positivity assumption for our theory as well as how our procedure could be influenced by the presence of selection bias.

2.D.1. Dependent mechanisms

We wish to demonstrate examples of DAGs where the main condition

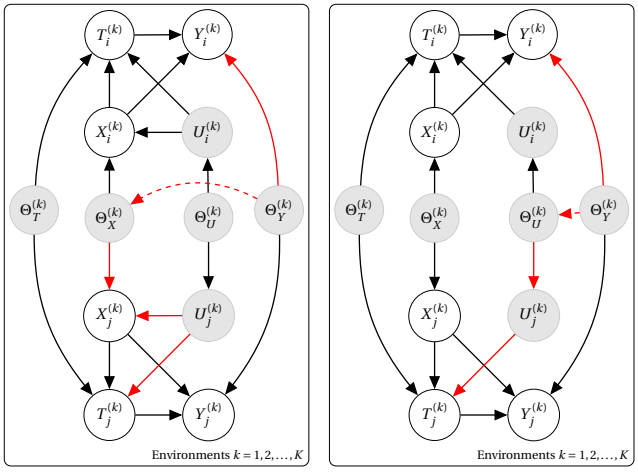
$$T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$$

in Theorem 2.1 is violated despite that there is no confounding when some of the causal mechanisms are pairwise dependent. We are able to find examples of violations with all pairwise dependencies, except for $\Theta_X \not\perp\!\!\!\perp_P \Theta_U$. Hence, we conjecture that a dependency between the mechanism of the observed and unobserved confounder does not influence the test proposed by our theory. Figures 2.7a, 2.7b, 2.7c, 2.8a and 2.8b show found examples where violations occur between the mechanisms.



(a) $\Theta_T \not\perp\!\!\!\perp_P \Theta_Y$ (b) $\Theta_T \not\perp\!\!\!\perp_P \Theta_X$ (c) $\Theta_T \not\perp\!\!\!\perp_P \Theta_U$

Figure 2.7: Violations of $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$ with dependent mechanisms despite that X is a valid adjustment set for estimating $P(Y \mid do(T))$. Open paths between $T_j^{(k)}$ and $Y_i^{(k)}$ after conditioning on T_i, X_i and X_j are marked in red.



(a) $\Theta_Y \not\perp\!\!\!\perp_P \Theta_X$ (b) $\Theta_Y \not\perp\!\!\!\perp_P \Theta_U$

Figure 2.8: Violations of $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$ with dependent mechanisms despite that X is a valid adjustment set for estimating $P(Y \mid do(T))$. Open paths between $T_j^{(k)}$ and $Y_i^{(k)}$ after conditioning on T_i, X_i and X_j are marked in red.

2.D.2. Degenerate mechanisms

What happens if one or more of the distributions $P(\Theta_T)$, $P(\Theta_Y)$, $P(\Theta_X)$, and $P(\Theta_U)$ are constant across all environments? We investigate these scenarios by first adding Θ_T, Θ_Y, Θ_X and/or Θ_U to the conditioning set in $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$ - the testable implication for confounding. We then follow the same procedure as for proving Theorem 2.1 and noting whether this conditional independence still discriminates between cases when U is and is not a confounder. Out of 15 possible cases with degenerate mechanisms, we identify that the theorem fails every time we condition on both Θ_T and Θ_U, as demonstrated in Table 2.3.

2.D.3. Faithfulness violation

What happens if the conditional independencies we test do not correspond to the dependencies in the underlying causal graph? If that is the case, we would fail to detect dependencies implying the presence of confounders. Knowing when a faithfulness violation occurs is not possible, we can only reason about its plausibility. In this section, we first showcase an example where a faithfulness violation occurs in linear-Gaussian structural causal models for particular configurations as well as when the confounder effect sizes – that is γ and λ in eq. (2.7) – become very large.

It was proved by Meek (1995) that all graphs with discrete and linear-Gaussian data distributions fulfill faithfulness in a measure-theoretic sense; distributions that violate faithfulness have measure zero.

However, even if we restrict T and Y to be categorical we might not want to assume that U or the causal mechanisms follow a discrete distribution. The following example also demonstrates the practical issues stemming from faithfulness violations even when the data is jointly Gaussian.

Example 2.1. *Let $k = 1, \ldots, K$ and $i = 1, \ldots, N_k$. Consider the structural causal model*

$$\begin{aligned} U_i^{(k)} &= \Theta_U^{(k)} + \varepsilon_{U,i}, & \varepsilon_{U,i} &\sim \text{Normal}(0, \sigma_U^2), \\ T_i^{(k)} &= \gamma U_i^{(k)} + \Theta_T^{(k)} + \varepsilon_{T,i}, & \varepsilon_{T,i} &\sim \text{Normal}(0, \sigma_T^2), \\ Y_i^{(k)} &= \lambda U_i^{(k)} + \beta T_i^{(k)} + \Theta_Y^{(k)} + \varepsilon_{Y,i}, & \varepsilon_{Y,i} &\sim \text{Normal}(0, \sigma_Y^2), \end{aligned} \tag{2.7}$$

where $\Theta_V^{(k)} \sim \text{Normal}(0, \sigma_{\Theta_V}^2)$ and $\varepsilon_V \perp\!\!\!\perp_P \Theta_V^{(k)}$ for $V \in \{T, Y, U\}$. Further, subscript i for the noise variables indicates that they are sampled independently for each observation i. Then, despite the presence of confounding, $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}$ for any $i \neq j$ when $\sigma_{\Theta_U} = \frac{\sigma_U}{\sigma_T}\sigma_{\Theta_T}$. In the finite-sample setting, it noticeably influences our ability to detect confounding when the distribution parameters come close to this equality, as illustrated in Figure 2.9.

To create Figure 2.9, we generate data according to eq. (2.7) with the following parameters fixed: $\beta = \gamma = \lambda = 1$, $\sigma_{\Theta_Y} = 1$, $\sigma_Y = \sigma_U = 1$ and $\sigma_T = \frac{2}{3}$. Meanwhile, we vary σ_{Θ_T} and σ_{Θ_U}

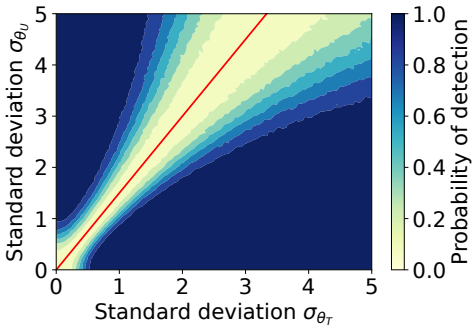


Figure 2.9: Faithfulness violation from Example 2.1. Varying σ_{Θ_T} and σ_{Θ_U} with $\sigma_T = \frac{2}{3}$ and $\sigma_U = 1$, the probability for detecting confounding goes to zero as we get closer to $\sigma_{\Theta_U} = \frac{\sigma_U}{\sigma_T}\sigma_{\Theta_T}$ (red line).

between 0 and 5. We test $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}$ using the partial correlation (Baba *et al.*, 2004). The experiment is repeated 1000 times with 1000 environments and a significance level $\alpha = 0.05$.

2.D.4. Derivation of example 2.1

We consider the structural causal model in eq. (2.7). Now, we want to prove that $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}$ for any $i \neq j$ when $\sigma_{\Theta_U} = \frac{\sigma_U}{\sigma_T}\sigma_{\Theta_T}$. For ease of notation, we will drop the superscript (k), as the results hold for any k.

Crucially, we note that the partial correlation

$$\rho_{T_j,Y_i \cdot T_i} = \frac{\rho_{T_j,Y_i} - \rho_{T_j,T_i}\rho_{T_i,Y_i}}{\sqrt{1-\rho_{T_j,T_i}^2}\sqrt{1-\rho_{T_i,Y_i}^2}}\,, \tag{2.8}$$

is zero if and only if $T_j \perp\!\!\!\perp_P Y_i \mid T_i$ when the data is jointly Gaussian (Baba *et al.*, 2004), which is the case for eq. (2.7) because $p(T_i, Y_i, U_i) = P(Y_i \mid T_i, U_i)P(T_i \mid U_i)P(U_i)$ where each factor is a Gaussian density.

To check when the partial correlation is zero, we need to find out when

$$\rho_{T_j,Y_i} - \rho_{T_j,T_i}\rho_{T_i,Y_i} = 0\,.$$

Since $\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$ for some random variables X and Y, we can write this as

$$\rho_{T_j,Y_i} - \rho_{T_j,T_i}\rho_{T_i,Y_i} = \frac{\mathrm{Cov}(T_j,Y_i)}{\sqrt{\mathrm{Var}(T_j)\mathrm{Var}(Y_i)}} - \frac{\mathrm{Cov}(T_j,T_i)}{\sqrt{\mathrm{Var}(T_j)\mathrm{Var}(T_i)}}\frac{\mathrm{Cov}(T_i,Y_i)}{\sqrt{\mathrm{Var}(T_i)\mathrm{Var}(Y_i)}} \tag{2.9}$$

$$= \frac{\mathrm{Var}(T_i)\mathrm{Cov}(T_j,Y_i) - \mathrm{Cov}(T_j,T_i)\mathrm{Cov}(T_i,Y_i)}{\sqrt{\mathrm{Var}(T_i)^3\mathrm{Var}(Y_i)}}\,, \tag{2.10}$$

where we used the fact that $\mathrm{Var}(T_i) = \mathrm{Var}(T_j)$ for any samples i and j.

First, we need to determine all the (co)variances, for which we need to know

$$\begin{aligned} T_i &= \gamma\Theta_U + \gamma\varepsilon_{U,i} + \Theta_T + \varepsilon_{T,i} \\ Y_i &= \lambda\Theta_U + \lambda\varepsilon_{U,i} + \beta\gamma\Theta_U + \beta\gamma\varepsilon_{U,i} + \beta\Theta_T + \beta\varepsilon_{T,i} + \Theta_Y + \varepsilon_{Y,i} \end{aligned}$$

Note that $\mathbb{E}[T_i] = \mathbb{E}[Y_i] = 0$. Consequently, we can write out the covariances, for any i, j, as follows:

$$\begin{aligned} \mathrm{Cov}(T_j,Y_i) &= \mathbb{E}[T_jY_i] = (\gamma\lambda + \beta\gamma^2)\mathbb{E}[\Theta_U^2] + \beta\mathbb{E}[\Theta_T^2] \\ \mathrm{Cov}(T_j,T_i) &= \mathbb{E}[T_jT_i] = \gamma^2\mathbb{E}[\Theta_U^2] + \mathbb{E}[\Theta_T^2] \\ \mathrm{Cov}(T_i,Y_i) &= \mathbb{E}[T_iY_i] = (\gamma\lambda + \beta\gamma^2)\mathbb{E}[\Theta_U^2] + (\gamma\lambda + \beta\gamma^2)\mathbb{E}[\varepsilon_U^2] + \beta\mathbb{E}[\Theta_T^2] + \beta\mathbb{E}[\varepsilon_T^2] \\ \mathrm{Var}(T_i) &= \mathbb{E}[T_iT_i] = \gamma^2\mathbb{E}[\Theta_U^2] + \gamma^2\mathbb{E}[\varepsilon_U^2] + \mathbb{E}[\Theta_T^2] + \mathbb{E}[\varepsilon_T^2] \\ \mathrm{Var}(Y_i) &= \mathbb{E}[Y_iY_i] = 2(\lambda^2 + \beta^2\gamma^2)\mathbb{E}[\Theta_U^2] + 2(\lambda^2 + \beta^2\gamma^2)\mathbb{E}[\varepsilon_U^2] + \beta^2\mathbb{E}[\Theta_T^2] \\ &\qquad + \beta^2\mathbb{E}[\varepsilon_T^2] + \mathbb{E}[\Theta_Y^2] + \mathbb{E}[\varepsilon_Y^2] \end{aligned}$$

Now, we look at the numerator in eq. (2.10) and we want to know when it could be zero since that makes the partial correlation zero:

$$\begin{aligned} 0 &= \mathrm{Var}(T_i)\mathrm{Cov}(T_j,Y_i) - \mathrm{Cov}(T_j,T_i)\mathrm{Cov}(T_i,Y_i) \\ &= \gamma\lambda(\mathbb{E}[\Theta_U^2]\mathbb{E}[\varepsilon_T^2] - \mathbb{E}[\varepsilon_U^2]\mathbb{E}[\Theta_T^2]) \end{aligned}$$

The solution is given by

$$\sigma_{\Theta_U} = \frac{\sigma_U}{\sigma_T}\sigma_{\Theta_T}\,, \tag{2.11}$$

where the square root of the second moments are equal to the standard deviations. This is the same equality as demonstrated in the example.

2.D.5. Asymptotic behavior of partial correlation

We also look at the partial correlation and ask what happens when the confounder effect sizes γ or λ become very large. The numerator in eq. (2.10) grows linearly with respect to

both γ and λ, and the other variances can be rewritten as

$$\begin{aligned}
\text{Cov}(T_j, T_i) &= \gamma^2 \mathbb{E}[\Theta_U^2] + O(1) \\
\text{Cov}(T_i, Y_i) &= (\gamma\lambda + \beta\gamma^2)(\mathbb{E}[\Theta_U^2] + \mathbb{E}[\varepsilon_U^2]) + O(1) \\
\text{Var}(T_i) &= \gamma^2 (\mathbb{E}[\Theta_U^2] + \mathbb{E}[\varepsilon_U^2]) + O(1) \\
\text{Var}(Y_i) &= 2(\lambda^2 + \beta^2\gamma^2)(\mathbb{E}[\Theta_U^2] + \mathbb{E}[\varepsilon_U^2]) + O(1)
\end{aligned}$$

where $O(1)$ is a constant with respect to γ and λ.

We rewrite the partial correlation eq. (2.8) as

$$\rho_{T_j, Y_i \cdot T_i} = \frac{\left(\text{Var}(T_i)\text{Cov}(T_j, Y_i) - \text{Cov}(T_j, T_i)\text{Cov}(T_i, Y_i)\right) / \sqrt{\text{Var}(T_i)^3 \text{Var}(Y_i)}}{\sqrt{1 - \frac{\text{Cov}(T_j, T_i)^2}{\text{Var}(T_i)^2}} \sqrt{1 - \frac{\text{Cov}(T_i, Y_i)^2}{\text{Var}(T_i)\text{Var}(Y_i)}}} .$$

Assuming that all second moments are non-zero and finite, it is possible to show that

$$\rho_{T_j, Y_i \cdot T_i} \propto \begin{cases} \gamma^{-3} & \text{for } |\gamma| >> 1, \\ 1 & \text{for } |\lambda| >> 1 \end{cases} .$$

Hence, when either $|\gamma|$ or $|\lambda|$ goes to infinity we have

$$\rho_{T_j, Y_i \cdot T_i} \underset{|\gamma| \to \infty}{\longrightarrow} 0$$

$$\rho_{T_j, Y_i \cdot T_i} \underset{|\lambda| \to \infty}{\longrightarrow} C \text{ for some } C \in [-1, 1]$$

Note that C could be zero, for instance, when $\sigma_{\Theta_U} = \frac{\sigma_U}{\sigma_T} \sigma_{\Theta_T}$, although we demonstrate with simulation studies in Appendix 7.D a case where C is non-zero as well.

Interestingly, in this case, the bias from estimating the causal effect without adjusting for the confounder U is

$$\mathbb{E}[Y \mid do(T)] - \mathbb{E}[Y \mid T] = \beta T - (\beta T + \frac{\lambda}{\gamma} T) = -\frac{\lambda}{\gamma} T . \tag{2.12}$$

We note that when $\gamma \to \infty$ the bias goes to zero, similar to the partial correlation. Meanwhile, the bias increases with λ which also is consistent with the asymptotic behavior of the partial correlation as $\lambda \to \infty$.

2.D.6. Positivity violations in the sampling of mechanisms

In this section, we discuss another potential issue that can come up in our problem setting, namely that there could be positivity violations in the sampling of the mechanisms. Particularly unique to our setting is that the support for $(\Theta_T, \Theta_Y, \Theta_X, \Theta_U)$ must be the same for different environments, if not then this would be a direct violation of Assumption 2.4.

We start by pointing out that there exist two categories of positivity violations: structural violations and random violations (Hernan & Robins, 2023). Structural violations occur for instance when a certain range of values of a variable never will be observed, and they may restrict the population for which we can draw causal conclusions. Meanwhile, random violations are due to having a finite number of samples. Random violations are perhaps also less problematic, as they can go away as we collect more data.

In our setting with multi-environment data, we can make the same analogy with structural and random positivity violations. For instance, a structural violation could occur if we have only collected data from multiple hospitals in country A but there is also another country B such that $P(\Theta_T, \Theta_Y, \Theta_X, \Theta_U \mid \text{country A})$ does not overlap with $P(\Theta_T, \Theta_Y, \Theta_X, \Theta_U \mid \text{country B})$. Then, we have a structural violation between environments in countries A and B. Meanwhile, a random positivity violation could come from not observing enough environments in country A, assuming that we are only interested in studying data from that country.

So based on what we have discussed so far, one could ask how to reason about positivity violations when trying to detect hidden confounding. In this case, we think it is important to first ask ourselves: In what population are we trying to detect hidden confounding? If the answer is the population in country A, then we should not include data from country B, as there is a structural positivity violation; and vice versa. For random violations, it is harder to anticipate what problems can come up, but these issues are mainly avoided by ensuring that we have sampled enough environments. This is what, in the end, will affect the quality of the conditional independence test in our method.

2.D.7. Selection bias

In our theory, we assume there is no selection bias. That is, for instance, that there are no unobserved colliders that we have conditioned on in our causal DAG. In principle, we can study such scenarios by adding colliders in our original problem setup to introduce different types of selection mechanisms. While we leave this topic for future work, we present here a quick illustration of how selection bias can (or can not) hurt our procedure.

Example 2.2. *Consider the graph $\mathcal{G}$ in Figure 2.10 where C is a collider between treatment T and outcome Y. Now we consider the corresponding augmented DAG $\mathcal{G}^*$, add the causal mechanism $\Theta_C \sim P(\Theta_C)$ as parent to C, and check how (unknowingly) conditioning on C influences our ability to detect the presence of the unobserved U. The fact that Θ_C is a random variable would reflect that we have different selection mechanisms in different environments. We note in this case that the conditional independence $T_j^{(k)} \perp\!\!\!\not\perp_P Y_i^{(k)} \mid T_i^{(k)}, C_i^{(k)}, C_j^{(k)}$ will be violated regardless of whether U is present or not. In other words, selection bias would in this case lead to false positives of our procedure. This is because there will always be an open path between $T_j^{(k)}$ and $Y_i^{(k)}$ through Θ_C in $\mathcal{G}^*$. But if, on the other hand, the selection mechanism remains fixed across all environments (meaning Θ_C is constant), this would close that path. That means we do not have this problem of false*

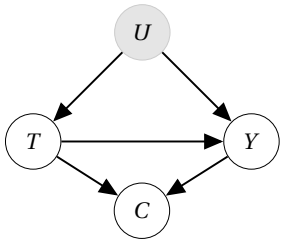


Figure 2.10: Grey variables are unobserved.

detection anymore, and the proposed approach would still work.

2.E. Generation of Twins semi-synthetic dataset

We use data from twin births in the USA between 1989-1991 (Almond *et al.*, 2005; Louizos *et al.*, 2017) to construct an observational dataset with a known causal structure. The dataset contains 46 covariates related to pregnancy, birth, and parents, out of which we select ten as potential confounders in our experiments. Many covariates are highly imbalanced and have low variance, some even on the border of being constant. Due to this, we wish to exclude those in data generation. The selection is performed by first excluding any binary variables from the list of covariates, and then further removing the remaining ones that have an empirical variance smaller than one.

In the end, the following covariates are used from the Twins dataset for our data generation: birth month, father's age, mom's age, mom's education, mom's place of birth, number of prenatal visits, number of live births before twins, the total number of births before twins, period of gestation, state of birth occurrence, and state of registered residence. Note that all of these are reported as categorical/discrete variables in the dataset.

Among the covariates, we use the state in which the birth took place as environment label, hence obtaining 51 different environments. We simulate a continuous treatment and outcome by randomly selecting $p \in [1,\dots,10]$ features $(X_1, X_2, \dots, X_p)$ and a set of functions to generate the data as follows:

$$T = \sum_{d=1}^{p} \alpha_d f_d(X_{d,\text{scaled}}) + \varepsilon_T \text{ and } Y = \sum_{d=1}^{p} \beta_d g_d(X_{d,\text{scaled}}) + \delta T + \varepsilon_Y \tag{2.13}$$

For each d, α_d is sampled from a uniform distribution Unif$(1,5)$, β_d is sampled from a uniform distribution Unif$(1,5)$, and f_d and g_d are sampled from the set of functions $\{\tanh(x), x, x^2\}$ with equal probability. The treatment effect, δ, is also randomly sampled from a uniform distribution Unif$(1,2)$, and we have noise variables $\varepsilon_T \sim \text{Normal}(0, 1/4)$ and $\varepsilon_Y \sim \text{Normal}(0, 1/4)$. The features from the Twins dataset are also scaled as

$$X_{d,\text{scaled}} = 5 * (X_d - \text{mean}(X_d))/(\max(X_d) - \min(X_d)) \,, \tag{2.14}$$

where mean/max/min are taken over all observed values of the covariate X_d. Note that the functions are fixed across all environments in this setup, and variations between environments only stem from the real-world distribution shifts of the covariates between birth states.

2.F. Additional experiments

We present additional simulation studies, mainly replicating the experiments on synthetic data from Section 2.5.1 with continuous data. In addition, we further investigate the asymptotic behavior of the partial correlation from Appendix 2.D.3 with both the binary and continuous data.

The continuous data is generated from a linear-Gaussian DAG as described in eq. (2.7). Unless otherwise stated, we use $\beta = 1$, $\sigma_T = \sigma_U = \sigma_Y = 1$, $\sigma_{\Theta_T} = \sigma_{\Theta_Y} = 1$ and $\sigma_{\Theta_T} = 5$. To vary the influence of the hidden confounder, we can adjust either γ or λ. We test $T_j^{(k)} \perp\!\!\!\perp_P Y_i^{(k)} \mid T_i^{(k)}$ using the partial correlation (Baba *et al.*, 2004) with $N_k = 2$ samples per environment.

For the first experiment, we vary the number of environments and the confounder influence by setting $\gamma = \lambda$. In the main part of the paper, we only considered what happens when varying λ with $\gamma = 1$. Similarly, we do the same experiment with the binary data for comparison. The results are seen in Figure 2.11. Notably, the probability of detecting hidden confounding starts decreasing when $\gamma = \lambda$ goes above a certain threshold. This is consistent with our previous conclusions from Appendix 2.D.3, where we noted that partial correlation is proportional to γ^{-3} for $\gamma >> 1$ while remaining constant for $\lambda >> 1$. Hence, we would expect the partial correlation to shrink as both γ and λ grow. Notably, the effect is more pronounced with the continuous data although it can also be seen with the binary data.

Secondly, we perform the experiment with continuous data where we only vary λ while fixing $\gamma = 1$. The results are shown in Figure 2.12, we note that the probability of detection no longer decreases as the confounder effect size increases. The results with binary data are also shown again for comparison. Once again, this is predicted by the asymptotic behavior of the partial correlation.

Finally, we compare our statistical testing procedure to testing $Y \perp\!\!\!\perp_P E \mid T$ with continuous data in Figure 2.13. We run the experiment with 10000 environments, 100 samples per environment, and $\sigma_{\Theta_U} = 10$ to avoid the issues of faithfulness violations which we have discussed before. Similar to the case with binary data, in Figure 2.14b, we see that the probability of false detection when testing $Y \perp\!\!\!\perp_P E \mid T$ grows as the standard deviation of Θ_Y increases. Meanwhile, we do not observe this problem in our testing procedure. The case when confounding is present is shown in Figure 2.13a to confirm that our method is able to detect confounding. For completion, we also include a case where $\lambda = 10$ (confounding is present) for the binary data experiment presented in the main part of the paper, this is shown in Figure 2.14a.

Table 2.1: Conditional d-separations in combinations of DAGs with variables (T, Y, X, U). For the d-separation, (✓) indicates that it holds and (✗) otherwise. The shaded rows are the cases where U is a confounder to T and Y in $\mathcal{G}$.

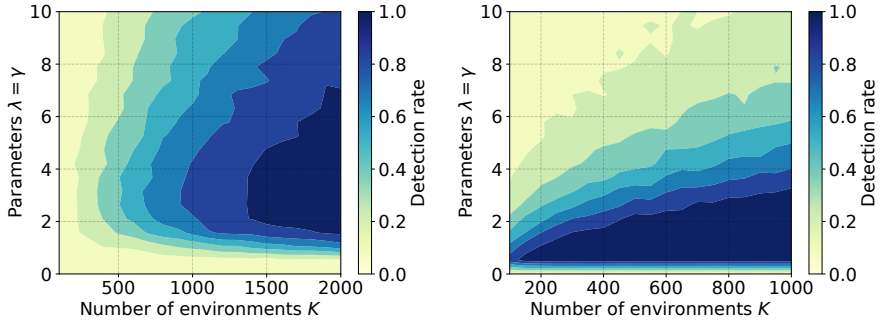
id	$X-T$	$X-Y$	$T-Y$	$U-T$	$U-Y$	$U-X$	$T_j^{(k)} \perp_d Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}$	U is confounder
1	→	→	→	→	→	→	✗	✓
2	→	→	→	→	→	←	✗	✓
3	→	→	→	→	→		✗	✓
4	→	→	→	→		→	✓	✗
5	→	→	→	→		←	✓	✗
6	→	→	→	→			✓	✗
7	→	→	→	←	→	←	✓	✗
8	→	→	→	←	→		✓	✗
9	→	→	→	←	←	←	✓	✗
10	→	→	→	←	←		✓	✗
11	→	→	→	←		←	✓	✗
12	→	→	→	←			✓	✗
13	→	→	→		→	→	✓	✗
14	→	→	→		→	←	✓	✗
15	→	→	→		→		✓	✗
16	→	→	→		←	←	✓	✗
17	→	→	→		←		✓	✗
18	→	→	→			→	✓	✗
19	→	→	→			←	✓	✗
20	→	→	→				✓	✗
21	→	→		→	→	→	✗	✓
22	→	→		→	→	←	✗	✓
23	→	→		→	→		✗	✓
24	→	→		→		→	✓	✗
25	→	→		→		←	✓	✗
26	→	→		→			✓	✗
27	→	→		←	→	←	✓	✗
28	→	→		←	→		✓	✗
29	→	→		←	←	←	✓	✗
30	→	→		←	←		✓	✗
31	→	→		←		←	✓	✗
32	→	→		←			✓	✗
33	→	→			→	→	✓	✗
34	→	→			→	←	✓	✗
35	→	→			→		✓	✗
36	→	→			←	←	✓	✗
37	→	→			←		✓	✗
38	→	→				→	✓	✗
39	→	→				←	✓	✗
40	→	→					✓	✗

Table 2.2: Conditional d-separations in combinations of DAGs with variables (T, Y, U) (this excludes any observed confounder X). For the d-separation, (✓) indicates that it holds and (✗) otherwise. The shaded rows are the cases where U is a confounder to T and Y in $\mathcal{G}$.

id	$T-Y$	$U-T$	$U-Y$	$T_j^{(k)} \perp_d Y_i^{(k)} \mid T_i^{(k)}$	$T_j^{(k)} \perp_d Y_i^{(k)} \mid Y_j^{(k)}$	U is confounder	Y ancestor of T
1	→	→	→	✗	✗	✓	✗
2	→	→		✓	✗	✗	✗
3	→	←	→	✓	✗	✗	✗
4	→	←	←	✓	✗	✗	✗
5	→	←		✓	✗	✗	✗
6	→		→	✓	✗	✗	✗
7	→		←	✓	✗	✗	✗
8	→			✓	✗	✗	✗
9	←	→	→	✗	✗	✓	✓
10	←	→	←	✗	✓	✗	✓
11	←	→		✗	✓	✗	✓
12	←	←	←	✗	✓	✗	✓
13	←	←		✗	✓	✗	✓
14	←		→	✗	✓	✗	✓
15	←		←	✗	✓	✗	✓
16	←			✗	✓	✗	✓
17		→	→	✗	✗	✓	✗
18		→	←	✗	✓	✗	✓
19		→		✓	✓	✗	✗
20		←	→	✓	✗	✗	✗
21		←	←	✓	✓	✗	✗
22		←		✓	✓	✗	✗
23			→	✓	✓	✗	✗
24			←	✓	✓	✗	✗
25				✓	✓	✗	✗

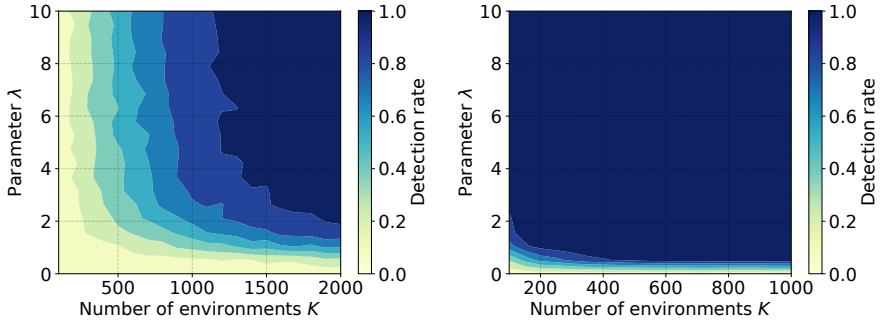
Table 2.3: Conditional d-separations in combinations of DAGs with variables (T, Y, X, U). We test $T_j^{(k)} \perp\!\!\!\perp_d Y_i^{(k)} \mid T_i^{(k)}, X_i^{(k)}, X_j^{(k)}, Z$ where Z contains the mechanisms that are degenerate, as noted per column. For the d-separation, (✓) indicates that it holds and (✗) otherwise. The shaded rows are the cases where U is a confounder to T and Y in $\mathcal{G}$.

id	$X-T$	$X-Y$	$T-Y$	$U-T$	$U-Y$	$U-X$	U is confounder	Deg. θ_T	Deg. θ_Y	Deg. θ_X	Deg. θ_U	Deg. θ_T,θ_Y	Deg. θ_T,θ_X	Deg. θ_T,θ_U	Deg. θ_Y,θ_X	Deg. θ_Y,θ_U	Deg. θ_X,θ_U	Deg. $\theta_T,\theta_Y,\theta_X$	Deg. $\theta_T,\theta_Y,\theta_U$	Deg. $\theta_T,\theta_X,\theta_U$	Deg. $\theta_Y,\theta_X,\theta_U$	Deg. $\theta_T,\theta_Y,\theta_X,\theta_U$
1	→	→	→	→	→	→	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓
2	→	→	→	→	→	←	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓
3	→	→	→	→	→		✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓
4	→	→	→	→		→	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	→	→	→	→		←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	→	→	→	→			✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7	→	→	→	←	→	←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
8	→	→	→	←	→		✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
9	→	→	→	←	←	←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	→	→	→	←	←		✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
11	→	→	→	←		←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
12	→	→	→	←			✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
13	→	→	→		→	→	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
14	→	→	→		→	←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
15	→	→	→		→		✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
16	→	→	→		←	←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
17	→	→	→		←		✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
18	→	→	→			→	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
19	→	→	→			←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
20	→	→	→				✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
21	→	→		→	→	→	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓
22	→	→		→	→	←	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓
23	→	→		→	→		✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓
24	→	→		→		→	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
25	→	→		→		←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
26	→	→		→			✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
27	→	→		←	→	←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
28	→	→		←	→		✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
29	→	→		←	←	←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
30	→	→		←	←		✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
31	→	→		←		←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
32	→	→		←			✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
33	→	→			→	→	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
34	→	→			→	←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
35	→	→			→		✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
36	→	→			←	←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
37	→	→			←		✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
38	→	→				→	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
39	→	→				←	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
40	→	→					✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



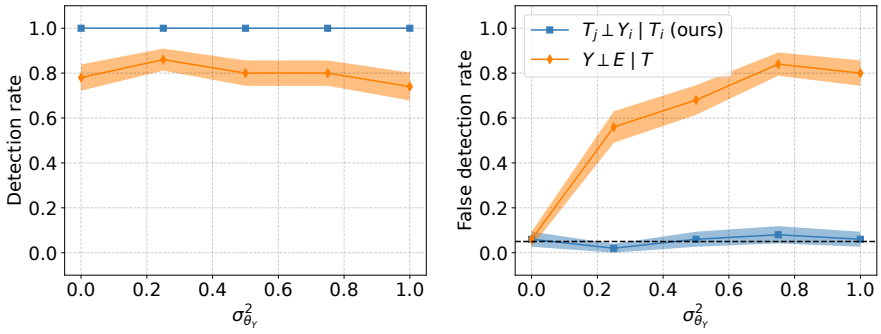
(a) Binary data (b) Continuous data

Figure 2.11: Probability of detecting hidden confounding for varying the number of environments K and confounder effect sizes where $\lambda = \gamma$ are varied jointly. 500 repetitions.



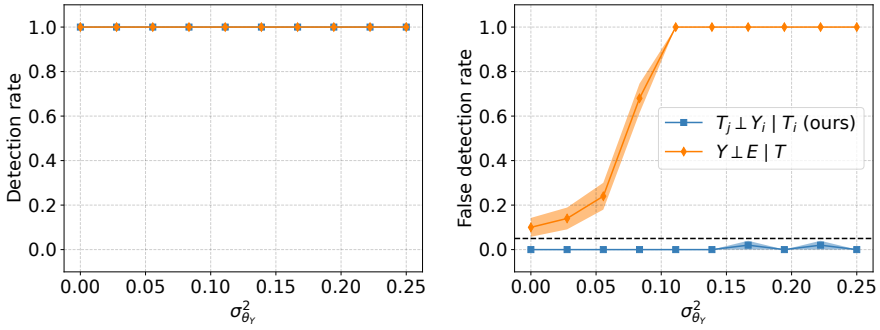
(a) Binary data (b) Continuous data

Figure 2.12: Probability of detecting hidden confounding for varying the number of environments K and confounder effect sizes where λ is varied while $\gamma = 1$ is fixed. 500 repetitions



(a) Confounding is present ($\lambda = 10$) (b) No confounding is present ($\lambda = 0$)

Figure 2.13: Comparison on continuous linear-Gaussian data between the proposed procedure and an alternative testing procedure by varying the standard deviation of Θ_Y in both the presence and absence of confounding. The black dashed line corresponds to the desired type 1 error control $\alpha = 0.05$. The shaded area shows the standard error from 50 repetitions.



(a) Confounding is present ($\lambda = 10$) (b) No confounding is present ($\lambda = 0$)

Figure 2.14: Comparison on binary data between the proposed procedure and an alternative testing procedure by varying the standard deviation of Θ_Y in both the presence and absence of confounding. The black dashed line corresponds to the desired type 1 error control $\alpha = 0.05$. The shaded area shows the standard error from 50 repetitions.

3

Efficient Falsification Through Testing Independence of Causal Mechanisms

A major challenge in estimating treatment effects in observational studies is the reliance on untestable conditions such as the assumption of no unmeasured confounding. In this work, we propose an algorithm that can falsify the assumption of no unmeasured confounding in a setting with observational data from multiple heterogeneous sources, which we refer to as environments. Our proposed falsification strategy leverages a key observation that unmeasured confounding can cause observed causal mechanisms to appear dependent. Building on this observation, we develop a novel two-stage procedure that detects these dependencies with high statistical power while controlling false positives. The algorithm does not require access to randomized data and, in contrast to other falsification approaches, functions even under transportability violations when the environment has a direct effect on the outcome of interest. To showcase the practical relevance of our approach, we show that our method is able to efficiently detect confounding on both simulated and semi-synthetic data.

3.1. Introduction

Using observational studies to estimate treatment effects is a ubiquitous yet challenging task in many disciplines, such as medicine (Hernán & Robins, 2006) or social sciences (Athey & Imbens, 2017). Whereas there exists a rich literature of methods for treatment effect estimation in the observational setting (Bang & Robins, 2005; Chernozhukov

This chapter appears as: Karlsson, R., & Krijthe, J. (2025). Falsification of unconfoundedness by testing independence of causal mechanisms. *Proceedings of the 42nd International Conference on Machine Learning, 267*, 29128–29147

et al., 2018; Wager & Athey, 2018), all methods have in common that before a causal effect can be estimated, often untestable conditions need to hold. One such condition is that we assume there is *no unmeasured confounding*, meaning that there are no unobserved factors that have both an influence on the treatment and on the outcome of interest that are not accounted for by the method. If unmeasured confounders are present, our treatment effect estimates are likely to be biased and inconsistent (Greenland *et al.*, 1999). This can have serious downstream consequences such as unknowingly recommending a non-effective or, even worse, potentially harmful treatment policy. Unfortunately, without making further assumptions, it is in general impossible to verify all assumptions needed to identify treatment effects from observational data.

In this work, we investigate a novel strategy for falsifying unconfoundedness. Specifically, we focus on the common scenario where observational datasets are collected from different heterogeneous sources, which we refer to as *environments*. Each environment corresponds to distinct study populations, due to factors such as geographical differences that results in distribution shifts between the environments. We propose a falsification strategy based on the assumption that these distribution shifts stem from independent changes in the underlying causal mechanisms. This idea is grounded in the principle of independent causal mechanisms (ICM) (Janzing *et al.*, 2012; Peters *et al.*, 2017), which posits that a causal system comprises autonomous modules that do not inform or influence each other. Assuming independent causal mechanisms has been leveraged to, for instance, improve causal structure learning (Guo *et al.*, 2023; B. Huang *et al.*, 2020) and understand model behavior in statistical machine learning (Schölkopf *et al.*, 2012). However, the implications of assuming independent causal mechanisms to treatment effect estimation problems has received far less attention.

Our proposed falsification strategy leverages a key observation that unmeasured confounding can cause observed mechanisms to appear dependent (Janzing & Schölkopf, 2018; Karlsson & Krijthe, 2023; Mameche, Vreeken & Kaltenpoth, 2024; Reddy & Balasubramanian, 2024). If we assume that the underlying causal mechanisms should be independent, contrary to what is observed, it follows that any apparent dependencies could be the result of unmeasured confounding. This observation motivates the central research question of this paper: *How can we efficiently test causal mechanism independencies to falsify the conditions required for treatment effect estimation in settings with multi-environment data?*

Contributions By formalizing the problem using a Neyman-Rubin causal model for multi-environment data, we show that falsification of unconfoundedness is possible by testing dependencies between causal mechanisms directly by combining the principle of independent causal mechanisms with functional assumptions on the mechanisms. In this model, we prove that the presence of unmeasured confounding has testable implications in the form of dependencies between the model's observed parameters. Using our theoretical results, we introduce new algorithmic ideas that can be used for falsification: in particular, we propose a two-stage algorithm that statistically tests statistical dependencies between learned model parameters of the treatment assignment and outcome

mechanism. We show that our algorithm performs favorably compared to alternative approaches on both simulated and semi-synthetic data.. To showcase the potential applications of our algorithm and clarify what constitutes an "environment", we provide two illustrative examples where we envision our algorithm being used.

Example 3.1. *In a meta-analysis of multiple observational studies with individual participant data (Riley et al., 2010), our algorithm can jointly test whether an unmeasured confounder is present between the treatment and outcome across all studies. Here, each observational study serves as a distinct environment.*

Example 3.2. *In a single observational analysis involving a multi-level structure in which individuals are nested in clusters and non-randomly assigned to a treatment/control on an individual level, such as students from different schools (Leite et al., 2015) or patients from different hospitals (Goldstein et al., 2002), our algorithm can test whether the conditions necessary to identify treatment effects are violated within each cluster due to unmeasured confounding. In this context, the environment refers to the sub-populations within each cluster of the same observational study.*

3.2. Related works

When discussing the validity of causal assumptions, sensitivity analysis might come to mind. In sensitivity analysis, one hypothesizes departures from the assumption of no unmeasured confounding and investigates how different biases would arise depending on the hypothesized confounder's relationship with treatment and outcome (Cornfield *et al.*, 1959; Tan, 2006; VanderWeele & Ding, 2017). This typically results in bounds on the treatment effect, which is an instance of partial identification (Manski, 2003). However, while sensitivity analysis probes 'what-if' scenarios regarding potential unmeasured confounding (a process that can always be undertaken), falsification aims to empirically test whether assumptions are violated, based on the observable implications of those assumptions (which is not always feasible). For instance, falsification may involve testing the validity of instrumental variables (Pearl, 1995) or evaluating the compatibility of learned causal structures with observed data (Faller *et al.*, 2024). In this way, sensitivity analysis and falsification are complementary: the former explores possible scenarios, while the latter seeks direct empirical evidence for these scenarios. Despite this, falsification has received comparatively less attention in the literature.

One line of research on falsification in observational causal inference assumes that certain transportability conditions hold, allowing causal effects to be transferred between different environments (Dahabreh, Robins & Hernán, 2020; Z. Hussain *et al.*, 2023; Z. M. Hussain *et al.*, 2022). The basic premise is that, under transportability conditions, comparing treatment effect estimates from multiple observational studies, or from a single observational study and a randomized one, should yield consistent results. If inconsistencies are found, this can be used to falsify the identifiability conditions, assuming the transportability assumptions hold. This idea has been further extended to time-to-event

outcomes with censoring (Demirel *et al.*, 2024), as well as for quantifying bias from unmeasured confounding (De Bartolomeis, Abad *et al.*, 2024; De Bartolomeis, Martinez *et al.*, 2024). In contrast, our approach assumes independence of causal mechanisms, which does not require transportable treatment effects or access to randomized data.

Testing for independence of causal mechanisms has been applied in previous work to falsify causal assumptions, such as detecting hidden confounding (Karlsson & Krijthe, 2023) or testing the validity of instrumental variables (Burauel, 2023; Karlsson *et al.*, 2023). Most similar to our work is that of Karlsson and Krijthe (2023), though their method relies on conditional independence testing which is a notoriously difficult statistical problem in itself (Shah & Peters, 2020). To avoid the challenges of conditional independence testing–for instance, losing statistical power as the adjustment set becomes larger–we address this problem by proposing an alternative method that does not rely on conditional independence testing.

Parallel to our ideas on falsification, other approaches have been proposed for detecting or addressing unmeasured confounding, under various assumptions on the setting and data-generating process. For example, when multiple causes are observed (D'Amour, 2019; Y. Wang & Blei, 2019) or when a negative control is available (Lipsitch *et al.*, 2010).

Finally, our work investigates the implications of the principle of independent causal mechanisms, which has a rich literature in causal discovery, particularly in settings with data from multiple environments (Guo *et al.*, 2023; B. Huang *et al.*, 2020; Mameche, Kaltenpoth & Vreeken, 2024; Perry *et al.*, 2022). A closely related line of research assumes the existence of invariant mechanisms across environments (Peters *et al.*, 2016). In contrast, our approach explicitly allows these mechanisms to vary–and, as we will show, such variation is sometimes necessary to enable falsification. Rather than aiming to learn the entire causal structure as typically done in causal discovery, our approach focuses on verifying specific aspects of a partially known structure that is relevant for treatment effect estimation. Recently, Guo *et al.* (2023) examined how independent causal mechanisms can lead to identification of certain treatment effects, though they did not address scenarios where causal assumptions are violated, such as in the presence of unmeasured confounders, which we study here.

3.3. Setup

3.3.1. Notation & data structure

For each individual $i = 1, \dots, n$, we observe baseline covariates X_i in $\mathscr{X} \subseteq \mathbb{R}^d$, a treatment A_i in $\mathscr{A} \subseteq \mathbb{R}$ and outcome Y_i in $\mathscr{Y} \subseteq \mathbb{R}$. We allow the treatment and outcome to be binary or continuous; but to simplify exposition, we will mainly show our results for the continuous case and then discuss how to modify our theory for binary treatments and outcomes when appropriate. We consider observations to be collected from K different environments labeled with $S_i \in \{1, \dots, K\}$ where $K \geq 2$. We denote n_s as the number of observations from environment $\{S = s\}$ and we define $n = \sum_{s=1}^{K} n_s$ as the total number of observations. Each

observation therefore consists of the tuple $O_i = (X_i, S_i, A_i, Y_i)$. Throughout the paper, we will use capitalized letters to denote random variables and small letters to denote their realized values.

We are considering a setting with a composite dataset of observations from separate environments. Each environment represent a different study population where the sampling probability of individual i belonging to environment $\{S = s\}$ can be unknown; this setting is referred to as a non-nested study design (Dahabreh, Robertson *et al.*, 2020) . Formally, we consider observations within an environment $\{S = s\}$ to be sampled independently and identically (i.i.d) according to some distribution $(X, A, Y) \sim P(X, A, Y \mid S = s)$. This distribution may vary across the different environments $s \in \{1, \ldots, K\}$. Importantly, observations are not assumed to be i.i.d. if we consider the marginal distribution $P(X, A, Y)$ over all environments. Furthermore, we assume that the environments are related to each other by having a shared, albeit unknown, causal structure, that is: the causal directed acyclic graph (DAG) between the variables (X, S, A, Y) is the same for all $S \in \{1, \ldots, K\}$.

3.3.2. Assumptions for identification of causal effects

To define causal effects of interest, we use potential (counterfactual) outcomes (Rubin, 1974). For an individual i, we posit the potential outcome Y_i^a for $a \in \mathscr{A}$ which denotes the outcome under an intervention to set treatment A_i to a. For the typical causal analysis in a non-nested study design, the goal is often to estimate the average treatment effect or conditional average treatment effect between two different treatments $a, a' \in \mathscr{A}$ in the underlying population from an environment $\{S = s\}$, that is $\tau_s = \mathbb{E}[Y^a - Y^{a'} \mid S = s]$, resp. $\tau_s(x) = \mathbb{E}[Y^a - Y^{a'} \mid X = x, S = s]$. It is well-known that under certain conditions τ_s and $\tau_s(x)$ are identified from the observations in environment $\{S = s\}$.

Assumption 3.1. *We assume the following conditions for each environment $s = 1, \ldots, K$.*
Consistency: if $A_i = a$, then $Y_i^a = Y_i$, for every individual i and every treatment $a \in \mathscr{A}$.
Positivity: for each treatment $a \in \mathscr{A}$, if $f(x, S = s) \neq 0$, then $\Pr(A = a | X = x, S = s) > 0$.
Unconfoundedness: for each $a \in \mathscr{A}$, $Y^a \perp\!\!\!\perp A | (X, S = s)$.

Consistency is satisfied when the treatment is clearly defined, ensuring that no hidden treatment variation exist and that there is no interference between individuals. Positivity requires that every possible covariate pattern in the environment $S = s$ has a nonzero probability of receiving each possible treatment option. Unconfoundedness, also referred to as conditional exchangeability, implies there is no unmeasured confounding. That is, the covariates X are sufficient to adjust for in order to identify the causal effect of A on Y. In observational studies, assuming unconfoundedness is often considered controversial, requiring strong domain expertise to justify its validity.

When the conditions in Assumption 3.1 are met, both the average treatment effect and the conditional average treatment effect can be identified from the observed data (Hernan & Robins, 2023). Let $\mu_{a,s}(X) = \mathbb{E}[Y \mid X, A = a, S = s]$, then a statistical estimand for the ATE

is $\tau_s = \mathbb{E}\left[\mu_{a,s}(X) - \mu_{a',s}(X) \mid S = s\right]$ and for the CATE is $\tau_s(X) = \mu_{a,s}(X) - \mu_{a',s}(X)$. Rather than focusing on how to estimate these estimands from data, we will concentrate on how to assess the validity of the conditions that allow us to identify them in the first place. Specifically, in the context of data from multiple environments, we will demonstrate that Assumption 3.1 can be falsified under certain conditions related to distributional shifts across the different environments.

3.4. A novel falsification strategy

3.4.1. Assumptions on environment changes

We consider a general class of models of the treatment and potential outcomes, namely: all linear functions of the feature representations $\psi(X) : \mathscr{X} \to \mathbb{R}^z$ and $\phi(X, A) : \mathscr{X} \times \mathscr{A} \to \mathbb{R}^{z'}$,

$$\begin{aligned} A &= \alpha_s^\top \psi(X) + \varepsilon_A \\ Y^a &= \beta_s^\top \phi(X, A = a) + \varepsilon_Y \end{aligned} \tag{3.1}$$

where the noise variables fulfill $\mathbb{E}[\varepsilon_A \mid X, S = s] = 0$ and $\mathbb{E}[\varepsilon_Y \mid X, A, S = s] = 0$. Additionally, under Assumption 3.1, the noise variables are independent $\varepsilon_A \perp\!\!\!\perp \varepsilon_Y \mid S$.

The function class described in eq. (3.1) encompasses a wide range of complex models, particularly because the representations $\psi(X)$ and $\phi(X, A)$ can involve nonlinear transformations of the variables (X, A). Although we focus on continuous treatment and outcome values to illustrate the core ideas of our falsification strategy, this framework can be extended to generalized linear models that accommodates binary/categorical values. For instance, we can include binary treatment by defining $P(A = 1 \mid X, S = s) = h^{-1}(\alpha_s^\top \psi(X))$, where $h(p) = \ln(p/(1-p))$ is the logit link function (Mccullagh & Nelder, 1989).

Distributional changes between environments are accounted for by eq. (3.1) through allowing the parameters $\alpha_s \in \mathbb{R}^z$ and $\beta_s \in \mathbb{R}^{z'}$ to change for different environments $s \in \{1, \ldots, K\}$, in addition to changes in the covariate distribution $P(X \mid S = s)$. Changes in the parameters (α_s, β_s) correspond to shifts in the treatment assignment mechanism $\mathbb{E}[A \mid X, S = s] = \alpha_s^\top \psi(X)$ and outcome mechanism $\mathbb{E}[Y^a \mid X, S = s] = \beta_s^\top \phi(X, A = a)$; the feature representations $\psi(X)$ and $\phi(X, A)$ are considered to be fixed across environments. In practice, changes in the treatment assignment and outcome mechanisms are often expected. For example, if an unmeasured effect modifier exists, the outcome mechanism $\mathbb{E}[Y^a \mid X, S = s]$ will vary if the distribution of unmeasured effect modifiers differs across environments (Dahabreh, Robertson *et al.*, 2020). Similarly, variation in the treatment assignment $\mathbb{E}[A \mid X, S = s]$ can be expected due to factors like differences in treatment policies across environments.

We will now pose the main assumption on how changes in the mechanism parameters (α_s, β_s) occur. Specifically, we assume there exists an unknown *prior distribution* $P(\alpha, \beta)$ that fulfills the following condition.

Assumption 3.2. *The parameters* $(\alpha_s, \beta_s) \sim P(\alpha, \beta)$ *are drawn independently for each* $s = 1, \ldots, K$. *Furthermore, the parameters are independent from each other such that* $P(\alpha, \beta) = P(\alpha)P(\beta)$.

Following the principle of independent causal mechanisms, Assumption 3.2 states that the parameters (α_s, β_s) are *uninformative* of each other as they are sampled independently, and furthermore, that changing α has *no influence* on β, and vice versa. In the language of structural causal models, sampling (α_s, β_s) should be seen as independent soft interventions on the distribution $P(X, A, Y \mid S = s)$. In the broader statistical context, Assumption 3.2 can also be related to hierarchical regression models with a prior independence assumption, see e.g. Gelman (2007, Chapter 11). Here, the independent sampling of the parameters (α_s, β_s) resembles the way hierarchical models account for variability between environments.

Finally, we will contrast our approach with falsification strategies based on transportability, which for instance would assume that the outcome mechanism $\mathbb{E}[Y^a \mid X, S = s]$ remains invariant across environments S. This assumption can be violated when unmeasured effect modifiers differ in distribution across environments, causing $\mathbb{E}[Y^a \mid X, S = s]$ to vary. In contrast, Assumption 3.2 does not require such invariance and explicitly allows this causal mechanism to vary. As a result, even when transportability fails to hold, Assumption 3.2 may still hold. We will later show that this makes our proposed falsification robust to violations of transportability, whereas transportability-based strategies may yield false positives: that is, incorrectly rejecting unconfoundedness despite the absence of unmeasured confounding. For a more detailed discussion of transportability-based falsification strategies, see Appendix 3.A.

3.4.2. A testable implication under the independence assumption

To focus on the core ideas and limits of our falsification strategy, we assume the feature representations ϕ and ψ are known up to some permutation and element-wise scaling. Moreover, to allow the use of standard estimation techniques, we will require the dimensionality of the feature representations to not be larger than any of the individual sample sizes among the different environments. We formalize these two conditions as follows.

Assumption 3.3. *We have access to* $\widetilde{\phi}(X) = C\phi(X)$ *and* $\widetilde{\psi}(X, A) = D\psi(X, A)$ *where* $C \in \mathbb{R}^{z \times z}$ *and* $D \in \mathbb{R}^{z' \times z'}$ *are invertible matrices. The dimensionality of the feature representations* $z, z' \in \mathbb{N}$ *is finite and lower than the smallest sample size across environments, i.e,* $z, z' < \min_s n_s$.

Our proposed falsification strategy will rely on estimating $\mathbb{E}[A \mid X, S = s]$ and $\mathbb{E}[Y \mid X, A, S = s]$ which under Assumption 3.1 corresponds to the true treatment and outcome mechanism. To estimate these conditional expectations, we employ two statistical working models $e_s(X) = \omega_s^\top \widetilde{\phi}(X)$ and $h_s(X, A) = \gamma_s^\top \widetilde{\psi}(X, A)$, respectively. Since we replaced the unknown feature representations $\{\phi, \psi\}$ with the observed feature representations $\{\widetilde{\phi}, \widetilde{\psi}\}$,

the mechanism parameters (α_s, β_s) are replaced by (ω_s, γ_s). Our falsification strategy will test a statement equivalent to Assumption 3.2 but, again, substituting (α_s, β_s) with (ω_s, γ_s) as follows,

$$H_0 : P(\omega, \gamma) = P(\omega)P(\gamma) \,. \tag{3.2}$$

To understand how our falsification strategy will be centered around testing this null hypothesis, we begin by establishing the following key result (see Appendix 3.B.1 for the proof).

Theorem 3.1. *Under the functional class described in eq.* (3.1)*, assumptions 3.1, 3.2 and 3.3, and with $e_s(X)$ and $h_s(X, A)$ being correctly specified models for $\mathbb{E}[A \mid X, S = s]$ and $\mathbb{E}[Y \mid X, A, S = s]$, we have that H_0 in eq.* (3.2) *is true.*

The above theorem suggests that if we reject the null hypothesis H_0, it is likely because at least one of the conditions in the theorem is violated. While rejecting H_0 does not tell which condition in the theorem could be false, it still provides valuable information about the validity of the conditions in Assumption 3.1 which are necessary for treatment effect estimation. Before introducing our algorithm to statistically test H_0, we first explore a setting where violating Assumption 3.1 provably leads to the falsity of H_0.

3.4.3. Unmeasured confounding leads to mechanism dependencies

We examine a setting involving a linear causal model that includes both a main effect of treatment and interaction effects between treatment and covariates. While linearity may not always hold in real-world scenarios, this setting offers valuable insights into the conditions necessary to falsify causal assumptions in a multi-environment context.

To understand what effect an unmeasured confounder has on the independence of mechanisms, we introduce another unmeasured covariate $U \in \mathbb{R}$ as follows,

$$\begin{aligned} A &= \alpha_s^\top \psi(X) + \alpha_s^{(U)} U + \varepsilon_A \\ Y^a &= \beta_s^\top \phi(X, A = a) + \left(\beta_s^{(U)} + a\beta_s^{(AU)}\right) U + \varepsilon_Y \end{aligned} \tag{3.3}$$

We let $\psi(X) = [1, X]^\top$ and $\phi(X, A) = [1, X, A, AX]^\top$ such that we can define the parameters $\alpha_s = [\alpha_s^{(0)}, \alpha_s^{(X)}]$ and $\beta_s = [\beta_s^{(0)}, \beta_s^{(X)}, \beta_s^{(A)}, \beta_s^{(AX)}]$. Throughout this example, we assume that $X \perp\!\!\!\perp U \mid S$.

The above causal model is partially observed because U is an unmeasured covariate. If U is a common cause to both the treatment A and potential outcome Y^a, that is $\{\alpha_s^{(U)} \neq 0, \beta_s^{(U)} \neq 0\}$ and/or $\{\alpha_s^{(U)} \neq 0, \beta_s^{(AU)} \neq 0\}$, then we say that U is an unmeasured confounder.

Whereas it is in general impossible to determine the presence of U, if we have correctly specified working models for $\mathbb{E}[A \mid X, S = s]$ and $\mathbb{E}[Y \mid A, X, S = s]$, we note that there exists dependencies between the observable parameters ω_s and γ_s when U is an unmeasured confounder (see Appendix 3.B.2 for the proof).

Lemma 3.1. *Assume U has a normal distribution with mean $\mu_s^U \in \mathbb{R}$ and standard deviation $\sigma_s^{(U)} \in \mathbb{R}^+$, and the noise variables $(\varepsilon_A, \varepsilon_Y)$ are normally distributed with mean zero and standard deviations $\sigma^{(A)} \in \mathbb{R}^+$ and $\sigma^{(Y)} \in \mathbb{R}^+$. Consider the well-specified working models $e_s(X) = \omega_s^\top \widetilde{\phi}(X)$ and $h_s(X,A) = \gamma_s^\top \widetilde{\psi}(X,A)$ with $\widetilde{\phi}(X) = [1, X]^\top$ and $\widetilde{\psi}(X,A) = [1, X, A, AX, A^2]^\top$. Under the model in eq.* (3.3) *with U being an unmeasured confounder, we then have that the observable parameters are $\omega_s = \alpha_s + [\alpha_s^{(U)}\mu_s^{(U)}, 0]^\top$ and $\gamma_s = [\beta_s, 0]^\top + \Gamma_s$ where*

$$\Gamma_s = \delta_s \begin{bmatrix} -\beta_s^{(U)}\left(\frac{\alpha_s^{(0)}(\sigma_s^{(U)})^2}{\alpha_s^{(U)}} - \mu_s^{(U)}\left(\frac{\sigma_s^{(A)}}{\alpha_s^{(U)}}\right)^2\right) \\ -\beta_s^{(U)}\frac{\alpha_s^{(X)}(\sigma_s^{(U)})^2}{\alpha_s^{(U)}} \\ \beta_s^{(U)}\frac{(\sigma_s^{(U)})^2}{\alpha_s^{(U)}} - \beta_s^{(AU)}\left(\frac{\alpha_s^{(0)}(\sigma_s^{(U)})^2}{\alpha_s^{(U)}} - \mu_s^{(U)}\left(\frac{\sigma_s^{(A)}}{\alpha_s^{(U)}}\right)^2\right) \\ -\beta_s^{(AU)}\frac{\alpha_s^{(X)}(\sigma_s^{(U)})^2}{\alpha^{(U)}} \\ \beta_s^{(AU)}\frac{(\sigma_s^{(U)})^2}{\alpha_s^{(U)}} \end{bmatrix}$$

and $\delta_s = \left((\sigma_s^{(U)})^2 + \left(\frac{\sigma_s^{(A)}}{\alpha_s^{(U)}}\right)^2\right)^{-1}$.

The lemma, which holds for any $P(X \mid S = s)$, states that if U is an unmeasured confounder then the observable parameters (γ_s, ω_s) have shared dependencies on the true parameters of the underlying data-generating process: the parameters $(\alpha_s^{(0)}, \alpha_s^{(X)}, \alpha_s^{(U)}, \mu_s^{(U)})$ appear in both the expressions of ω_s and γ_s.

The above results hold for both single-environment data ($K = 1$) and multi-environment data ($K > 1$), and does not rely on Assumption 3.2. Next, we show that our proposed falsification strategy allows us to detect the presence of the unmeasured confounder U under certain conditions on the multi-environment structure when invoking Assumption 3.2 (see Appendix 3.B.3 for the proof).

Theorem 3.2. *Under the assumptions stated in Lemma 3.1 and Assumption 3.2, if at least one of the following parameters $(\alpha_s^{(0)}\alpha_s^{(X)}, \alpha_s^{(U)}, \mu_s^{(U)})$ are i.i.d. sampled from a non-degenerate distribution for $s = 1, \ldots, K$, then H_0 is false if and only if U is a confounder for all $s \in \{1, \ldots, K\}$.*

The theorem establishes that H_0 can be false due to violations of Assumption 3.1, which can be understood in terms of following statement: unmeasured confounding can create dependencies between observable parameters. The reason at least one of the parameters $(\alpha_s^{(0)}\alpha_s^{(X)}, \alpha_s^{(U)}, \mu_s^{(U)})$ must be sampled from a non-degenerate distribution is that this creates a statistical dependence between ω_s and γ_s in the presence of an unmeasured confounder. Detecting this dependence becomes crucial for falsifying unconfoundedness.

The parameters $(\alpha_s^{(0)}, \alpha_s^{(X)}\alpha_s^{(U)}, \mu_s^{(U)})$ are related to the distributions $P(A \mid X, S = s)$ and $P(U \mid S = s)$. Thus, the non-degeneracy condition implies that falsifying Assumption 3.1

requires changes in either the treatment assignment or the distribution of the unmeasured confounder across environments. This observation motivates the requirement of having multi-environment data: that is, without multiple environments there are no distributional changes that enable falsification to happen.

The same non-degeneracy condition was observed by Karlsson and Krijthe (2023) despite using a different formalism based on causal graphs. Their approach relies on identifying a specific d-separation via a conditional independence test between the treatment and outcome variables, which also allows for falsification of unconfoundedness across multiple environments. Their test can be interpreted as an indirect test of independence of causal mechanisms, as it operates solely on observed variable relationships. In contrast, our approach directly tests independence at the level of mechanism parameters. As a consequence, a further key difference is that their approach needs to make additional structural assumptions on the covariate distribution $P(X \mid S)$, while our theoretical results impose no such constraint.

3.5. Algorithm

We now introduce the Mechanism INdependent Test (MINT) algorithm, which operationalizes our falsification strategy for testing mechanism independence using data from multiple environments. We will use the following notation: for all environments $s = 1,\dots,K$, we denote the observed data matrices as $\mathbf{A}_s = [A_1,\dots,A_{n_s}]^\top$, $\mathbf{Y}_s = [Y_1,\dots,Y_{n_s}]^\top$, $\widetilde{\Psi}_s = [\widetilde{\psi}(X_1),\dots,\widetilde{\psi}(X_{n_s})]^\top$, and $\widetilde{\Phi}_s = [\widetilde{\phi}(X_1,A_1),\dots,\widetilde{\phi}(X_{n_s},A_{n_s})]^\top$.

The MINT algorithm can be divided into two steps: In the first stage, for all $s = 1,\dots,K$, we estimate the parameters (ω_s,γ_s). The estimates are obtained through solving the least-squares problems $\widehat{\omega}_s = \arg\min_\omega ||\mathbf{A}_s - \widetilde{\Psi}_s\omega||_2^2$ and $\widehat{\gamma}_s = \arg\min_\gamma ||\mathbf{Y}_s - \widetilde{\Phi}_s\gamma||_2^2$ where $||\cdot||_2^2$ denotes the l^2-norm. We denote all estimated parameters as $\widehat{\boldsymbol{\omega}} = [\widehat{\omega}_1,\dots,\widehat{\omega}_K]$ and $\widehat{\boldsymbol{\gamma}} = [\widehat{\gamma}_1,\dots,\widehat{\gamma}_K]$. In the second stage, we perform a statistical independence test for the null hypothesis $H_0 : P(\omega,\gamma) = P(\omega)P(\gamma)$ using the estimated parameters $\widehat{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\gamma}}$. If we accept H_0 then we should consider Assumption 3.1 and 3.2 to hold. On the other hand, if we reject H_0 then both assumptions are falsified jointly.

For the statistical independence test in the second stage, we study the co-variability of (γ_s,ω_s) across all environments by analyzing the covariance matrix $\Sigma = \mathrm{Cov}(\omega,\gamma)$. We propose using the test statistic $T = \sqrt{\sum_{i,j}|\Sigma_{ij}|^2}$ which is the Frobenius norm of the covariance matrix; crucially, this test statistic is always non-negative and $T = 0$ under H_0. The estimated test statistic becomes

$$\widehat{T}(\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\gamma}}) = \frac{1}{K}\sqrt{\sum_{i=1}^{z}\sum_{j=1}^{z'}\left[\sum_{s=1}^{K}(\widehat{\omega}_{s,i}-\bar{\omega}_i)(\widehat{\gamma}_{s,j}-\bar{\gamma}_j)\right]^2}$$

where $\bar{\omega}_i = K^{-1}\sum_{s=1}^{K}\widehat{\omega}_{s,i}$ and $\bar{\gamma}_j = K^{-1}\sum_{s=1}^{K}\widehat{\gamma}_{s,j}$.

Lastly, we need to calibrate a rejection threshold R such that we reject H_0 if $\widehat{T}(\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{\gamma}}) > R$

while ensuring guarantees on the Type I error $\Pr(\widehat{T}(\widehat{\omega},\widehat{\gamma}) > R \mid H_0) \leq \alpha$ for some $\alpha \in (0,1)$. While this can be done using a permutation-based procedure with M resamples, we have to take into account the uncertainty of the estimates from the model fitting in the first step of our algorithm.

To address this problem, we introduce an additional modification in the permutation-based calibration. Specifically, in the first step, we use bootstrapping and resample M datasets with replacement to obtain estimates $\{(\widehat{\omega}^{(m)},\widehat{\gamma}^{(m)})\}_{m=1}^{M}$. Then, for each $m = 1,\ldots,M$, we compute $T_m = T(\widetilde{\omega}^{(m)},\widehat{\gamma}^{(m)})$ where $\widetilde{\omega}^{(m)}$ is a random permutation of $\widehat{\omega}^{(m)}$. Finally, we determine the rejection threshold as

$$R = \underset{t\in(0,\infty)}{\arg\max}\{t : M^{-1}\sum_{m=1}^{M} 1(T_m > t) \leq \alpha\},$$

where $1(T_m > t)$ equals 1 if $T_m > t$ and otherwise 0. Throughout the remainder of the paper, we let $M = 1000$. To highlight the importance of bootstrapping in the calibration, we present an ablation study in Appendix 3.D.5 where we show that bootstrapping is essential for ensuring Type 1 errors remain below α.

3.6. Experiments

We conducted a series of experiments to compare the proposed MINT algorithm with alternative baseline approaches. First, we investigate efficiency with respect to number of samples and number of environments. Next, we validate our theoretical findings by investigating necessary mechanism changes that allow for falsification. We then assessed the sensitivity of our algorithm to (mis)specification in its working models. Lastly, we evaluated all methods under more realistic conditions using semi-synthetic data based on the real-world Twins dataset (Almond *et al.*, 2005), which includes birth data across different geographical locations used as environment labels.

We measured performance using the falsification rate (probability of falsification) and set the significance level $\alpha = 0.05$ to control Type 1 errors. In the absence of unmeasured confounding, the falsification rate reflects the Type I error rate and should remain below the significance level $\alpha = 0.05$. Conversely, in the presence of unmeasured confounding, the falsification rate corresponds to the statistical power of the algorithms, and thus, a higher rate is desirable. The code for reproducing our experiments is available at our GitHub repository.[1]

3.6.1. Baselines

We compare the proposed MINT algorithm to two baselines. The first, referred to as the transportability test, evaluates whether the independence $Y \perp\!\!\!\perp S \mid A, X$ holds, allowing for

[1] https://github.com/RickardKarl/falsification-unconfoundedness

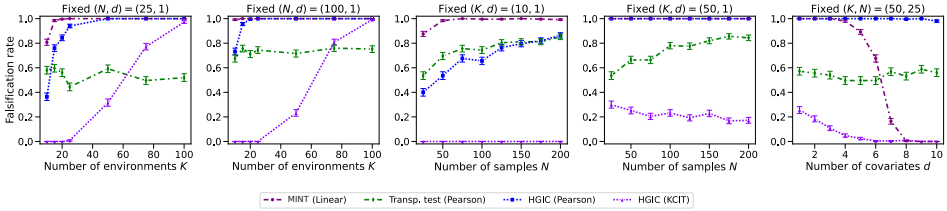


Figure 3.1: Comparison of falsification rate when varying either the number of environment K, the number of samples per environment N, or the number of observed covariates d. The error bars show the standard error over 250 repetitions.

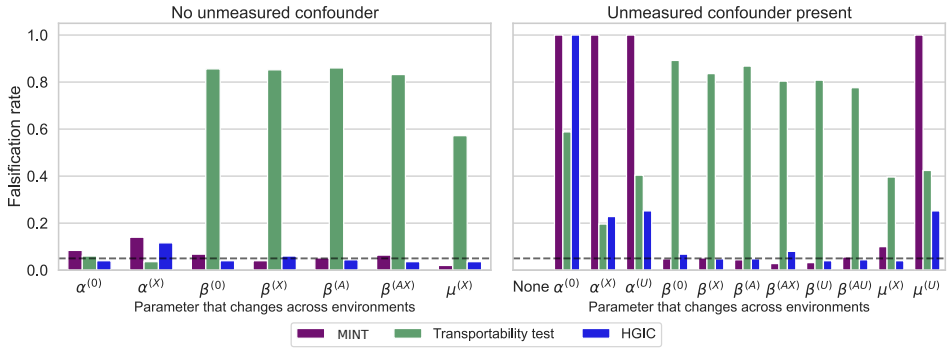


Figure 3.2: Comparison of falsification rate when different mechanisms vary across the environment: The parameters on the x-axis correspond to those of the data-generating process in eq. (3.3). The black dotted line corresponds to the chosen significance level $\alpha = 0.05$. Average falsification rate and standard errors are reported over 250 repetitions. In the absence of unmeasured confounding, the falsification rate reflects the Type I error rate and should remain below the significance level $\alpha = 0.05$. Conversely, in the presence of unmeasured confounding, the falsification rate corresponds to the power of the test, and thus, a higher rate is desirable.

the joint falsification of Assumption 3.1 and the transportability condition $Y^a \perp\!\!\!\perp S \mid X$ (Dahabreh, Robins & Hernán, 2020). A detailed overview of transportability-based falsification strategies is provided in Appendix 3.A. The second baseline is the hierarchical graph independence constraint (HGIC) approach (Karlsson & Krijthe, 2023). This approach tests a conditional independence statement based on a hierarchical description of the data. Unlike the transportability test, HGIC remains valid even when transportability conditions are violated, as it relies on an independence of causal mechanisms assumption similar to ours. This makes HGIC a strong candidate for comparison.

Since both baselines require selecting a conditional independence testing method, we

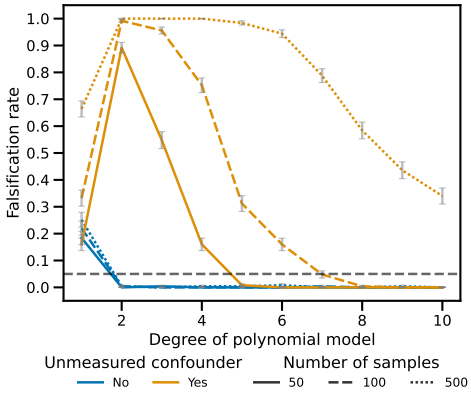


Figure 3.3: Our algorithm's performance is evaluated using polynomial basis functions as feature representation. The falsification rate is plotted against polynomial degree, with the true data-generating process including polynomials up to degree 2. The black dotted line corresponds to the chosen significance level $\alpha = 0.05$. Average falsification rate and standard errors are reported over 250 repetitions.

evaluated them using either the Pearson partial correlation test, which is suitable for linear data, or the non-parametric kernel conditional independence test (KCIT) (K. Zhang *et al.*, 2011) with a radial basis function kernel, which is better suited for nonlinear data. For HGIC, we encountered some issues with KCIT that required minor modifications to the original implementation used by Karlsson and Krijthe (2023). These issues and the differences between our implementation and the original are discussed in detail in Appendix 3.D.4.

3.6.2. Synthetic data

Which method is most efficient?

In the first experiment, we aimed to evaluate the efficiency of each method in a well-specified linear setting (see Appendix 3.D.1 for more details on data generation). To make our method well-specified to the underlying data generating process, we used linear feature representations for MINT, and for the two baselines methods we used the partial Pearson correlation test which is suitable for conditional independence testing with linear data. Additionally, following Karlsson and Krijthe (2023), we tested HGIC using KCIT, as it is also well-specified in this context.

We evaluated the falsification rate of each method under an unmeasured confounder while varying the number of environments K, the number of samples per environment N, or the number of observed covariates d, keeping the other factors fixed. The results are

shown in Figure 3.1. When varying the number of environments K, our proposal MINT consistently outperformed HGIC. The transportability test was most effective when K was small, though both MINT and HGIC showed higher falsification rates as K increased. HGIC performed better with the Pearson test than with KCIT, highlighting the advantage of a parametric test in a well-specified setting. Increasing the number of samples N improved falsification rates for all methods except HGIC with KCIT, although the gains were slower compared to increasing K. Finally, when varying d, HGIC with KCIT lost power the fastest, followed by MINT, while Pearson-based methods remained robust up to $d = 10$.

Additionally, in Appendix 3.D.5, we confirm that all methods controlled Type 1 error in the absence of an unmeasured confounder, with the falsification rate remaining below the significance level $\alpha = 0.05$.

What are necessary mechanism changes to detect confounding?

In the second experiment, we validated the theory behind our proposed MINT algorithm by generating various types of independent mechanism changes across the environments, following the model in eq. (3.3) (see Appendix 3.D.2 for details on data generation). We also applied the baseline methods to the same data to provide further insights into the necessary conditions for them to serve as a valid falsification strategy.

The different parameters on the x-axis in Figure 3.2 represent which of the parameters in eq. (3.3) are varied across different environments, while all other parameters are kept fixed. This is done under both the absence and presence of unmeasured confounding. We observed that environmental changes in the parameters $(\alpha_s^{(0)}, \alpha_s^{(X)}, \alpha_s^{(U)}, \mu_s^{(U)})$, which influence either the treatment mechanism $\mathbb{E}[T \mid X, S = s]$ or the distribution of the unmeasured confounder $P(U \mid S = s)$, were sufficient for MINT to falsify unconfoundedness. This observation supports the claim we proved in Theorem 3.2.

Furthermore, both HGIC and the transportability test falsified under the same conditions when unmeasured confounders were present. However, the transportability test showed a notable issue with false positives in the absence of unmeasured confounding. The most likely explanation is that these false positives result from mechanism changes that violate the transportability condition, a key assumption for applying the transportability test.

Model Specification & Performance

In the third experiment, we sampled data from a process with a polynomial basis function (see Appendix 3.D.1 for more details on the data-generating process). We examined how changing the specification of the working models $e_s(X)$ and $h_s(X)$ in MINT affected its performance. The true polynomials in the data-generating process had a degree of 2, while MINT used a representation with polynomial basis functions with degrees ranging from 1 to 10. If the degree was set to 1, this introduced misspecification. For degrees of 2 or

higher, the model was well-specified but became increasingly flexible as the polynomial degree increased.

As shown in Figure 3.3, misspecified models led to an elevated false positive rate in the absence of unmeasured confounding. However, once the models were well-specified, false positives dropped below the nominal level $\alpha = 0.05$ even as model flexibility increased. When an unmeasured confounder was present, MINT successfully detected it, though its power (i.e., true positive rate) declined with higher model flexibility. We observed, however, that this reduction in power could be mitigated by increasing the number of samples per environment.

We also compared the transportability test and HGIC under both well-specified and misspecified settings. Using the Pearson partial correlation test on nonlinear data allowed us to assess their performance under misspecification. Similar to MINT, both exhibited higher false positive rates in the absence of unmeasured confounding when misspecified. Full results are provided in Table 3.2 in Appendix 3.D.5.

3.6.3. Twins data

In the final experiment, we used data from twin births in the USA between 1989-1991 (Almond *et al.*, 2005) to construct a multi-environment observational dataset with a known causal structure. The environment corresponds to the birth state of each pair of twins. We generated treatment and outcome variables using the covariates from this dataset, providing a ground-truth causal structure to validate our methods while emulating realistic distributions with real-world covariates (see Appendix 3.D.3 for details on dataset construction). The outcomes and treatment were modeled using a quadratic polynomial, and all methods were well-specified through either a quadratic polynomial feature representation or KCIT for conditional independence testing.

We examined a scenario with five confounders, varying the number of observed covariates from one to five. When all five confounders were observed, no unmeasured confounders remained; otherwise, some confounders were unmeasured. As a control, we repeated the experiment while varying the number of observed confounders but ensuring no unmeasured confounders. The results, shown in Figure 3.4, indicate that MINT outperforms both the transportability test and HGIC in terms of power. Additionally, when all five confounders were observed, MINT achieved the nominal falsification error below $\alpha = 0.05$.

3.7. Discussion

Our falsification strategy is not a silver bullet to detect unmeasured confounding. As we have demonstrated, our proposed algorithm is a joint falsification test that assesses both the conditions necessary for causal identification and the assumption of independent causal mechanisms. Thus, the limit to how informative this falsification test can be will

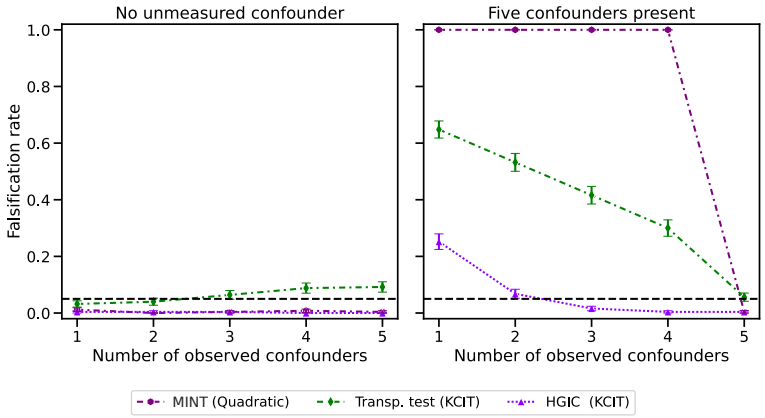


Figure 3.4: Comparison on the Twins semi-synthetic dataset. Average falsification rate and standard errors are reported over 250 repetitions; the black dotted line correspond to the chosen significance level $\alpha = 0.05$.

depend on the plausibility of the independent causal mechanism assumption.

One reason our proposed algorithm performs well, especially compared to HGIC with KCIT, could be because of the parametric nature of our approach. While parametric assumptions can be incorporated into both HGIC and the transportability test by selecting an appropriate conditional independence test, such as the Pearson test used in our experiment, our algorithm encodes these assumptions differently. Specifically, it explicitly incorporates the parametric assumptions for both the treatment and outcome models. Interestingly, this approach aligns more closely with the common practice of specifying both models when estimating treatment effects in observational studies.

A drawback of relying on parametric assumptions for the treatment and outcome models is the increased Type 1 error under misspecification. This happened to our algorithm when $\{\tilde{\psi}, \tilde{\phi}\}$ were misspecified. So far, we have assumed these representations are known a priori. To mitigate the risk of misspecification, one strategy is to construct feature representations that apply a broad set of transformations to the observed covariates. This ensures the representation is sufficiently expressive to capture the underlying relationships in the data. However, increasing the richness of the feature representation introduces a trade-off: while it reduces the risk of misspecification, it can also decrease statistical power due to increased model complexity. This trade-off was evident in our experiments (Figure 3.3), where increasing model complexity lead to a reduction in power of the test.

Hence, a key future direction is to address the case where $\{\tilde{\psi}, \tilde{\phi}\}$ are unknown and attempt to learn them from data. In this case, we have shown that it would be sufficient to learn them up to some permutation and element-wise scaling. Alternatively, we could attempt to use more flexible (implicit) feature representations through the use of kernel methods (Schölkopf & Smola, 2002). Whereas further work is needed to adapt our theory to

a kernelized algorithm, we provide a sketch as a starting point for such an approach in Appendix 3.C.

3.8. Conclusion

We propose novel algorithmic ideas to directly exploit observed dependencies in causal mechanisms for falsification of the assumptions necessary for causal effect identification. Specifically, we propose a two-stage algorithm that can be applied to multi-environment data. Although there are no universal solutions for addressing untestable assumptions in causal inference, we believe that our proposal has an important place in evaluating the necessary conditions to enable more trustworthy causal conclusions.

Appendices

3.A. Falsification with transportability conditions

A common way of using data from multiple environments to falsify the validity of Assumption 3.1 is to assume a transportability condition that relates the different environments to each other. One of the most common way of formalizing the transportability condition is as follows.

Assumption 3.4 (Conditional exchangeability between environments)**.** *We assume for all* $a \in \mathscr{A}$, $Y^a \perp\!\!\!\perp S \mid X$.

Other variations of the transportability condition is to assume conditional mean exchangeability $\mathbb{E}[Y^a \mid X, S = s] = \mathbb{E}[Y^a \mid X]$ or that an effect measure such as the conditional average treatment effect $\mathbb{E}[Y^1 - Y^0 \mid X, S = s] = \mathbb{E}[Y^1 - Y^0 \mid X]$ is transportable (Colnet *et al.*, 2024).

It is well-known that Assumption 3.1 and 3.4 together have a testable implication in the law of the observed data, see e.g. Dahabreh, Robins and Hernán (2020). More specifically, Assumption 3.1 and 3.4 together imply that the following conditional independence must be true

$$Y \perp\!\!\!\perp S \mid X, A\,. \tag{3.4}$$

Testing eq. (3.4) can be done with any suitable conditional independence test and efficient procedures also exists for testing implications from the other variations of the transportability condition, see e.g. Z. Hussain *et al.* (2023). However, the underlying premise is always the same: if one would conclude that eq. (3.4) does not hold, then this could be due to either a violation of Assumption 3.1 or 3.4. Thus, if one believes that Assumption 3.4 must hold yet observes eq. (3.4) to be false, that means that Assumption 3.1 must be violated in at least one of the environments. This argument becomes particularly strong if treatment has been randomized in one of the environments since any difference between the environments is more likely to be explained by an unmeasured confounder in the remaining environments with observational data. However, Assumption 3.4 itself can be controversial as it is also untestable. More specifically, it would be violated if there are unmeasured so-called effect modifiers which are covariates that differ in distribution between environments and modulate treatment heterogeneity. Effect modifiers are distinct from confounders as effect modifiers only need to be a cause of the outcome of interest. Thus, confounders can be effect modifiers but not vice versa, meaning that we often might expect to have unmeasured effect modifiers present even where are no unmeasured confounders.

The primary distinction between our falsification strategy and a transportability-based falsification strategy lies in the assumption our strategy relies on: instead of using Assumption 3.4, our strategy employs Assumption 3.2 to derive an alternative jointly testable implication. This comparison also highlights their underlying similarity: both strategies

aim to combine two untestable assumptions to generate a testable implication, enabling the joint falsification of these otherwise untestable assumptions.

3.B. Proofs

3.B.1. Proof of Theorem 3.1

Proof. Using the conditions from Assumption 3.1 and that $h_s(X) = \gamma_s^\top \widetilde{\phi}(X, A = a)$ is a correctly specified model for $\mathbb{E}[Y \mid A, X, S = s]$, we can write

$$\begin{aligned}
\beta_s^\top \phi(X, A = a) &= \mathbb{E}[Y^a \mid X, S = s] \\
&= \mathbb{E}[Y^a \mid X, A = a, S = s] && Y^a \perp\!\!\!\perp A \mid (X, S = s) \\
&= \mathbb{E}[Y \mid X, A = a, S = s] && A = a \Rightarrow Y^a = Y \\
&= \gamma_s^\top \widetilde{\phi}(X, A = a) \,.
\end{aligned}$$

Because we assumed $\widetilde{\phi}(X, A = a) = C\phi(X, A = a)$ for some invertible matrix C (Assumption 3.3), it follows for $s = 1, \ldots, K$ that $\gamma_s = (C^{-1})^\top \beta_s$. Furthermore, using that $e_x(X) = \omega_s^\top \widetilde{\psi}(X, A = a)$ is a correctly specified model for $\mathbb{E}[A \mid, X, S = s]$ and $\widetilde{\psi}(X, A) = D\psi(X, A)$ for some invertible matrix D (again, Assumption 3.3), it follows using similar arguments that $\omega_s = (D^{-1})^\top \alpha_s$ for $s = 1, \ldots, K$. To conclude the proof, using Assumption 3.2 which states that there exists a distribution $P(\alpha, \beta) = P(\alpha)P(\beta)$, we observe that (ω_s, γ_s) are distributed according to a distribution defined as $P(\omega, \gamma) := P((C^{-1})^\top \alpha, (D^{-1})^\top \beta)$. It is well-known that if α_s and β_s are independent random variables, then their transformations $(C^{-1})^\top \alpha_s$ and $(D^{-1})^\top \beta_s$ are also independent; see Grimmett and Stirzaker (2020, Chapter 4.2). Thus, we have $P(\alpha, \beta) = P(\alpha)P(\beta) \iff P(\omega, \gamma) = P(\omega)P(\gamma)$. □

3.B.2. Proof of Lemma 3.1

Before we can prove the lemma, we need the following auxiliary result.

Lemma 3.2. *Consider two Normal probability densities* $f_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2}\left(x - \mu_1\right)^2\right)$ *and* $f_2(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2\sigma_2^2}\left(x - \mu_2\right)^2\right)$ *with* $\sigma_1, \sigma_2 > 0$. *We then have that the product of the densities is* $f_1(x) \cdot f_2(x)$ *is proportional to a Normal density* $\frac{1}{\sqrt{2\pi\sigma_{12}^2}} \exp\left(-\frac{1}{2\sigma_{12}^2}\left(x - \mu_{12}\right)^2\right)$ *where*

$$\mu_{12} = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \text{ and } \sigma_{12}^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \,.$$

Proof. Note that

$$f_1(x) \cdot f_2(x) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \underbrace{\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(x-\mu_2)^2}{\sigma_2^2} \right]}_{Q} \right),$$

where the expression inside the exponential function can be written as

$$\begin{aligned} Q &= \frac{\left(\sigma_1^2+\sigma_2^2\right)x^2 - 2\left(\mu_1\sigma_2^2+\mu_2\sigma_1^2\right)x + \mu_1^2\sigma_2^2+\mu_2^2\sigma_1^2}{\sigma_1^2\sigma_2^2} \\ &= \frac{x^2 - 2\frac{\left(\mu_1\sigma_2^2+\mu_2\sigma_1^2\right)}{\left(\sigma_1^2+\sigma_2^2\right)}x + \frac{\mu_1^2\sigma_2^2+\mu_2^2\sigma_1^2}{\left(\sigma_1^2+\sigma_2^2\right)}}{\frac{\sigma_1^2\sigma_2^2}{\left(\sigma_1^2+\sigma_2^2\right)}} \\ &= \frac{\left(x - \frac{\left(\mu_1\sigma_2^2+\mu_2\sigma_1^2\right)}{\left(\sigma_1^2+\sigma_2^2\right)}\right)^2}{\frac{\sigma_1^2\sigma_2^2}{\left(\sigma_1^2+\sigma_2^2\right)}} + \frac{\frac{\left(\mu_1\sigma_2^2+\mu_2\sigma_1^2\right)}{\left(\sigma_1^2+\sigma_2^2\right)} + \frac{\mu_1^2\sigma_2^2+\mu_2^2\sigma_1^2}{\left(\sigma_1^2+\sigma_2^2\right)}}{\frac{\sigma_1^2\sigma_2^2}{\left(\sigma_1^2+\sigma_2^2\right)}}. \end{aligned}$$

As the second term on the last line is independent of x, we finish the proof by observing that $f_1(x) \cdot f_2(x)$ is up to some constant proportional to

$$\frac{1}{\sqrt{2\pi\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}} \exp\left(-\frac{1}{2} \frac{\left(x - \frac{\left(\mu_1\sigma_2^2+\mu_2\sigma_1^2\right)}{\left(\sigma_1^2+\sigma_2^2\right)}\right)^2}{\frac{\sigma_1^2\sigma_2^2}{\left(\sigma_1^2+\sigma_2^2\right)}} \right).$$

□

Next, we proceed with the proof of Lemma 3.1.

Proof. To simplify notation, we will drop the subscript s for all parameters. We start with $\mathbb{E}[A \mid X, S=s]$, which can be written as

$$\begin{aligned} \mathbb{E}[A \mid X, S=s] &= \mathbb{E}[\alpha^{(0)} + \alpha^{(X)}X + \alpha^{(U)}U + \varepsilon_A \mid X, S=s] \\ &= \alpha^{(0)} + \alpha^{(X)}X + \alpha^{(U)}\mathbb{E}[U \mid X, S=s] + \underbrace{\mathbb{E}[\varepsilon_A \mid X, S=s]}_{=0} \\ &\overset{(a)}{=} \alpha^{(0)} + \alpha^{(X)}X + \alpha^{(U)}\mu^{(U)} \\ &= \left(\begin{bmatrix} \alpha_0 \\ \alpha_X \end{bmatrix} + \begin{bmatrix} \alpha^{(U)}\mu^{(U)} \\ 0 \end{bmatrix} \right)^\top \begin{bmatrix} 1 \\ X \end{bmatrix} \end{aligned}$$

where in (a) we use that $X \perp\!\!\!\perp U \mid (S=s)$ meaning that $\mathbb{E}[U \mid X, S=s]\mathbb{E}[U \mid S=s] = \mu^{(U)}$.

Next, we continue with $\mathbb{E}[Y \mid X, A, S = s]$, which can be expressed as

$$\begin{aligned}
&\mathbb{E}\left[\beta^{(0)} + \beta^{(X)}X + \beta^{(U)}U + A\left(\beta^{(A)} + \beta^{(AX)}X + \beta^{(AU)}U\right) + \varepsilon_Y \mid X, A, S = s\right] \\
&\quad = \beta^{(0)} + \beta^{(X)}X + A\left(\beta^{(A)} + \beta^{(AX)}X\right) \\
&\qquad + \left(\beta^{(U)} + A\beta^{(AU)}\right)\mathbb{E}[U \mid X, A, S = s] + \underbrace{\mathbb{E}[\varepsilon_Y \mid X, A, S = s]}_{=0}.
\end{aligned} \tag{3.5}$$

To evaluate the conditional expectation $\mathbb{E}[U \mid X, A, S = s]$, we use Bayes rule to rewrite the probability density function

$$f(U \mid X, A, S) = \frac{f(A \mid X, U, S) f(U \mid X, S)}{f(A \mid X, S)}.$$

Firstly, note that $f(U \mid X, S) = f(U \mid S)$ follows from that $X \perp\!\!\!\perp U \mid (S = s)$. The density $f(U \mid S)$ is corresponds to the density of $N(\mu^{(U)}, (\sigma^{(U)})^2)$. Secondly, we can express $f(A \mid X, U, S)$ differently by exploiting that $A = \alpha^{(0)} + \alpha^{(X)}X + \alpha^{(U)}U + \varepsilon_A$ as follows,

$$\begin{aligned}
f(A = a \mid X = x, U = u, S = s) &= f(\varepsilon_A = a - \alpha^{(0)} - \alpha^{(X)}x - \alpha^{(U)}u \mid S = s) \\
&\overset{(b)}{=} \frac{1}{\sqrt{2\pi(\sigma^{(A)})^2}} \exp\left(-\frac{1}{2(\sigma^{(A)})^2}(a - \alpha^{(0)} - \alpha^{(X)}x - \alpha^{(U)}u)^2\right) \\
&\overset{(c)}{=} \frac{1}{\sqrt{2\pi(\sigma^{(A)})^2}} \exp\left(-\frac{1}{2\left(\frac{\sigma^{(A)}}{\alpha^{(U)}}\right)^2}\left((\alpha^{(U)})^{-1}\left(a - \alpha^{(0)} - \alpha^{(X)}x\right) - u\right)^2\right)
\end{aligned}$$

where (b) follows from that $\varepsilon_A \mid (S = s) \sim N(0, (\sigma^{(A)})^2)$. In (c) we reshuffle terms to explicitly break out u inside the exponential function. From inspecting the expression on the last line, we note that it looks like an unnormalized probability density function w.r.t u for a Normal distribution. If we rescale $f(A = a \mid X = x, U = u, S = s)$ by $\frac{1}{\sigma^{(U)}}$, we obtain an probability density function w.r.t u for the Normal distribution

$$N\left((\alpha^{(U)})^{-1}\left(a - \alpha^{(0)} - \alpha^{(X)}x\right), \left(\frac{\sigma^{(A)}}{\alpha^{(U)}}\right)^2\right).$$

This observation together with the results from Lemma 3.2 allows us to show that the product of densities $f(A \mid X, U, S) f(U \mid X, S)$ also corresponds to an unnormalized, scaled probability density function of a Normal distribution with mean equal to

$$\frac{(\alpha^{(U)})^{-1}\left(a - \alpha^{(0)} - \alpha^{(X)}x\right)(\sigma^{(U)})^2 + \mu^{(U)}\left(\frac{\sigma^{(A)}}{\alpha^{(U)}}\right)^2}{(\sigma^{(U)})^2 + \left(\frac{\sigma^{(A)}}{\alpha^{(U)}}\right)^2}. \tag{3.6}$$

Since we can re-normalize $f(A \mid X, U, S) f(U \mid X, S)$ with $1/f(A \mid X, S)$, we have that the conditional expectation

$$\mathbb{E}[U \mid X, A, S = s] = \int u \frac{f(A \mid X, U = u, S = s) f(U = u \mid X, S = s)}{f(A \mid X, S = s)} du$$

is equal to eq. (3.6).

Plugging eq. (3.6) back into eq. (3.5), we have that

$$\mathbb{E}[Y \mid X, A, S = s] = \beta^{(0)} + \beta^{(X)} X + A\left(\beta^{(A)} + \beta^{(AX)} X\right) + \left(\beta^{(U)} + A\beta^{(AU)}\right)\left(\frac{(\alpha^{(U)})^{-1}\left(A - \alpha^{(0)} - \alpha^{(X)} X\right)(\sigma^{(U)})^2 + \mu^{(U)}\left(\frac{\sigma^{(A)}}{\alpha^{(U)}}\right)^2}{(\sigma^{(U)})^2 + \left(\frac{\sigma^{(A)}}{\alpha^{(U)}}\right)^2}\right).$$

To conclude the proof, we simplify the above expression to the form $\mathbb{E}[Y \mid X, A, S = s] = \gamma^\top [1, X, A, AX, A^2]^\top$ where

$$\gamma = \begin{bmatrix} \beta^{(0)} \\ \beta^{(X)} \\ \beta^{(A)} \\ \beta^{(AX)} \\ 0 \end{bmatrix} + \delta \begin{bmatrix} -\beta^{(U)}\left(\frac{\alpha^{(0)}(\sigma^{(U)})^2}{\alpha^{(U)}} - \mu^{(U)}\left(\frac{\sigma^{(A)}}{\alpha^{(U)}}\right)^2\right) \\ -\beta^{(U)}\frac{\alpha^{(X)}(\sigma^{(U)})^2}{\alpha^{(U)}} \\ \beta^{(U)}\frac{(\sigma^{(U)})^2}{\alpha^{(U)}} - \beta^{(AU)}\left(\frac{\alpha^{(0)}(\sigma^{(U)})^2}{\alpha^{(U)}} - \mu^{(U)}\left(\frac{\sigma^{(A)}}{\alpha^{(U)}}\right)^2\right) \\ -\beta^{(AU)}\frac{\alpha^{(X)}(\sigma^{(U)})^2}{\alpha^{(U)}} \\ \beta^{(AU)}\frac{(\sigma^{(U)})^2}{\alpha^{(U)}} \end{bmatrix}.$$

and $\delta = \left((\sigma^{(U)})^2 + \left(\frac{\sigma^{(A)}}{\alpha^{(U)}}\right)^2\right)^{-1}$. □

3.B.3. Proof of Theorem 3.2

Before proving the theorem, we present the following result that we will need later.

Lemma 3.3. *Under the data generating process in eq.* (3.3) *and* $\{\alpha^{(U)} = 0\}$*, we have that* $\mathbb{E}[Y \mid X, A, S = s] = \alpha_s^\top [1, X]^\top$ *and* $\mathbb{E}[Y \mid X, A, S = s] = \gamma_s^\top [1, X, A, AX, A^2]^\top$ *where*

$$\alpha_s = \begin{bmatrix} \alpha_s^{(0)} \\ \alpha_s^{(X)} \end{bmatrix}, \quad \gamma_s = \begin{bmatrix} \beta_s^{(0)} \\ \beta_s^{(X)} \\ \beta_s^{(A)} \\ \beta_s^{(AX)} \\ 0 \end{bmatrix} + \begin{bmatrix} \beta_s^{(U)}\mu_s^{(U)} \\ 0 \\ \beta_s^{(AU)}\mu_s^{(U)} \\ 0 \\ 0 \end{bmatrix}.$$

On the other hand, if instead of $\{\alpha^{(U)} = 0\}$ *we have that* $\{\beta^{(U)} = 0, \beta^{(AU)} = 0\}$ *then*

$$\alpha_s = \begin{bmatrix} \alpha_s^{(0)} + \alpha_s^{(U)}\mu_s^{(U)} \\ \alpha_s^{(X)} \end{bmatrix}, \quad \gamma_s = \begin{bmatrix} \beta_s^{(0)} \\ \beta_s^{(X)} \\ \beta_s^{(A)} \\ \beta^{(AX)} \\ 0 \end{bmatrix}.$$

Proof. For the first case with $\{\alpha^{(U)} = 0\}$, we have that $\mathbb{E}[A \mid X, S = s] = \mathbb{E}[\alpha_s^{(0)} + \alpha_s^{(X)} X + \varepsilon_A \mid S = s] = \alpha_s^{(0)} + \alpha_s^{(X)} X$. Further, we can show

$$\begin{aligned}\mathbb{E}[Y \mid X, A, S = s] &= \beta_s^{(0)} + \beta_s^{(X)} X + \beta_s^{(A)} A + \beta_s^{(AX)} AX + \mathbb{E}[\beta_s^{(U)} U + \beta_s^{(AU)} AU \mid X, A, S = s] \\ &= \beta_s^{(0)} + \beta_s^{(X)} X + \beta_s^{(A)} A + \beta_s^{(AX)} AX + \beta_s^{(U)} \mu_s^{(U)} + \beta_s^{(AU)} \mu_s^{(U)} A\end{aligned}$$

where the second equality from that $U \perp\!\!\!\perp X \mid A, S = s$ holds in eq. (3.3) if $\{\alpha^{(U)} = 0\}$. This concludes the first case.

For the second case with $\{\beta^{(U)} = 0, \beta^{(AU)} = 0\}$, it follows from the above equation that

$$\mathbb{E}[Y \mid X, A, S = s] = \beta_s^{(0)} + \beta_s^{(X)} X + \beta_s^{(A)} A + \beta_s^{(AX)} AX \,.$$

Meanwhile, for the treatment mechanism we now instead have that

$$\begin{aligned}\mathbb{E}[A \mid X, S = s] &= \alpha_s^{(0)} + \alpha_s^{(X)} X + \mathbb{E}[\alpha_s^{(U)} U \mid S = s] \\ &= \alpha_s^{(0)} + \alpha_s^{(X)} X + \alpha_s^{(U)} \mu_s^{(U)} \,.\end{aligned}$$

□

Now, we can proceed with the proof of the Theorem 3.2.

Proof. Throughout the proof, we define $\phi(X) = [1, X]^\top$ and $\psi(X) = [1, X, A, AX, A^2]^\top$. We will show that regardless of the presence of the confounder U, we can write $\mathbb{E}[A \mid X, S = s] = \omega_s^\top \phi(X)$ and $\mathbb{E}[Y \mid X, A, S = s] = \gamma_s^\top \psi(X, A)$ for some (ω_s, γ_s) and only if and only if U is a confounder will $\omega_s \not\perp\!\!\!\perp \gamma_s$ under some conditions on the parameters $(\alpha_s^{(0)}, \alpha_s^{(X)}, \alpha_s^{(U)}, \mu_s^{(U)})$.

Note that under Assumption 3.2, it follows that there exists a distribution $(\omega_s, \gamma_s) \sim P(\omega, \gamma)$ since the parameters (ω_s, γ_s) are directly dependent on the parameters $(\alpha_s, \beta_s) \sim P(\alpha, \beta)$, for $s = 1, \ldots, K$. To determine if $P(\omega, \gamma) = P(\omega)P(\gamma)$, we have to determine whether ω_s and γ_s both depend on the same parameters from the underlying data-generating process and under what conditions this can create statistical dependencies between ω_s and γ_s.

No unmeasured confounding present First, consider the condition that U is not a confounder. There are three cases for which this happens: (1) we have $\{\alpha^{(U)} = 0\}$, (2) we have $\{\beta^{(U)} = 0, \beta^{(AU)} = 0\}$, and (3) we have both $\{\alpha^{(U)} = 0\}$ and $\{\beta^{(U)} = 0, \beta^{(AU)} = 0\}$. For case (1) and (2), it follows immediately from Lemma 3.3 that ω_s and γ_s have no shared parameters. For the final case (3), it is easy to see that $\omega_s = [\alpha_s^{(0)}, \alpha_s^{(X)}]^\top$ and $\gamma_s^\top = [\beta_s^{(0)}, \beta_s^{(X)}, \beta_s^{(A)}, \beta^{(AX)}, 0]$ where, again, there are no shared parameters between ω_s and γ_s. Thus, we can conclude that under all of the cases when U is not a confounder, (ω_s, γ_s) have no shared parameter and thus $\omega_s \perp\!\!\!\perp \gamma_s$.

Unmeasured confounding present Next, consider the condition that U is a confounder. It follows from Lemma 3.1 that if U is a confounder, then both ω_s and γ_s depend on the parameters $(\alpha_s^{(0)}, \alpha_s^{(X)}, \alpha_s^{(U)}, \mu_s^{(U)})$. Thus, if any of the parameters $(\alpha_s^{(0)}, \alpha_s^{(X)}\alpha_s^{(U)}, \mu_s^{(U)})$ vary across different environments, which happens if we assume there exist a non-degenerate distribution for at least of one these parameters, it follows that $\omega_s \not\perp\!\!\!\perp \gamma_s$. This concludes the proof. □

3.C. Extension to implicit feature representations

In this section, we provide a sketch for replacing the features representation $\{\widetilde{\phi}, \widetilde{\psi}\}$ in our falsification algorithm with an implicit feature representation through the use of kernel methods (Schölkopf & Smola, 2002). More specifically, we let $\widetilde{\phi}(x)$ be the implicit feature representation whose inner product is given by the kernel $k(x,x) = \langle \widetilde{\phi}(x), \widetilde{\phi}(x) \rangle_{\mathscr{H}}$ defined on $\mathscr{X}$ with corresponding RKHS $\mathscr{H}$ and, similarly, let $\widetilde{\psi}(x,a)$ be the implicit feature representation with inner product given by $h((x,a),(x,a)) = \langle \widetilde{\psi}(x,a), \widetilde{\psi}(x,a) \rangle_{\mathscr{G}}$ defined on $\mathscr{X} \otimes \mathscr{A}$ with corresponding RKHS $\mathscr{G}$.

With some minor modifications, we can run our falsification algorithm without having to compute $\widetilde{\phi}(x)$ and $\widetilde{\psi}(x)$. To illustrate this, we impose the restriction that we use the same number of observations from each environment, denoted with n. We shall focus on estimators given by $\widehat{\omega}_s = \arg\min_\omega = ||\mathbf{A}_s - \Phi_s\omega||_2^2 + \lambda_1||\omega||_2$ and $\widehat{\gamma}_s = \arg\min_\omega = ||\mathbf{Y}_s - \Psi_s\gamma||_2^2 + \lambda_2||\gamma||_2$ for some constants $\lambda_1, \lambda_2 > 0$. The above optimization problems correspond to kernel ridge regression problem, for which it is well-known that the estimators have a closed form solution, namely

$$\widehat{\omega}_s = (K_s + n\lambda_1 I_n)^{-1}\mathbf{A}_s \text{ and } \widehat{\gamma}_s = (H_s + n\lambda_2 I_n)^{-1}\mathbf{Y}_s$$

where $K_s = k(\mathbf{X}_s, \mathbf{X}_s)$ and $H_s = h((\mathbf{X}_s, \mathbf{A}_s), (\mathbf{X}_{s'}, \mathbf{A}_{s'}))$ are the Gram matrices. The above estimators are however not always be computable since they can, depending on the choice of kernel, be infinite-dimensional. This also makes it infeasible to directly compute the covariance matrix $\Sigma = \text{Cov}(\omega, \gamma)$. However, it is luckily still possible to compute the Frobenius norm $||\Sigma||_2$. Using results from Lemma 1 in Gretton *et al.* (2005), we can rewrite $||\Sigma||_2 = \mathbb{E}_{P(\omega,\gamma)}[\omega^\top\omega\,\gamma^\top\gamma] + \mathbb{E}_{P(\omega)}[\omega^\top\omega]\mathbb{E}_{P(\gamma)}[\gamma^\top\gamma] - 2\mathbb{E}_{P(\omega,\gamma)}[\mathbb{E}_{P(\omega)}[\omega^\top\omega]\mathbb{E}_{P(\gamma)}[\gamma^\top\gamma]]$. From this equality, it follows that we can compute $||\Sigma||_2$ by inspecting the inner products $\widehat{\omega}_s^\top\widehat{\omega}_{s'}$ and $\widehat{\gamma}_s^\top\widehat{\gamma}_{s'}$ for all $s, s' \in \{1, \ldots, K\}$. Interestingly, these inner products can be computed as follows

$$\widehat{\omega}_s^\top\widehat{\omega}_{s'} = \mathbf{A}_s^\top(K_s^\top + n\lambda_1 I_n)^{-1}(K_{s'} + n\lambda_1 I_n)^{-1}\mathbf{A}_{s'}$$
$$\widehat{\gamma}_s^\top\widehat{\gamma}_{s'} = \mathbf{Y}_s^\top(H_s^\top + n\lambda_1 I_n)^{-1}(H_{s'} + n\lambda_1 I_n)^{-1}\mathbf{Y}_{s'}$$

which means that $||\Sigma||_2$ can be computed and we can in principle statistically test for independence of ω and γ with some implicit feature representations $\{\widetilde{\phi}, \widetilde{\psi}\}$. For future work, it remains unknown how to best implement this algorithm and investigate how our theory needs to be modified for it.

3.D. Experimental details

3.D.1. Sampling from data-generating process with polynomial basis functions

We generated observational datasets as follows. For each environment $s = 1, \dots, K$, we obtain $i = 1, \dots, N$ individuals by first sampling a set of d-dimensional covariates according to $X_i \sim N(\mu_s^{(X)}, \frac{1}{\sqrt{d}}\Sigma)$ with the mean $\mu_s^{(X)} \in \mathbb{R}^d \sim N(\mathbf{0}, \frac{1}{4}\mathbf{I}_d)$ where $\mathbf{I}_d$ was a $d \times d$ identity matrix and the covariance matrix Σ of shape $d \times d$ had its diagonal elements set to 2 and its off-diagonal elements set to 0.1. Thereafter, we sampled the treatment T_i and outcome Y_i according to eq. (3.1) with the features representations $\psi(X) = [1, X_1, \dots, X_d, X_1^p, \dots, X_d^p]^\top$ and $\phi(X, A) = [1, X_1, \dots, X_d, X_1^p, \dots, X_d^p, A]^\top$ being polynomial basis functions of degree p. The noise variables ε_A and ε_Y were mean-zero Normal distributed with their standard deviation set to 0.5. Each element in α_s where sampled uniformly from the set $\{-1, 1\}$ while each element in β_s was set to 1, except for the elements in α_s corresponding to the intercepts which were sampled according to $N(0, 1)$. Only the intercept elements were resampled for each new environment, whereas the remaining coefficients in (α_s, β_s) where kept fixed for all environments. When introducing an unmeasured confounder, we additionally sampled an one-dimensional covariate $U_i \sim N(\mu_s^{(U)}, 2)$ with its mean $\mu_s^{(U)} \sim N(0, 1)$. Then, we added U_i directly to T_i and Y_i. For simplicity, we let each environment have the same number of samples $N = n_1 = \cdots = n_K$ even though all methods also work if the number of samples per environment differ.

3.D.2. Sampling from data-generating process in eq. (3.3)

We sampled a multi-environment dataset with $K = 250$ environments and 1000 samples per environment according to the data-generating process described in Section 3.4.3:

$$\begin{aligned} A &= \alpha_s^\top \psi(X) + \alpha_s^{(U)} U + \varepsilon_A \\ Y^a &= \beta_s^\top \phi(X, A = a) + \left(\beta_s^{(U)} + a\beta_s^{(AU)}\right) U + \varepsilon_Y \end{aligned} \tag{3.7}$$

where we let $\psi(X) = [1, X]^\top$ and $\phi(X, A) = [1, X, A, AX]^\top$, the noise variables were sampled according to $\varepsilon_A \sim N(0, \frac{1}{8})$ and $\varepsilon_Y \sim N(0, \frac{1}{8})$, and the covariates were sampled according $X \sim N(\mu_X, 1)$ and $U \sim N(\mu_U, 1)$. By default, we set the parameters as $\alpha_s = [\frac{1}{2}, \frac{1}{3}]^\top$, $\beta_s = [\frac{1}{2}, \frac{1}{3}, \frac{1}{2}, \frac{1}{3}]^\top$, $\mu_X = 1$, and $\mu_U = 1$. To impose the presence of an unmeasured confounder, we would set the remaining parameters $(\alpha_s^{(U)}, \beta_s^{(U)}, \beta_s^{(AU)})$ to $\frac{1}{4}$ and otherwise set them to 0. To introduce changes in the parameters among environments, we would select one of the parameter values and overrule the above default values by sampling from a uniformly from the range $[0.1, 3.0]$ for each new environment.

3.D.3. Generating Twins semi-synthetic dataset

We use data from twin births in the USA between 1989-1991 (Almond *et al.*, 2005) to construct an multi-environment observational dataset with a known causal structure. The dataset contains 46 covariates related to pregnancy, birth, and parents. As many covariates are highly imbalanced and have low variance, we select a subset of the covariates for generating the semi-synthetic dataset.

As the environment label we used the birth state and as covariates we used the following variables (variable names from the dataset documentation are shown in parenthesis): birth month (birmon), father's age (dfageq), number of live births before twins (dlivord_min), total number of births before twins (dtotord_min), gestation type (gestat10), mom's age (mager8), mom's education (meduc6), mom's place of birth (mplbir), and number of prenatal visits (nprevistq).

The treatment and outcome were generated using the same procedure described in Section 3.D.1, with one key difference: the synthetic covariates were replaced by real-world covariates from the Twins dataset. Prior to generating the treatment and outcome, the covariates were standardized. Each time a semi-synthetic dataset was created using the Twins dataset covariates, five of the chosen covariates were randomly selected as the confounders. A polynomial degree of $p = 2$ was consistently used throughout all experiments.

3.D.4. Kernel conditional independence testing with the HGIC approach

When using the original implementation of the hierarchical graph independence constraint (HGIC) approach described in Karlsson and Krijthe (2023) as a baseline in our experiments, we noted that their implementation sometimes would not be properly calibrated (i.e., elevated Type 1 error above $\alpha = 0.05$). For this reason, we introduced some modifications to their method that resolved this issue. Note that the modification we describe below were only necessary when using HGIC with the kernel conditional independence test, and not the Pearson conditional independence test used in some other experiments.

The elevated Type 1 errors was caused by that HGIC combines p-values from multiple independence tests using Fisher's method, which employs the test statistic $T = \sum_k \log p_k$ where p_k where are the p-values from the tests (Fisher, 1925). This modification allowed HGIC to use all observations in the multi-environment dataset and was observed to help increase the falsification test's power. However, we noticed in our own simulations that a poorly calibrated conditional independence test can cause the combination of p-values to amplify type I errors. So to address this issue, we made two modifications to the original HGIC implementation.

First, we improved calibration by switching to permutation-based calibration, replacing the original Gamma distribution approximation that is also commonly used for KCIT. Furthermore, we adopted the KCIT implementation from the *causal-learn* Python pack-

age (Zheng *et al.*, 2024) which allowed for optimizing the kernel width hyperparameter in the test using Gaussian process regression.

Second, although the above changes improved KCIT's calibration, Fisher's method still sometimes amplified type I errors. To mitigate this, we replaced it with Tippett's method, which uses $T = \min_k p_k$ as the test statistic (Tippett, 1931). Like Fisher's method, Tippett's method emphasizes the smallest p-values (Heard & Rubin-Delanchy, 2018) but we found Tippett's method to be more conservative under the null. Testing on a simple conditional independence scenario confirmed that Tippett's method worked better with KCIT, while retaining the benefits of increased power in combining p-values.

To illustrate the difference between our modified implementation and the original implementation described in Karlsson and Krijthe (2023), we include an experiment using the data-generating process with polynomial basis functions described in Appendix 3.D.1. We used $K = 100$ environments with $N = 50$ samples per environment and $d = 1$ observed confounder, and set the polynomial degree to $p = 2$. Here, both implementations combine 25 p-values. As the results in Table 3.1 show, our implementation achieved a better Type 1 error while retaining similar power to the original implementation.

3.D.5. Additional experiments

We have included three additional experiments in this section to complement the experiments in the main paper.

First, repeating the same setup as in the experiment presented in Section 3.6.2, we include results that confirmed that all methods have controlled Type 1 errors in the well-specified setting. These results are shown in Figure 3.6.

Secondly, we conducted an ablation study to highlight the importance of using bootstrapping on top of the permutation-based test in our procedure. We repeated the same setup as in Section 3.6.2, but implemented permutation-based testing without bootstrapping. As shown in Figure 3.5, bootstrapping is crucial for maintaining Type 1 errors below $\alpha = 0.05$, even with an increased sample size.

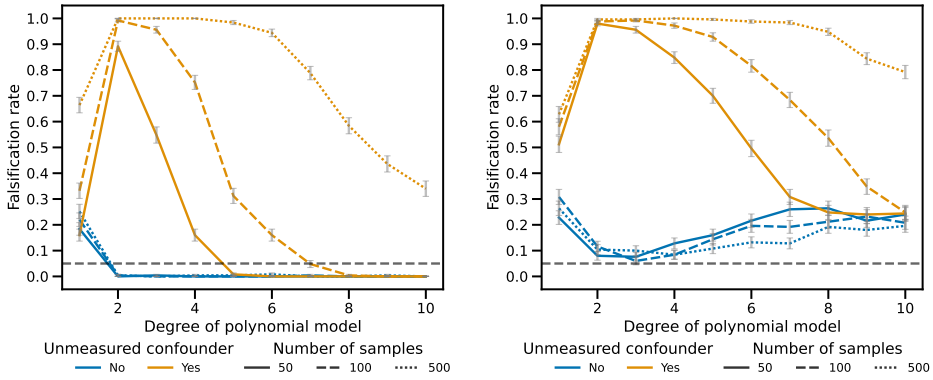
Lastly, we compared all methods under misspecification across several scenarios in the data-generating process described in Appendix 3.D.1. These scenarios included the presence or absence of an unmeasured confounder, whether transportability holds by sampling the intercept term in β_s from $N(0,1)$ across environments, and whether the underlying data-generating process (DGP) was linear ($p = 1$) or nonlinear ($p = 3$) The results in Table 3.2 show that misspecification led to elevated Type 1 errors (falsifying without an unmeasured confounder) for all methods. Additionally, transportability violations caused higher Type 1 errors for the transportability test, while our proposed algorithm remained unaffected.

Table 3.1: Comparison of the new and old HGIC implementation using the data-generating process with polynomial basis functions. The average falsification rate and standard error (in parenthesis) is reported from 250 repetitions.

Method	No unmeasured confounder	Unmeasured confounder present
Modified HGIC implementation	0.04 (.01)	0.88 (.02)
Original HGIC implementation	0.28 (.03)	0.80 (.03)

Table 3.2: Comparison of different approaches under various scenarios with $K = 100$ environments and $N = 100$ samples per environment. The average falsification rate and standard error (in parenthesis) is reported from 250 repetitions.

	No unmeasured confounder				Unmeasured confounder present			
Transportability	Holds		Violated		Holds		Violated	
DGP	Cubic	Linear	Cubic	Linear	Cubic	Linear	Cubic	Linear
MINT (Linear)	0.62 (.03)	0.01 (.01)	0.65 (.03)	0.05 (.01)	0.60 (.03)	1.00 (.00)	0.53 (.03)	1.00 (.00)
MINT (Cubic)	0.00 (.00)	0.02 (.01)	0.04 (.01)	0.06 (.01)	1.00 (.00)	1.00 (.00)	1.00 (.00)	1.00 (.00)
Transp. test (Pearson)	0.69 (.03)	0.04 (.01)	0.68 (.03)	0.81 (.02)	0.65 (.03)	0.75 (.03)	0.71 (.03)	0.86 (.02)
Transp. test (KCIT)	0.05 (.01)	0.07 (.02)	0.38 (.03)	0.43 (.03)	0.16 (.02)	0.24 (.03)	0.34 (.03)	0.42 (.03)
HGIC (Pearson)	0.66 (.03)	0.03 (.01)	0.58 (.03)	0.04 (.01)	0.44 (.03)	1.00 (.00)	0.43 (.03)	1.00 (.00)
HGIC (KCIT)	0.02 (.01)	0.02 (.01)	0.11 (.02)	0.02 (.01)	0.76 (.03)	0.99 (.01)	0.35 (.03)	0.70 (.03)



(a) With bootstrapping

(b) Without bootstrapping

Figure 3.5: An ablation study showing the falsification rate our proposed algorithm using permutation-based testing with bootstrapping versus without bootstrapping. We resample 1000 times when using bootstrapping. The error bars show the standard error over 250 repetitions. The black dotted lines correspond to the chosen significance level $\alpha = 0.05$.

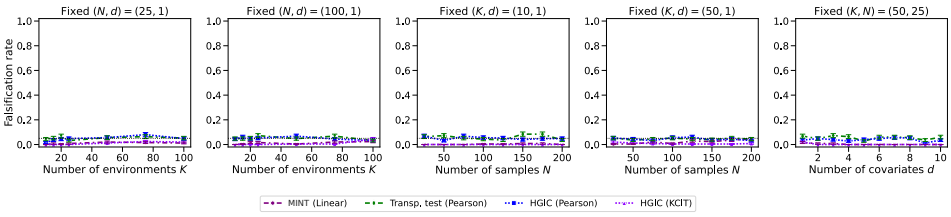


Figure 3.6: Same experiment as in Figure 3.1 but with no unmeasured confounding being present. Comparison of falsification rate when varying either the number of environment K, the number of samples per environment N, or the number of observed covariates d. The error bars show the standard error over 250 repetitions.

4

Falsification of Front-Door and IV Approaches

We study the problem of falsifying the assumptions behind a set of broadly applied causal identification strategies: namely back-door adjustment, front-door adjustment, and instrumental variable estimation. While these assumptions are untestable from observational data in general, we show that with access to data coming from multiple heterogeneous environments, there exist novel independence constraints that can be used to falsify the validity of each strategy. Most interestingly, we make no parametric assumptions, instead relying on that changes between environments happen under the principle of independent causal mechanisms.

4.1. Introduction

A common theme within the field of causal inference has been to study settings with data collected from multiple environments. This type of data often tends to be heterogeneous due to e.g. changing circumstances or time shifts. While data heterogeneity is sometimes seen as an obstacle in data science, it is possible to turn it to one's advantage. For instance, it can allow us to learn invariant predictors that better generalize to unseen environments (Peters *et al.*, 2016; Rothenhäusler *et al.*, 2021), improve causal discovery (Ghassami *et al.*, 2018; B. Huang *et al.*, 2020; Mooij *et al.*, 2020), and enable new causal effect identification strategies (Athey *et al.*, 2020; Bareinboim & Pearl, 2016). In this paper, we focus on the last two ideas together.

We study the problem of falsifying a set of broadly applied graphical conditions under the possible presence of latent variables: namely the *1. back-door criterion*, *2. front-door*

This chapter appears as: Karlsson, R., Creastă, Ș., & Krijthe, J. H. (2023). Putting causal identification to the test: Falsification using multi-environment data. *Causal Representation Learning Workshop at NeurIPS 2023*

criterion, and *3. instrumental variable criterion*. These conditions are crucial when we want to estimate the effects of interventions from observational data. Unfortunately, in the most general case, these conditions can not be verified from a single observational dataset alone (Pearl, 2009). However, we will show that when we have multiple datasets stemming from different environments or clusters – such as different locations, time periods, or studies – some of these conditions can be tested.

Our contribution is to demonstrate that a novel type of independence constraints (Guo *et al.*, 2023; Karlsson & Krijthe, 2023) can be used to falsify the above-mentioned conditions when we have access to multi-environment data under the assumption of independent causal mechanisms (Peters *et al.*, 2017; Schölkopf *et al.*, 2012, 2021). In particular, we aim to do this without access to interventional data. We believe our findings are of direct interest to those who want to test the validity of their causal identification strategy and have access to multi-environment data. However, the technique we use to obtain our results may be of independent interest to the broader causality community.

4.2. Related works

This paper contributes to the growing body of literature on doing causal inference from heterogeneous, multi-environment data (Bareinboim & Pearl, 2016; B. Huang *et al.*, 2020; Mooij *et al.*, 2020; Peters *et al.*, 2016; Shi *et al.*, 2021; Squires *et al.*, 2023). Most closely related to our work are Guo *et al.* (2023) and Karlsson and Krijthe (2023). Assuming independent causal mechanisms, these works showcase novel independence constraints that can be used for causal discovery in multi-environment settings. Guo *et al.* (2023) focused on the setting with all variables observed (i.e. having causal sufficiency): they show in this case that we can go beyond the Markov Equivalence Class and uniquely determine the causal DAG from observational data. Meanwhile, in a similar setting, Karlsson and Krijthe (2023) relaxed the causal sufficiency assumption and showed how to detect the presence of latent confounders. The technique used in both works shares similarities to the twin network method for counterfactual reasoning by Balke and Pearl (1994) by looking at independence constraints in a "twinned" graph. In contrast to Balke and Pearl (1994), this "twinning" technique is applied to a setting with different environments having the same causal structure. We build further on these developments, showing new non-parametric identification results for widely applied identification strategies.

In this paper, we explore the possibilities for falsification implied by the independent causal mechanism assumption. There do however also exist other techniques for falsification. For instance, under mild conditions involving an auxiliary variable, Bhattacharya and Nabi (2022) demonstrate testable conditions for the front-door criterion. In addition, there are the well-known instrumental inequalities that sometimes can falsify the validity of instrumental variables (Kédagni & Mourifie, 2017; Pearl, 1995). We believe our work can be used together with previously proposed tests like the ones mentioned, strengthening our toolbox to (in certain cases) falsify our causal assumptions.

4.3. Problem setting

We start with some preliminaries of the causal terminology used in this paper.

Definition 4.1 (Causal Graphical Model (CGM)). *A causal graphical model $M = (\mathcal{G}, P)$ over d random variables $\mathbf{V} = (V_1, V_2, \ldots, V_d)$ comprises (i) a directed acyclic graph (DAG) $\mathcal{G}$ with vertices $\mathbf{V}$ and edges $V_j \to V_j'$ iff V_j is a direct cause of V_j', and (ii) a joint distribution P such that it has the following Markov or causal factorization over $\mathcal{G}$:*

$$P(V_1, V_2, \ldots, V_d) = \prod_{j=1}^{d} P(V_j \mid \mathrm{Pa}(V_j)) \tag{4.1}$$

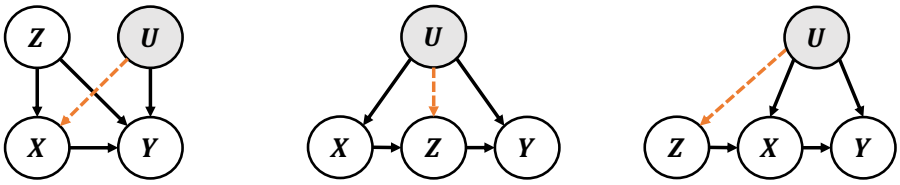
where $\mathrm{Pa}(V_j)$ denotes the parents (direct causes) of V_j in $\mathcal{G}$ and $P(V_j \mid \mathrm{Pa}(V_j))$ is the causal mechanism of V_j.

The DAG $\mathcal{G}$ encodes various conditional independences (or d-separations) between the variables which we write as $\mathbf{A} \perp\!\!\!\perp_d \mathbf{B} \mid \mathbf{C}$ over some disjoint sets of variables $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$. We shall assume that conditional independencies in $\mathcal{G}$ imply the same conditional independencies in P, and vice versa.

Assumption 4.1 (Faithfulness & Causal Markov Property). *For P and $\mathcal{G}$ we have (i) the faithfulness property that $\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \Rightarrow \mathbf{A} \perp\!\!\!\perp_d \mathbf{B} \mid \mathbf{C}$, and (ii) the causal Markov property that $\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \Leftarrow \mathbf{A} \perp\!\!\!\perp_d \mathbf{B} \mid \mathbf{C}$.*

We will consider a setting with the following variables: We have treatment $X \in \mathcal{X}$ and outcome $Y \in \mathcal{Y}$, as well as an auxiliary variable $Z \in \mathcal{Z}$. In addition, we allow the presence of an unobserved confounder $U \in \mathcal{U}$ between X and Y. We shall further assume that we know that $Y \notin \mathrm{Ancestors}(X)$. This setting comes up when we are interested in estimating the interventional effect of X on Y, denoted as $P(Y \mid \mathrm{do}(X))$ using do-calculus (Pearl, 2009); here we are often sure X "happens" before Y and we wish to learn if X has an effect on Y.

Depending on how (X, Y) relates with respect to Z, we can use different strategies to estimate the interventional effect from observational data: back-door adjustment if Z fulfills the back-door criterion; front-door adjustment if Z is a mediator fulfilling the front-door criterion; or instrumental variable estimation if Z is a valid instrument. These different settings are illustrated in Figure 4.1. While domain knowledge often informs us which strategy to use, no independence constraint exists between (X, Y, Z) that allows us to verify any of these conditions (Pearl, 1995). We demonstrate that such conditions exist, however, when we have data from multiple environments and assume independent causal mechanisms. We will now formalize this assumption.



(a) Back-door (b) Front-door (c) Instrumental variable

Figure 4.1: Three common settings where observing Z allows for identification of $P(Y \mid \mathrm{do}(X))$; the shaded variables are unobserved. The addition of the dashed red arrow illustrates one way in which Z becomes insufficient for identification.

4.3.1. Assumptions for multi-environment data

We have observational datasets from multiple environments e_k, indexed by $k = 1, \ldots, K$. The datasets are sampled as $(X_i^{(k)}, Y_i^{(k)}, Z_i^{(k)}, U_i^{(k)}) \sim P^{(e_k)}(X, Y, Z, U)$ for $i = 1, \ldots, N_k$, where N_k is the number of observations in environment e_k. Note that in what follows, $U_i^{(k)}$ is not observed. We allow each environment to have a different joint distribution $P^{(e_k)}$ but assume they are related to each other through the following assumption:

Assumption 4.2 (Shared Causal Graph)**.** *All environments share the same causal DAG $\mathcal{G}$.*

Next, we specify how changes in $P^{(e_k)}(X, Y, Z, U)$ arise between the different environments. We shall assume that the conditional probabilities in eq. (4.1) – which we refer to as causal mechanisms – vary independently per environment. This is known as the independent causal mechanism principle (Peters *et al.*, 2017). We shall now describe the assumption that operationalizes this.

To model changes between environments with independent causal mechanisms, we parameterize each causal mechanism with a parameter $\Theta = \{\Theta_V \in \mathcal{O}_V : V \in \{X, Y, Z, U\}\}$.[1] In each environment, these are fixed and determine the distribution $P^{(e_k)}(X, Y, Z, U \mid \Theta) = \prod_{V \in \{X,Y,Z,U\}} P^{(e_k)}(V \mid \mathrm{Pa}(V), \Theta_V)$. One could see changes in Θ as different soft interventions on the causal mechanisms, similar to the settings considered by B. Huang *et al.* (2020) and Perry *et al.* (2022).

Further, we shall assume that environments are randomly sampled from a *distribution over mechanisms* by defining non-degenerate probability measures for each causal mechanism.

[1] For our intended purpose, note that we do not have to specify the explicit form of parameterization. This also means that, in principle, we do not specify the dimensionality of these parameters. While it is perhaps easier to imagine what independence between parameters looks like in the finite-dimensional case, one could also consider independence between infinite-dimensional parameters. This concept has been rigorously studied in nonparametric Bayesian inference, where one often constructs a prior over independent parameters (Ghosal & Van der Vaart, 2017).

Assumption 4.3 (Stochastic Independent Causal Mechanisms)**.** *The parameters Θ_V of the causal mechanisms are pair-wise independent random variables with non-degenerate probability measures $P(\Theta_V)$ for all $V \in \{X, Y, Z, U\}$.*

With the above assumption, when we say *independent* causal mechanisms, we refer to statistical independence between them. This independence is a strong assumption to make, which we will see gives us new testable implications in the data.

4.4. Testing causal identification strategies with multi-environment data

We are now ready to present the main theoretical tool that we will use: a hierarchical causal graphical model that incorporates the multi-environment structure of the data under our assumptions. Crucially, in contrast to the single-environment causal graphical model, the hierarchical graph encodes additional independence constraints among the observed variables. These can be used to falsify causal identification strategies. We start with the definition of the hierarchical causal graphical model before we go into examples and results using this model.

Definition 4.2 (Hierarchical Causal Graphical Model)**.** *For a given environment e_k, we have the causal graphical model $M^{(e_k)} = (P^{(e_k)}, \mathcal{G})$ with variables $\mathbf{V}^{(k)} = (V_1^{(k)}, V_2^{(k)}, \ldots, V_d^{(k)})$. We define the hierarchical causal graphical model $M^* = (P^*, \mathcal{G}^*)$ as follows: Let $\mathcal{G}^*$ be a DAG containing vertices $\{\mathbf{V}_i^{(k)} : k = 1, \ldots, K\}$ for all observations $i = 1, \ldots, N_k$. It has the edge $V_{i,j}^{(k)} \to V_{i,j'}^{(k)}$ for all i iff the same edge exists in $\mathcal{G}$ where $j, j' = 1, \ldots, d$. Furthermore, we posit the causal mechanism parameters $\Theta^{(k)} = (\Theta_{V_1}^{(k)}, \Theta_{V_2}^{(k)}, \ldots, \Theta_{V_d}^{(k)})$ to $\mathcal{G}^*$ so that $\Theta_{V_j}^{(k)} \to V_{i,j}^{(k)}$ for every i, j and k. The joint distribution P^* over all variables in $\mathcal{G}^*$ factorizes as*

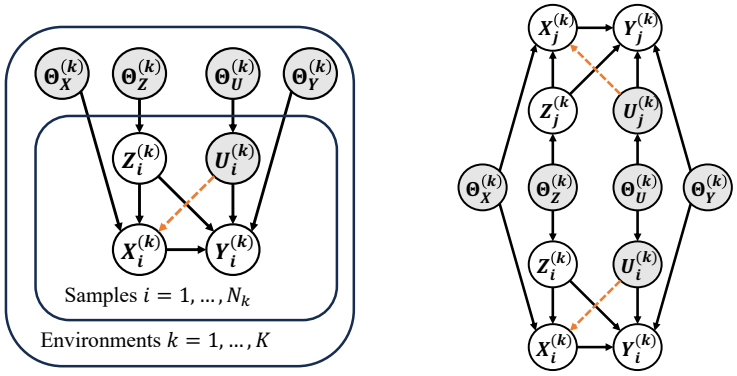
$$\prod_{k=1}^{K} \prod_{i=1}^{N_k} \prod_{j=1}^{d} P^*(V_{i,j}^{(k)} \mid \mathrm{Pa}(V_{i,j}^{(k)}), \Theta_{V_j}^{(k)}) P^*(\Theta_{V_j}^{(k)}) \tag{4.2}$$

where $P^*(V_{i,j}^{(k)} \mid \mathrm{Pa}(V_{i,j}^{(k)}), \Theta_{V_j}) = P^{(e_k)}(V_{i,j}^{(k)} \mid \mathrm{Pa}(V_{i,j}^{(k)}), \Theta_{V_j})$.

To illustrate why the hierarchical causal graphical model is helpful, we first revisit a result from Karlsson and Krijthe (2023), showing how it can be used to falsify the back-door criterion.

4.4.1. Testing the back-door criterion

Let all environments share the graph $\mathcal{G}$ from Figure 4.1a and construct its corresponding hierarchical DAG $\mathcal{G}^*$, seen in Figure 4.2a. For the original DAG $\mathcal{G}$, it is well-known that there exist no independence constraints between the observed variables for testing the



(a) (b)

Figure 4.2: **(a)**: The hierarchical causal graphical model for the DAG from Figure 4.1a. **(b)**: We unfold the hierarchical causal graphical model to obtain a "twin" structure. This allows us to study the dependency structure between two different observations (i, j) from the same environment k.

presence of the unobserved confounder U (Pearl, 1995). For the hierarchical $\mathcal{G}^*$, however, we will see that such constraints exist.

In graph $\mathcal{G}^*$, we can study dependencies between two different samples (i, j) within an environment k: that is, $(X_i^{(k)}, Y_i^{(k)}, Z_i^{(k)})$ and $(X_j^{(k)}, Y_j^{(k)}, Z_j^{(k)})$ where $i \neq j$. Interestingly, if we do not condition on the environment – or conversely, the causal mechanism parameters Θ in $\mathcal{G}^*$ – these two samples are dependent as they share parents in $\mathcal{G}^*$. This is illustrated in Figure 4.2b, where we unfold the hierarchical structure; or, one could say that we have created a "twin" of the original graph.

Now, one can verify graphically that if the dashed arrow is absent in the graph in Figure 4.2b then

$$X_i^{(k)} \perp\!\!\!\perp_{P^*} Y_j^{(k)} \mid X_j^{(k)}, Z_i^{(k)}, Z_j^{(k)} \,. \tag{4.3}$$

But if the dashed arrow is present, such that Z becomes an invalid back-door adjustment set as we have an open backdoor path between X and Y, then the independence in eq. (4.3) is violated. As eq. (4.3) only contains observed variables, the back-door criterion has testable implications according to the hierarchical model. In fact, this statement is true even if we consider a larger set of possible graphs.

Theorem 4.1 (Karlsson and Krijthe (2023))**.** *Consider Assumption 4.1, 4.2 and 4.3 where $Y \notin Ancestors(X)$ and that there is no selection bias. Let $\mathcal{G}$ be the shared causal DAG across environments and $\mathcal{G}^*$ its corresponding hierarchical DAG. Then, we have that eq.* (4.3) *holds for any k and $i \neq j$ in $\mathcal{G}^*$ iff Z blocks every back-door path between X and Y in G.*

How do we test this independence? We have shown here that opening a back-door path leads to the violation of a new independence constraint in the observed data distribution. For the rest of the paper, our goal is to provide more of these identification results, while constructing efficient tests for these dependencies is outside the scope of our paper. This problem has been studied by both Guo *et al.* (2023) and Karlsson and Krijthe (2023). For the interested reader, however, we provide an explanation of how to test independencies such as eq. (4.3) in the Appendix.

4.4.2. Testing the front-door criterion

The next graphical condition we will explore is the front-door criterion where Z is a mediator between X and Y, as demonstrated in Figure 4.1b and which is defined as follows:

Definition 4.3 (Front-door criterion (Pearl, 2009))**.** *A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) in a DAG $\mathcal{G}$ if: (i) Z intercepts all directed paths from X to Y; (ii) there is no unblocked back-door path from X to Z; and (iii) all back-door paths from Z to Y are blocked by X.*

If we know the causal ordering of (X, Z, Y), then we see that Theorem 4.1 can be directly applied to construct testable implications for both (ii) and (iii) in Definition 4.3.

Corollary 4.1. *Consider assumption 4.1, 4.2 and 4.3 with $Y \notin$ Ancestors(Z), $Z \notin$ Ancestors(X), and no selection bias, let $\mathcal{G}$ be the shared causal DAG across environments and $\mathcal{G}^*$ its corresponding hierarchical DAG. Then, for any k and $i \neq j$,*

$$X_i^{(k)} \perp\!\!\!\perp_{P^*} Z_j^{(k)} \mid X_j^{(k)} \text{ and } Z_i^{(k)} \perp\!\!\!\perp_{P^*} Y_j^{(k)} \mid Z_j^{(k)}, X_i^{(k)}, X_j^{(k)} \tag{4.4}$$

iff condition (ii) and (iii) in Definition 4.3 hold true for $\mathcal{G}$.

Proof. We apply Theorem 4.1 twice, noting that conditions (ii) and (iii) concern that there exist no unblocked back-door paths. For (ii), we need to check that there is no open back-door path between the ordered pair (X, Z) with an empty adjustment set; this results in the first independence. Similarly, for (iii), we get the second independence by having to check whether X is sufficient to block any back-door path between the ordered pair (Z, Y). □

Starting on a positive note, we have shown that it is in fact possible to verify two out of three conditions in the front-door criterion. While the "twinning" technique is very suitable to detect open back-door paths, we will see now that testing the remaining condition – whether Z intercepts all directed paths between (X, Y) – is more difficult to test; in fact, it is impossible to do it with this technique.

Theorem 4.2. *Consider the same assumptions as in Corollary 4.1 with $\mathcal{G}$ being the shared causal DAG across environments and $\mathcal{G}^*$ its corresponding hierarchical DAG. then there exist no independence constraints in $\mathcal{G}^*$ that imply whether condition (i) in Definition 4.3 holds.*

A proof of the theorem is provided in the Appendix.

4.4.3. Testing the instrumental variable criterion

Now, we turn our attention to the third identification strategy to see whether it is possible to reject the validity of an instrumental variable. As before, we start with the graphical definition of an instrumental variable.

Definition 4.4 (Graphical criterion for instrumental variable Pearl (2009))**.** *A variable Z is an instrument relative to the total effect of X on Y if (i) $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$ and (ii) $(Z \not\perp\!\!\!\perp_d X)_{\mathcal{G}}$. Here $\mathcal{G}_{\bar{X}}$ refers to the causal graph $\mathcal{G}$ where all incoming edges into X have been removed.*

We note that condition (ii) in the above definition is already a testable independence constraint. Thus, we put our attention on whether we can test condition (i) – which does not have any observable independence constraints in $\mathcal{G}$ – using the "twinning" technique. We start with observing a problematic special case.

Theorem 4.3. *Consider assumption 4.1, 4.2 and 4.3 with $Y \notin Ancestors(X)$, $X \notin Ancestors(Z)$, and no selection bias, let $\mathcal{G}$ be the shared causal DAG across environments where the (testable) condition $(Z \not\perp\!\!\!\perp_d Y)$ holds and $\mathcal{G}^*$ its corresponding hierarchical DAG. Then, there exist no independence constraints in $\mathcal{G}^*$ for whether the edge $Z \to Y$ is present or not.*

We provide proof in the Appendix. The consequence of this theorem is that without further assumptions, we can not find an independence constraint in $\mathcal{G}^*$ that implies $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$. The reason is that the presence of the edge $Z \to Y$ implies that $(Z \not\perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$. That Z may not have a direct effect on Y is also referred to as the exclusion restriction for instrumental variables (Angrist *et al.*, 1996). The result itself might not come as a surprise, as the impossibility result we proved for the front-door criterion in Theorem 4.2 also relates to the presence of such direct edges. Despite this, we can still in some scenarios falsify if $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$ is true.

Theorem 4.4. *Consider the same assumptions as in Theorem 4.3, let $\mathcal{G}$ be the shared causal graph across environments and $\mathcal{G}^*$ its corresponding hierarchical DAG. Then, for any k and $i \neq j$, we have $Z_i^{(k)} \not\perp\!\!\!\perp_{P^*} Y_j^{(k)} \mid Z_j^{(k)} \Rightarrow (Z \not\perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$.*

The theorem presents an approach to falsify the validity of an instrument. As shown in the proof of the theorem, which is found in the Appendix, falsification is possible

when the unobserved confounder U is a cause of Z. In literature, this relates to the necessary condition that Z must be independent of any exogenous variable between X and Y (Angrist *et al.*, 1996). We note however that falsification is not possible if it is the other way around, i.e. $Z \to U$. This means that if one would conclude that $Z_i^{(k)} \perp\!\!\!\perp_{P^*} Y_j^{(k)} \mid Z_j^{(k)}$, one still needs to think carefully about the assumptions that have been made.

4.5. Discussion

In this paper, we have studied a new type of hierarchical causal model for data from multiple environments and its use in deriving testable implications of violations of common identification strategies. We learned that there exist independence constraints in this new class of DAGs that can be used to falsify (parts of) three common identification strategies in causal inference: the back-door, front-door, and instrumental variable criterion. If one of the testable conditions we have presented is violated, this could be informative to us that not all of our assumptions are valid for identification.

It is important to note that, although these hierarchical models expand the possibilities of testing assumptions, they are not a silver bullet. Firstly, our theory relies on a new untestable assumption: the independent causal mechanisms varying across environments. This assumption should not be taken for granted, yet a more conservative interpretation of the tests presented in this paper would be that they are a joint test to detect either a violation in the identification assumptions or that the mechanisms are dependent. Secondly, we demonstrated some limits of using the "twinning" technique with the hierarchical models. In particular, we learned that we can not test for the presence of a direct edge in the front-door and the instrumental variable setting. Still, we believe that showing we can test parts of these conditions constitutes important progress in the falsification of causal assumptions.

The hierarchical causal graphical model was a useful model in this setting that may be insightful in other causal inference settings as well. Interesting directions in this regard are investigating other identification strategies or combining this model with traditional independence-based causal discovery.

Appendices

4.A. Practical testing of independence constraints

In this section, we outline the procedure for testing a conditional independence relationship like $X_i^{(k)} \perp\!\!\!\perp_{P^*} Y_j^{(k)} \mid X_j^{(k)}, Z_i^{(k)}, Z_j^{(k)}$ or those in eq. (4.4), utilizing multi-environment data. We denote this data with $\{x_i^{(k)}, y_i^{(k)}, z_i^{(k)}\}_{i=1}^{N_k}$ with $k = 1, \ldots, K$.

To test such independencies, we want to simulate sampling from the joint distribution $P^*(X_i^{(k)}, Y_i^{(k)}, Z_i^{(k)}, X_j^{(k)}, Y_j^{(k)}, Z_j^{(k)})$ for some $i \neq j$. It is worth noting here that we do not condition on the environment, because otherwise the sample pair (i, j) would always be independent. Here's the approach we follow:

1. We select two distinct observations, denoted as i and j, from all environments. This selection yields vectors of observed treatments $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(K)})$, outcomes $y_i = (y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(K)})$, and so on for the vectors for z_i, x_j, y_j and z_j.

2. Subsequently, we apply a suitable conditional independence testing method, using the data points in $(x_i, y_i, z_i, x_j, y_j, z_j)$ as samples of each respective random variable.

It's important to note that the choice of observations within each environment is arbitrary, as long as we avoid selecting the same observation for both i and j. This flexibility arises from the assumption that observations are independent and identically distributed within each environment.

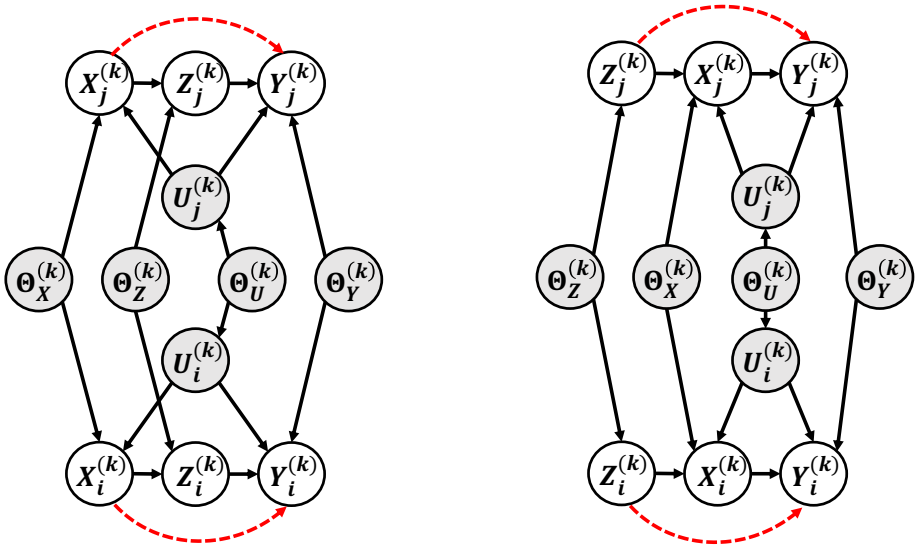
We see that, in principle, we only need two observations per environment to perform this independence test. The "sample size" of the test is the number of environments. However, it is possible to construct a procedure that uses all available data by combining multiple independence tests using Fisher's method, as long as we select different observations for each test (Karlsson & Krijthe, 2023).

4.B. Proofs

4.B.1. Proof of Theorem 4.2

Proof. We will show that there exists no independence constraint in the hierarchical graph $\mathcal{G}^*$, illustrated in Figure 4.3a, that is affected by the presence or absence of the red dashed edge. This edge corresponds to a violation of the fact that Z must intercept all directed paths between X and Y according to the front-door criterion (see Defintion 4.3). We construct our proof by showing that for the graph $\mathcal{G}^*$ in Figure 4.3a, the presence or absence of a red dashed arrow changes no independence constraint in $\mathcal{G}^*$.

First, we note that because of Assumption 4.2, any independence constraint in $\mathcal{G}^*$ holds



(a) Front-door setting

(b) Instrumental variable setting

Figure 4.3: Graphs to illustrate the claims by Theorem 4.2 and Theorem 4.3. We compare the independence constraints in the hierarchical DAG with either the red dashed edge present or absent; this corresponds to a violation of either the front-door or instrumental variable criterion respectively.

for all environments k. Secondly, because of Assumption 4.3, there exist open paths between the sample pair (i, j). Thirdly, Assumption 4.1 allows us to connect d-separation in G^* with the independence statement in the data distribution P^*. Finally, to show our claim, we only have to consider independence constraints between (i, j) samples, in contrast to for instance $(X_i^{(k)} \perp\!\!\!\perp_d Z_i^{(k)})_{\mathcal{G}^*}$, since $\mathcal{G}^*$ in this case otherwise does not provide anything extra compared to the corresponding non-hierarchical DAG $\mathcal{G}$.

We start by considering the independencies of the form $X_i^{(k)} \perp\!\!\!\perp_d Y_j^{(k)} \mid S$, where S is a set of the other observed variables. We note there always is a path between $X_i^{(k)}$ and $Y_j^{(k)}$ that traverses through $(U_i^{(k)}, U_j^{(k)})$, regardless of S. Thus, this type of independence does not change based on the presence of the edge $X \to Y$.

Next, we look at the independencies of the form $X_i^{(k)} \perp\!\!\!\perp_d Z_j^{(k)} \mid S$. We note that there is a path between $X_i^{(k)}$ and $Z_j^{(k)}$ through X_j that does not depend on the edge $X \to Y$, thus we should let $X_j^{(k)} \in S$ to block this path. The only way to now unblock the path between $X_i^{(k)}$ and $Z_j^{(k)}$ is to let $Y_j^{(k)} \in S$. However, this path does not go through the direct edge $X \to Y$ either. Thus, no independence of the form $X_i^{(k)} \perp\!\!\!\perp_d Z_j^{(k)} \mid S$ can detect the presence of this direct edge.

We look at the final form of independencies: that is $Y_i^{(k)} \perp\!\!\!\perp_d Z_j^{(k)} \mid S$. Using similar reasoning as above, it is clear that regardless of S, no path between $Y_i^{(k)}$ and $Z_j^{(k)}$ depends on the presence of the edge $X \to Y$.

As we have considered all possible types of independence constraints between observed variables, we see that no independence in $\mathcal{G}^*$ will change because of the presence of the direct edge $X \to Y$. This means that we can not test whether Z intercepts all directed paths between X and Y using this "twinning" technique. □

4.B.2. Proof of Theorem 4.3

Proof. We will show that there exists no observable independence constraint in the hierarchical graph $\mathcal{G}^*$, illustrated in Figure 4.3b, that depends on the presence of the red dashed edge $Z \to Y$. This edge corresponds to a violation of the fact that $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$ must hold for Z to be a valid instrument (see Definition 4.4). Here $\mathcal{G}_{\bar{X}}$ refers to the causal graph $\mathcal{G}$ where all incoming edges into X have been removed. We construct our proof by showing that for the graph $\mathcal{G}^*$ in Figure 4.3b, the presence or absence of a red dashed arrow changes no independence constraint in $\mathcal{G}^*$.

We use the same arguments as in the proof of Theorem 4.2 to conclude that we may look at independence constraint for any k and that there exist open paths between different pairs of samples (i, j). Once again, we will check all relevant independence constraints in the hierarchical DAG $\mathcal{G}^*$ and see if they would change if the red dashed edge is present or absent.

First, we look at the independencies of the form $X_j^{(k)} \perp\!\!\!\perp_d Y_i^{(k)} \mid S$ with S comprising the other observed variables. We note that this independence will always be violated, i.e. $X_j^{(k)} \not\perp\!\!\!\perp_d Y_i^{(k)} \mid S$ for any S. This is because we can always reach $(U_i^{(k)}, U_j^{(k)})$ without traversing $Z \to Y$.

Secondly, we look at the independencies of the form $X_i^{(k)} \perp\!\!\!\perp_d Z_j^{(k)} \mid S$. If $Z_i^{(k)} \notin S$, then we always have $X_i^{(k)} \not\perp\!\!\!\perp_d Z_j^{(k)} \mid S$, meaning we would need to only consider cases when $Z_i^{(k)} \in S$. In that case we have that $X_i^{(k)} \not\perp\!\!\!\perp_d Z_j^{(k)} \mid S$ holds whenever $X_j^{(k)}$ is also being conditioned on. Thus, it remains to check the cases where $X_j^{(k)} \notin S$, in which case we can see that conditioning on $Y_i^{(k)}$ and/or $Y_j^{(k)}$ does not change the conditional independence between $X_i^{(k)}$ and $Z_j^{(k)} \mid S$. Therefore, the validity of $X_i^{(k)} \perp\!\!\!\perp_d Z_j^{(k)} \mid S$ does not change based on the presence of the edge $Z \to Y$.

Finally, we look at $Y_i^{(k)} \perp\!\!\!\perp_d Z_j^{(k)} \mid S$. $Y_i^{(k)} \not\perp\!\!\!\perp_d Z_j^{(k)} \mid S$ will always hold if $Z_i^{(k)} \notin S$. Thus, we must consider cases when $Z_i^{(k)} \in S$, where we note that $Y_i^{(k)} \not\perp\!\!\!\perp_d Z_j^{(k)} \mid S$ if also $X_j^{(k)} \in S$ and/or $Y_j^{(k)} \in S$. This is because they both open a collider path through $X_j^{(k)}$. If $\{X_i^{(k)}, X_j^{(k)}\} \in S$, then the confounder association can be traversed. However, if $S = \{Z_i^{(k)}\}$

or $S = \{X_i^{(k)}, Z_i^{(k)}\}$, then $Y_i^{(k)}$ and $Z_j^{(k)}$ are independent. In none of these cases the edge $Z \to Y$ was used.

As we have considered all possible types of independence constraints between observed variables, we see that no independence in $\mathcal{G}^*$ will change because of the presence of the direct edge $Z \to Y$ in $\mathcal{G}$. □

4.B.3. Proof of Theorem 4.4

Proof. For this proof, we use a computational approach to iterate over different DAGs G while simultaneously searching for independence constraints in the corresponding hierarchical DAG G^* that can discriminate whether $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$ holds or not. Compared to the proofs of Theorem 4.2 and 4.3, where we only considered two graphs, we now must consider a much larger set of graphs to check whether an independence constraint gives the same value as $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$.

This approach consists of two parts: First, we iterate over a list of DAGs $\mathcal{G}$ that respect the following properties:

- U is always a confounder between X and Y, i.e. $U \to X$ and $U \to Y$ must be present;
- edges $X \to Y$, $Z \to Y$ and/or $Z \to X$ are present or absent (as we assume to know the causal ordering);
- and $U \to Z$ can either be present, absent or reversed;

This gives us a total of 24 graphs. In these graphs, the independence constraint $Z_j^{(k)} \perp\!\!\!\perp_d Y_i^{(k)} \mid Z_i^{(k)}$ often has the same values as $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$. As we illustrate in Table 4.1, we see that $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$ always holds if also $Z_j^{(k)} \perp\!\!\!\perp_d Y_i^{(k)} \mid Z_i^{(k)}$ is true, but not vice versa. We also see that $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$ is violated for graphs 0-11 as these have the direct edge between Z to Y, thus the most interesting cases are for graphs 12-23.

As $(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}} \Rightarrow (Z_j^{(k)} \perp\!\!\!\perp_d Y_i^{(k)} \mid Z_i^{(k)})_{\mathcal{G}^*}$, then must $(Z \not\perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}} \Leftarrow (Z_j^{(k)} \not\perp\!\!\!\perp_d Y_i^{(k)} \mid Z_i^{(k)})_{\mathcal{G}^*}$. Due to faithfulness (Assumption 4.1) we have that

$$(Z \not\perp\!\!\!\perp_P Y)_{\mathcal{G}_{\bar{X}}} \Leftarrow Z_j^{(k)} \not\perp\!\!\!\perp_{P^*} Y_i^{(k)} \mid Z_i^{(k)} .$$

□

Table 4.1: Each row corresponds to a different graph $\mathcal{G}$ considered in the proof for Theorem 4.4. The second column depicts the necessary conditions for the validity of Z being an instrument, while the third presents the independence constraint in $\mathcal{G}^*$ – a checkmark (✓) indicates that an independence hold. The remaining columns show the edges we change in the graphs.

Graph	$(Z \perp\!\!\!\perp_d Y)_{\mathcal{G}_{\bar{X}}}$	$Z_j \perp\!\!\!\perp_d Y_i \mid Z_i$	Z, X	Z, U	X, Y	Z, Y
0		✓	$Z \to X$	$Z \to U$	$X \to Y$	$Z \to Y$
1		✓	$Z \to X$		$X \to Y$	$Z \to Y$
2			$Z \to X$	$U \to Z$	$X \to Y$	$Z \to Y$
3		✓		$Z \to U$	$X \to Y$	$Z \to Y$
4		✓			$X \to Y$	$Z \to Y$
5				$U \to Z$	$X \to Y$	$Z \to Y$
6		✓	$Z \to X$	$Z \to U$		$Z \to Y$
7		✓	$Z \to X$			$Z \to Y$
8			$Z \to X$	$U \to Z$		$Z \to Y$
9		✓		$Z \to U$		$Z \to Y$
10		✓				$Z \to Y$
11				$U \to Z$		$Z \to Y$
12		✓	$Z \to X$	$Z \to U$	$X \to Y$	
13	✓	✓	$Z \to X$		$X \to Y$	
14			$Z \to X$	$U \to Z$	$X \to Y$	
15		✓		$Z \to U$	$X \to Y$	
16	✓	✓			$X \to Y$	
17				$U \to Z$	$X \to Y$	
18		✓	$Z \to X$	$Z \to U$		
19	✓	✓	$Z \to X$			
20			$Z \to X$	$U \to Z$		
21		✓		$Z \to U$		
22	✓	✓				
23				$U \to Z$		

Part Two

Trial Augmentation

5

Robust Integration of External Control Data in Randomized Trials

One approach for increasing the efficiency of randomized trials is the use of "external controls" – individuals who received the control treatment of the trial in routine practice or in prior experimental studies. Existing external control methods, however, can be biased if the populations underlying the trial and the external control data are not exchangeable. Here, we characterize a randomization-aware class of estimators for the treatment effect in the population underlying the trial that remain consistent and asymptotically normal when using external control data, even when exchangeability does not hold. We consider two members of this class of estimators: the well-known augmented inverse probability weighting trial-only estimator, which is the efficient estimator when only trial data are used; and a potentially more efficient member of the class when exchangeability holds and external control data are available, which we refer to as the optimized randomization-aware estimator. To achieve robust integration of external control data in trial analyses, we then propose a combined estimator based on the efficient trial-only estimator and the optimized randomization-aware estimator. We show that the combined estimator is consistent and no less efficient than the most efficient of the two component estimators, whether the exchangeability assumption holds or not. We examine the estimators' performance in simulations and we illustrate their use with data from two trials of paliperidone extended-release for schizophrenia.

This chapter appears as: Karlsson, R., Wang, G., De Bartolomeis, P., Krijthe, J. H., & Dahabreh, I. J. (2026). Robust integration of external control data in randomized trials [Forthcoming]. *Biometrics*. RK and GW contributed equally to this work.

5.1. Introduction

Randomized trials are the preferred approach for estimating treatment effects. However, trial conduct can be costly and time-consuming, and trials often have small sample sizes and imprecise results. One approach for improving the efficiency of trials involves augmenting them with data from *external* or *historical controls* (Jahanshahi *et al.*, 2021; Pocock, 1976) – individuals who received the control treatment as part of routine care or prior clinical investigations.

The challenges from small sample sizes in trials are particularly pronounced in studies of schizophrenia – a chronic, severe, and highly disabling condition, often leading to significant social and occupational impairment (National Institute of Mental Health, 2023). Difficulties with recruiting participants and monitoring their highly variable symptoms – even in clinical trials – result in sparse data in schizophrenia studies (Deckler *et al.*, 2022), leading to imprecise estimates of treatment effects. Incorporating external control data, as in the example presented in Section 5.7, is a promising approach for enhancing the efficiency of such clinical trials.

The task of augmenting trials with external controls is related to the task of transporting causal inferences from a trial to a target population (Dahabreh, Robertson *et al.*, 2020) because the former essentially "reverses the flow of information" compared with the latter (Ung *et al.*, 2024): instead of using information from the trial to learn about a target population, external control methods use information from an external population to improve inference in the trial. Consequently, trial analyses that use external control data often assume exchangeability conditions similar to those needed for transportability analyses (X. Li *et al.*, 2023; Schuler *et al.*, 2022; van Rosmalen *et al.*, 2018; van der Laan *et al.*, 2024; Ventz *et al.*, 2022; G. Wang *et al.*, 2024). However, when these exchangeability conditions do not hold, external control methods can introduce bias in the estimation of treatment effects. A natural, though imperfect, approach to address this challenge involves conducting a statistical test to assess whether the trial and external control populations are compatible for pooling (Yang *et al.*, 2023). Unfortunately, these tests have low statistical power, particularly when the trial sample size is small, which is precisely when using external control data would be most appealing. False negative results may result in substantial bias. The related approach of dynamically selecting valid external controls in a data-driven manner (Gao *et al.*, 2023; Viele *et al.*, 2014), is subject to the same risk of bias.

Here, we describe a class of "randomization-aware" estimators that can incorporate external control data and remain consistent and asymptotically normal, even when the external control population is not exchangeable with the trial population. We use optimization methods to identify a potentially more efficient member of this class when exchangeability holds; we refer to this member as the optimized randomization-aware estimator. Moreover, we propose a combined estimator that, asymptotically, is no less efficient than the most efficient between the trial-only estimator and the optimized randomization-aware estimator, whether the exchangeability assumption holds or not. In simulation studies, we verify that our estimators have good finite-sample performance

and are competitive with existing, less robust alternatives. Last, we apply the methods to data from two trials of paliperidone extended-release for schizophrenia.

5.2. Study design, data structure, and causal estimands

Study design and data structure: We assume that the trial and external control data are independently obtained simple random samples from different underlying populations, with unknown and possibly different sampling fractions. For the external data, this means that the sample is drawn randomly from an external population, such as the population underlying a registry or administrative database. We do not assume that individuals are randomly included in the registry itself. The trial and external control data are appended to form a composite dataset. In prior work, this sampling scheme has been referred to as a non-nested trial design because the proportions of trial participants and external controls in the composite dataset do not necessarily reflect the relative size of the underlying populations (Dahabreh *et al.*, 2021); see also Supplementary Material 5.B.

Simplifying assumptions: To focus on issues related to the integration of external controls, we make several simplifying assumptions: we consider only binary treatments and we assume complete adherence to treatment, no missing data, and no loss to follow-up. Standard methods for addressing these complications can be combined with the approaches we focus on.

Notation: We use italic capital letters for random variables and lowercase letters for specific values. We denote densities of random variables by $f(\cdot)$. Let $X \in \mathscr{X}$ denote baseline (pre-randomization and pre-treatment) covariates where $\mathscr{X}$ is the support over all possible covariate patterns, S the binary indicator for the study source ($S = 1$ for trial participants; $S = 0$ for the external controls), A the treatment strategy ($A = 1$ for the experimental treatment; $A = 0$ for the control treatment), and Y the binary, continuous, or count outcome measured at the end of the study.

Sampling model: We model the data on observation i with $S_i = s$ as independent and identically distributed, conditional on study source, realizations of the random tuple $O_i = (X_i, S_i = s, A_i, Y_i)$, for $i = 1, \ldots, n_s$, where n_s denotes the number of observations from source $S = s$. We define $n = n_1 + n_0$ as the sample size of the composite dataset. In the trial, treatment A is randomly assigned. In the population underlying the external control data, the only treatment in use may be the control treatment, in which case $\{S_i = 0\} \implies \{A_i = 0\}$, or treatment may be more variable, including the experimental and control treatments evaluated in the trial, as well as other treatments not examined in the trial. To simplify exposition, we mainly address the case of uniform use of the control treatment in the population underlying the external control data; we illustrate this data structure in Figure 5.1. Nevertheless, with small modifications, the methods we propose can also be applied when there exists variation in treatment in the population underlying the external data. Regardless of whether there is treatment variation, in many applied settings, the number of external controls, n_0, is much larger than the number of trial participants n_1. With increasing total sample size, we assume that the ratios of the sample

sizes of the trial and external control data over the total sample size converge to positive constants, that is, as $n \to \infty$, $n_s/n \to q_s > 0$.

Causal estimands: To define causal quantities, we use potential (counterfactual) outcomes (Robins & Greenland, 2000; Rubin, 1974). Specifically, for the ith individual and for $a \in \{0,1\}$, the potential outcome Y_i^a denotes the outcome under intervention to set treatment A to a, possibly contrary to fact. Our goal is to estimate the average treatment effect in the population underlying the trial, $\mathbb{E}[Y^1 - Y^0|S=1] = \mathbb{E}[Y^1|S=1] - \mathbb{E}[Y^0|S=1]$, and its constituent potential outcome means, $\mathbb{E}[Y^a|S=1]$, $a=0,1$.

5.3. Identification and estimation in the trial

5.3.1. Identification in the trial

Identifiability conditions: The following conditions suffice to identify potential outcome means and average treatment effect in the population underlying the trial:

Condition 5.1. *For every individual i and each treatment $a \in \{0,1\}$, if $A_i = a$, then $Y_i^a = Y_i$.*

Condition 5.2. *For each treatment $a \in \{0,1\}$, $Y^a \perp\!\!\!\perp A|(X, S=1)$.*

Condition 5.3. *For each treatment $a \in \{0,1\}$, if $f(x, S=1) \neq 0$, then* $\Pr[A=a|X=x, S=1] > 0$.

Condition 5.1 holds when the intervention is well-defined, such that there is no "hidden" or outcome-relevant treatment variation, and there is no interference. This condition is assumed on the basis of substantive knowledge, but aspects of experimental design (e.g., detailed treatment protocol) can increase plausibility. Implicit in our notation is an assumption that data source-specific effects (e.g., trial engagement effects (Dahabreh *et al.*, 2019)) are absent. Condition 5.2 is an assumption of no unmeasured confounding in the trial. This assumption is supported by study design in the context of a marginally or conditionally randomized trial. Condition 5.3, also supported by design, ensures that every covariate pattern in the trial population has a non-zero probability of receiving each treatment.

Identification: Under Conditions 5.1 to 5.3, the trial data alone can be used to identify the potential outcome mean under intervention to set treatment A to a in the trial population, $\mathbb{E}[Y^a|S=1]$, with $\psi_a \equiv \mathbb{E}\big[\mathbb{E}[Y|X,S=1,A=a]\big|S=1\big] = \dfrac{1}{\Pr[S=1]}\mathbb{E}\left[\dfrac{\mathbf{1}(S=1,A=a)Y}{\Pr[A=a|X,S=1]}\right]$. Furthermore, the average treatment effect in the population underlying the trial, $\mathbb{E}[Y^1 - Y^0|S=1]$, is identified with $\tau = \psi_1 - \psi_0$.

i	X_i	S_i	A_i	Y_i
1	x_1	1	1	y_1
2	x_2	1	0	y_2
⋮	⋮	⋮	⋮	⋮
n_1	x_{n_1}	1	1	y_{n_1}
n_1+1	x_{n_1+1}	0	0	y_{n_1+1}
n_1+2	x_{n_1+2}	0	0	y_{n_1+2}
⋮	⋮	⋮	⋮	⋮
n_1+n_0	$x_{n_1+n_0}$	0	0	$y_{n_1+n_0}$

Figure 5.1: The data contains information on baseline covariates X, an indicator for which population the observation belongs to (trial $\{S=1\}$ or external $\{S=0\}$ population), treatment A, and outcome Y. We have n_1 observations of the trial population and n_0 from the population underlying the external data, and $S=0$ implies $A=0$.

5.3.2. Estimation using trial data alone

To estimate ψ_a , we can use an outcome regression estimator $\left(\sum_{i=1}^n S_i\right)^{-1}\sum_{i=1}^n S_i\widehat{g}_a(X_i)$, where $\widehat{g}_a(X)$ is an estimator for $g_a(X)=\mathbb{E}[Y|X,S=1,A=a]$. When the model for $g_a(X_i)$ is correctly specified, the outcome regression estimator is consistent. However, correct specification of the outcome model is challenging. Alternatively, we can estimate ψ_a using the inverse probability weighting estimator $\left(\sum_{i=1}^n S_i\right)^{-1}\sum_{i=1}^n S_i\mathbf{1}(A_i=a)Y_i/e_a(X_i)$, where $e_a(X)=\Pr[A=a|X,S=1]$ is the conditional probability of treatment a in the trial (i.e., the propensity score (Rosenbaum & Rubin, 1983b)). This weighting estimator is consistent because the probability of treatment in the trial is known by design.

By combining outcome regression and weighting, we obtain the augmented inverse probability weighting estimator,

$$\widehat{\phi}_a=\left(\sum_{i=1}^n S_i\right)^{-1}\sum_{i=1}^n S_i\left[\frac{\mathbf{1}(A_i=a)}{e_a(X_i)}\left(Y_i-\widehat{g}_a(X_i)\right)+\widehat{g}_a(X_i)\right].$$

This estimator is asymptotically normal and *robust*, in the sense that it remains consistent even if the outcome regression model for $g_a(X)$ is misspecified, because the propensity score $e_a(X)$ is known. When the model for $g_a(X)$ is correctly specified, the estimator achieves the semiparametric variance bound using trial data alone (Robins & Rotnitzky, 1995; Robins *et al.*, 1994). A natural estimator for the average treatment effect in the trial τ is $\widehat{\tau}(\widehat{g})=\widehat{\phi}_1-\widehat{\phi}_0$; here, we index the estimator by $\widehat{g}$ to emphasize that it depends on $\widehat{g}_a(X)$, $a=0,1$. By construction, $\widehat{\tau}(\widehat{g})$ is consistent and asymptotically normal as well. We refer to $\widehat{\tau}(\widehat{g})$ as the *efficient trial-only estimator* of the treatment effect. In the next section, we consider strategies for incorporating external control data.

5.4. Using external controls under conditional exchangeability

We now present additional conditions that are often invoked when using external control data. In view of the results in the previous section, these additional conditions are not necessary for identification of the causal estimands; rather, they are invoked in the hopes of improving efficiency.

5.4.1. Identification using the external control data

In prior work (e.g., X. Li *et al.*, 2023; Valancius *et al.*, 2023), some version of the following two conditions has been invoked to allow the use of external control data:

Condition 5.4. $Y^0 \perp\!\!\!\perp S \mid X$.

Condition 5.5. $\{S = 0\} \implies \{A = 0\}$.

Alternatively, Condition 5.5 can be replaced with the following independence condition:

Condition 5.6. $Y^0 \perp\!\!\!\perp A | (X, S = 0)$.

Condition 5.4 is a condition of exchangeability in distribution that allows the external control data to contribute to the analysis of the trial. In our setting, it can be replaced by the weaker condition of exchangeability in expectation, $\mathbb{E}[Y^0 | X = x, S = 0] = \mathbb{E}[Y^0 | X = x, S = 1]$, for each x with $f(x, S = 1) \neq 0$. Condition 5.5 states that all individuals in the population underlying the external control data receive the control treatment in the trial. As noted, this condition can be replaced with Condition 5.6, a partial (only under $a = 0$) no-confounding assumption; this assumption may be plausible even in the presence of treatment variation in the population underlying the external data. We reiterate that Condition 5.4, and Condition 5.5 or Condition 5.6, are strong assumptions, not supported by study design, and of uncertain plausibility. However, when they hold, using external control data offers the potential for improved efficiency.

5.4.2. Estimation assuming exchangeability of populations

Estimators for integrating external control data under exchangeability of the trial and external population can be categorized into two types. The first category of methods, such as those in X. Li *et al.* (2023) and Valancius *et al.* (2023), fully pool trial and external control data under Conditions 5.4 and 5.5. While these methods offer efficiency gains when these conditions hold, they become biased if the conditions are violated. To address this, one can use the fact that Conditions 5.1 to 5.3, along with Conditions 5.4 and 5.5, impose

testable restrictions on the observed data (a case of overidentification). This allows for a test-then-pool approach – an example of pre-test estimation – that statistically tests the restrictions, and uses the trial-only estimator when the test rejects or a pooling estimator when the test does not reject. The performance of this approach depends on the statistical test used and, more importantly, the available sample size. A more detailed discussion on these methods is provided in Supplementary Material 5.C.

The second category of estimators borrows a subset of the external control data in a data-driven manner, attempting to approximate the exchangeability conditions by selecting external control data that is compatible with the trial (Gao *et al.*, 2023; Viele *et al.*, 2014). Borrowing external information in that way might reduce bias when exchangeability does not hold, but it still carries more risk than necessary: mistakenly pooling trial data with incompatible external control data can introduce bias that could have been avoided by limiting the analysis to the trial data.

5.5. A novel estimation approach using external controls

Our aim is to develop a *consistent* estimator that leverages the external control data to improve efficiency in the trial when exchangeability between sources holds, but does not rely on exchangeability assumptions for consistency. In outline, our strategy is as follows: First, we examine a class of randomization-aware estimators for τ that are consistent when Conditions 5.1 to 5.3 hold; the efficient trial-only estimator is a member of this class. Next, we identify a member within the class that has improved efficiency when Conditions 5.4 and 5.5 hold. Last, we introduce a combined estimator based on the efficient trial-only estimator and the optimized randomization-aware estimator, and show that it only requires Conditions 5.1 to 5.3 for consistency, and is no less efficient than the most efficient of the efficient trial-only estimator and the optimized randomization-aware estimator.

5.5.1. A class of consistent estimators

We consider the following class of estimators indexed by a function $h : \mathscr{X} \to \mathbb{R}$,

$$\widehat{\psi}_a(h) = \left(\sum_{i=1}^{n} S_i\right)^{-1} \sum_{i=1}^{n} S_i \left[\frac{\mathbf{1}(A_i = a)}{e_a(X_i)}\{Y_i - h(X_i)\} + h(X_i)\right],$$

which can be viewed as the solution to the estimating equation $\sum_{i=1}^{n} S_i\big[\frac{\mathbf{1}(A_i=a)}{e_a(X_i)}\{Y_i - h(X_i)\} + h(X_i) - \widehat{\psi}_a(h)\big] = 0$. Different choices of the function h correspond to different estimators $\widehat{\psi}_a(h)$, each with different properties. For instance, when $h \equiv 0$, the resulting estimator $\widehat{\psi}_a(0)$ is the inverse probability weighting estimator. Similarly, when $h \equiv \widehat{g}_a$, the resulting estimator $\widehat{\psi}_a(\widehat{g}_a)$ is the augmented inverse probability weighting estimator $\widehat{\phi}_a$.

We prove in Supplementary Material 5.A.1 that the estimator $\widehat{\psi}_a(h)$ is robust, that is, consistent regardless of the choice of h. We refer to this class of estimators as *randomization-*

aware because its members exploit randomization and knowledge of the probability of treatment to ensure consistency. Instances of specific randomization-aware estimators have appeared in previous work on using external data sources (e.g., Gagnon-Bartsch *et al.*, 2023).

To borrow information from the external control population, we attempt to "learn" $h(X)$ using both the trial and external control data. Throughout this paper, we assume that $h(X)$ satisfies certain regularity conditions; namely, that it is continuous, differentiable, smooth, and has finite expectation and variance at all points for $X = x$. Informally, the purpose of $h(X)$ in our estimators is to capture the relationship between the outcome and baseline covariates; flexible regression models such as splines and kernel smoothing methods can approximate the relationship and satisfy the regularity conditions.

Consider some fixed function h_{fix} that satisfies the above-mentioned regularity conditions, resulting in the estimator $\widehat{\psi}_a(h_{\text{fix}})$ that depends on the true propensity score and h_{fix}. We note the following properties of $\widehat{\psi}_a(h_{\text{fix}})$ (see Supplementary Material 5.A.2 for the proof).

Lemma 5.1. *Suppose Y has finite mean and variance, then the estimator $\widehat{\psi}_a(h_{fix})$ is consistent and asymptotically normal, with asymptotic variance* $\Pr[S=1]^{-1}[\mathscr{C}_a + L_a(h_{fix}(X), Y)]$, *where* $\mathscr{C}_a = Var[Y^a|S=1]$, *and*

$$L_a(h_{fix}(X), Y) = \mathbb{E}\left[\omega_a(X) \cdot \{Y - h_{fix}(X)\}^2 \middle| S=1, A=a\right] \tag{5.1}$$

with

$$\omega_a(X) = \Pr[A=a|S=1]\frac{1-e_a(X)}{e_a^2(X)}\,.$$

In other words, $\widehat{\psi}_a(h_{\text{fix}})$ is asymptotically normal with asymptotic variance whose only component that involves h_{fix} is $L_a(h_{\text{fix}}(X), Y)$. We will exploit Lemma 5.1 to develop a procedure for choosing h that attempts to improve the efficiency of our estimators.

5.5.2. An optimized randomization-aware estimator

Consider the class of randomization-aware estimators $\widehat{\psi}_a(h)$. The results in the previous sub-section suggest that we can "learn" an $h(X)$ function that results in a more efficient randomization-aware estimator. Specifically, we propose to use the trial and external control data together to find $h(X)$ that minimizes the asymptotic variance of $\widehat{\psi}_0(h_{\text{fix}})$, which is the same as minimizing the term $f_0(h_{\text{fix}}(X), Y)$ in eq. (5.1). Our approach is similar to the approach of Cao *et al.* (2009) for the problem of estimating an expectation of an outcome with missing data in a single data source setting.

In the results presented above, however, $f_0(h_{\text{fix}}(X), Y)$ is written as an expectation conditional on $S=1$. To incorporate external control data in the optimization, we use Conditions 5.4 and 5.5 to rewrite $f_0(h_{\text{fix}}(X), Y)$ as an expectation over both trial and external control data (see Supplementary Material 5.A.4 for proof).

Lemma 5.2. *Under Conditions 5.1 to 5.5, we have that*

$$L_0(h_{fix}(X), Y) = \mathbb{E}\left[\tilde{\omega}_0(X) \cdot \{Y - h_{fix}(X)\}^2 \middle| A = 0\right] \tag{5.2}$$

with

$$\tilde{\omega}_0(X) = \frac{\Pr[S = 1 \mid X, A = 0]}{\Pr[S = 1 \mid A = 0]} \frac{\Pr[A = 0|S = 1]e_1(X)}{e_0^2(X)} .$$

This lemma provides a way for choosing $h(X)$ using external control data to potentially improve the asymptotic variance of $\widehat{\psi}_0(h)$ when Conditions 5.4 and 5.5 hold. Ignoring normalizing constants, we define $h^*(X)$ as the minimizer of

$$R(\widetilde{h}) = \mathbb{E}\left[\eta_0(X)e_1(X)/e_0^2(X)\{Y - \widetilde{h}(X)\}^2 \middle| A = 0\right]$$

within a model class $\widetilde{h} \in \mathscr{H}$, where $\eta_0(X) = \Pr[S = 1 \mid X, A = 0]$ is the probability of participation in the trial, conditional on covariates, among individuals receiving treatment $A = 0$. We can estimate $h^*(X)$ by finding the minimizer of the sample analog of $R(\widetilde{h})$, replacing $\eta_0(X)$ with its estimator. Let $\mathscr{H}$ be a class of parametric models, and substitute $h(X;\widetilde{\gamma})$ for $\widetilde{h}(X)$, and $h(X;\gamma^*)$ for $h^*(X)$, where the additional notation is used to indicate dependence on the finite-dimensional vectors $\widetilde{\gamma}$ and γ^*, respectively. Estimating $h^*(X)$ is the same as estimating γ^*, which can be done by minimization: $\widehat{\gamma}^* = \arg\min_{\widetilde{\gamma}} \sum_{i=1}^n \widehat{R}(O_i;\widetilde{\gamma})$, where $\widehat{R}(O_i;\widetilde{\gamma}) = (1 - A_i)\widehat{\eta}_0(X_i)e_1(X_i)/e_0^2(X_i)\{Y_i - h(X_i;\widetilde{\gamma})\}^2$. We denote the resulting estimator as $\widehat{\psi}_0(\widehat{h}^*)$ and refer to it as the "optimized randomization-aware estimator."

5.5.3. Implementation using M-estimation methods

We use M-estimation (Stefanski & Boos, 2002) to implement the randomization-aware estimator $\widehat{\psi}_0(\widehat{h}^*)$ using $\widehat{h}^*(X)$ as the estimated optimized $h(X)$. For a set of smooth finite-dimensional target parameters $\boldsymbol{\theta}$, its M-estimator $\widehat{\boldsymbol{\theta}}$ is the solution to a stack of equation of the form $\sum_{i=1}^n \boldsymbol{m}(O_i;\widetilde{\boldsymbol{\theta}}) = \mathbf{0}$, where $\widetilde{\boldsymbol{\theta}}$ is the set of parameters with arbitrary values and $\boldsymbol{m}(O_i;\widetilde{\boldsymbol{\theta}})$ is the stack of estimating functions.

We consider parametric models for $\eta_0(X)$ and $h^*(X)$ denoted by $\eta_0(X;\beta)$ and $h(X;\gamma^*)$, respectively, and furthermore, we denote $q = \Pr[S = 1]$. The set of target parameters is $\boldsymbol{\theta} = \{q, \beta, \gamma^*, \psi_0\}$ and the stack of estimating functions becomes

$$\boldsymbol{m}(O_i;\widetilde{\boldsymbol{\theta}}) = \begin{pmatrix} m_q(O_i;\widetilde{q}) \\ m_{\eta_0}(O_i;\widetilde{\beta}) \\ m_{h^*}(O_i;\widetilde{\beta},\widetilde{\gamma}) \\ m_{\psi_0}(O_i;\widetilde{q},\widetilde{\gamma},\widetilde{\psi}_0) \end{pmatrix}. \tag{5.3}$$

To estimate q, we define $m_q(O_i;\widetilde{q}) = S_i - \widetilde{q}$. When $\eta_0(X)$ is estimated by logistic regression, $m_{\eta_0}(O_i;\widetilde{\beta})$ is the logistic regression score equation. For estimating γ^* and ψ_0, we define $m_{h^*}(O_i;\widetilde{\beta},\widetilde{\gamma}) = \frac{\partial}{\partial\widetilde{\gamma}}\widehat{R}(O_i;\widetilde{\beta},\widetilde{\gamma})$, which is equivalent to

$$(1 - A_i)\eta_0(X_i;\widetilde{\beta})\frac{e_1(X_i)}{\{1 - e_1(X_i)\}^2}\{Y_i - h(X_i;\widetilde{\gamma})\}\frac{\partial}{\partial\widetilde{\gamma}}h(X_i;\widetilde{\gamma}) ,$$

and define $m_{\psi_0}(O_i;\widetilde{q},\widetilde{\gamma},\widetilde{\psi}_0)$ as

$$\frac{S_i}{\widetilde{q}}\left[\frac{\mathbf{1}(A_i=0)}{1-e_1(X_i)}\{Y_i-h(X_i;\widetilde{\gamma})\}+h(X_i;\widetilde{\gamma})-\widetilde{\psi}_0\right].$$

We obtain a consistent estimator of $\widehat{\psi}_0(\widehat{h}^*)$ by jointly solving the stack of estimating functions, that is, letting $\widehat{\boldsymbol{\theta}}$ be the solution to $\sum_{i=1}^n \boldsymbol{m}(O_i;\widetilde{\boldsymbol{\theta}})=\mathbf{0}$. The following theorem summarizes the properties of $\widehat{\psi}_0(\widehat{h}^*)$ obtained from solving this optimization task (see Supplementary Material 5.A.5 for the proof).

Theorem 5.1. *Suppose $\boldsymbol{m}(O_i;\widetilde{\boldsymbol{\theta}})$ satisfies the regularity conditions in Supplementary Material 5.A.3; then, the M-estimator $\widehat{\psi}_0(\widehat{h}^*)$ is unbiased, consistent, and asymptotically normal. In addition, if Conditions 5.1 to 5.5 hold and the model for estimating $\eta_0(X)$ can be correctly specified, $\widehat{\psi}_0(\widehat{h}^*)$ minimizes the asymptotic variance of $\widehat{\psi}_0(h)$ over $\mathscr{H}$.*

Because $h^*(X)$ is not estimated using only trial data ($S=1$), Theorem 5.1 provides a practical solution for estimating ψ_0 when incorporating external control data. Furthermore, we can obtain an estimator for the average treatment effect in the trial population τ as

$$\widehat{\tau}(\widehat{h}^*)=\widehat{\psi}_1(\widehat{g}_1)-\widehat{\psi}_0(\widehat{h}^*)\,. \tag{5.4}$$

Because both of $\widehat{\psi}_1(\widehat{g}_1)$ and $\widehat{\psi}_0(\widehat{h}^*)$ are consistent estimators, $\widehat{\tau}(\widehat{h}^*)$ is also a consistent estimator. To construct asymptotically normal estimators, we again propose to obtain $\{\widehat{\psi}_1(\widehat{g}_1),\widehat{\psi}_0(\widehat{h}^*)\}$ via joint M-estimation (see details in Supplementary Material 5.A.6), so that $\widehat{\psi}_1(\widehat{g}_1)$ and $\widehat{\psi}_0(\widehat{h}^*)$ are asymptotically bivariate normal (Stefanski & Boos, 2002). Because a linear combination of two bivariate normally distributed random variables is normally distributed, $\widehat{\tau}(\widehat{h}^*)$ is asymptotically normal.

As a final remark, the consistency of $\widehat{\tau}(\widehat{h}^*)$ does not rely on Conditions 5.4 and 5.5; however, the efficiency improvement that this estimator hopes to offer depends on these conditions, as well as the specification for the model for $\Pr[S=1|X,A=0]$. If Conditions 5.4 and 5.5 do not hold, or if the model for $\Pr[S=1|X,A=0]$ is misspecified, then $\widehat{\tau}(\widehat{h}^*)$ may be less efficient than the efficient trial-only estimator, $\widehat{\tau}(\widehat{g})$. To further relax the dependence of the estimator's efficiency on these additional conditions that are not justified by randomization, in the next section we develop a new estimator that is asymptotically guaranteed to not perform worse than the efficient trial-only estimator.

5.5.4. Combined estimator

We have two different consistent estimators: the efficient trial-only estimator $\widehat{\tau}(\widehat{g})=\widehat{\phi}_1-\widehat{\phi}_0$; and the optimized randomization-aware estimator $\widehat{\tau}(\widehat{h}^*)=\widehat{\psi}_1(\widehat{g}_1)-\widehat{\psi}_0(\widehat{h}^*)$ that incorporates external control data. When Conditions 5.4 and 5.5 also hold, and necessary statistical models are correctly specified, $\widehat{\tau}(\widehat{h}^*)$ is expected to be more efficient (in finite-sample) than $\widehat{\tau}(\widehat{g})$ because it uses more observations to model the outcome conditional

on covariates, under the control treatment. However, when Conditions 5.4 and 5.5 do not hold, the relative efficiency of $\widehat{\tau}(\widehat{h}^*)$ and $\widehat{\tau}(\widehat{g})$ depends on many factors, including the sample sizes of the trial and external controls and the extent to which these conditions are violated. To avoid choosing between $\widehat{\tau}(\widehat{h}^*)$ and $\widehat{\tau}(\widehat{g})$, we consider combining these two estimators, in a way that may provide further efficiency gains (Graybill & Deal, 1959). Specifically, we propose the combined estimator

$$\widehat{\tau}(\lambda) = \lambda\widehat{\tau}(\widehat{h}^*) + (1-\lambda)\widehat{\tau}(\widehat{g}), \ \ \forall \lambda \in \mathbb{R}.$$

If $\lambda = 0$, then $\widehat{\tau}(\lambda)$ degenerates to the efficient trial-only estimator $\widehat{\tau}(\widehat{g})$; in all other cases, $\widehat{\tau}(\lambda)$ incorporates information from the external control data. We note some important properties of $\widehat{\tau}(\lambda)$ that hold for all $\lambda \in \mathbb{R}$ (see Supplementary Material 5.A.7 for the proof).

Lemma 5.3. *If* $\{\widehat{\tau}(\widehat{g}), \widehat{\tau}(\widehat{h}^*)\}$ *is obtained via joint M-estimation,* $\widehat{\tau}(\lambda)$ *is unbiased, consistent and asymptotically normal for all* $\lambda \in \mathbb{R}$.

With our combined estimator being consistent, we propose to select λ such that its asymptotic variance, denoted by v_λ^2, is minimized. This involves finding $\lambda^* = \arg\min_\lambda v_\lambda^2$. Denote the asymptotic variance of the estimators $\widehat{\tau}(\widehat{g})$ and $\widehat{\tau}(\widehat{h}^*)$ by v_g^2 and $v_{h^*}^2$, and the asymptotic covariance by v_{g,h^*}. By writing $v_\lambda^2 = \lambda^2 v_{h^*}^2 + (1-\lambda)^2 v_g^2 + 2\lambda(1-\lambda)v_{g,h^*}$, we see that v_λ^2 is a quadratic function of λ which has closed-form expressions for the minimizing λ^* and the corresponding variance $v_{\lambda^*}^2$,

$$\lambda^* = \frac{v_g^2 - v_{g,h^*}}{v_g^2 + v_{h^*}^2 - 2v_{g,h^*}} \quad \text{and} \quad v_{\lambda^*}^2 = \frac{v_g^2 v_{h^*}^2 - v_{g,h^*}^2}{v_g^2 + v_{h^*}^2 - 2v_{g,h^*}}.$$

In Supplementary Material 5.A.9, we show that λ^* is well-defined unless $\widehat{g}$ and $\widehat{h}^*$ converge to the same asymptotic limit. Because these models are likely to be misspecified in different ways, in what follows, we assume that $\widehat{g}$ and $\widehat{h}^*$ converge to different limits, and λ^* is well-defined. In practice, λ^* is unknown because the asymptotic variances and covariance of $\widehat{\tau}(\widehat{g})$ and $\widehat{\tau}(\widehat{h}^*)$ are unknown. We can estimate $(v_g^2, v_{h^*}^2, v_{g,h^*})$ by the sandwich estimators as usually done in M-estimation, denoted by $(\widehat{v}_g^2, \widehat{v}_{h^*}^2, \widehat{v}_{g,h^*})$. Then we can estimate λ^* and $v_{\lambda^*}^2$ using plug-in estimators, denoted by $\widehat{\lambda}^*$ and $\widehat{v^2}_{\widehat{\lambda}^*}$. Note that when estimating λ^*, if $\widehat{g}$ and $\widehat{h}^*$ are highly correlated – or, in the unlikely case that they converge to the same limit – the estimation can be stabilized by adding a small correction factor to both the numerator and denominator (see (Barnatchez *et al.*, 2025) for a similar idea in a different setting).

Because the sandwich estimators are consistent estimators of the respective asymptotic variances (Stefanski & Boos, 2002), $\widehat{\lambda}^*$ is a consistent estimator for λ^*. More importantly, we can show that the (feasible) estimator $\widehat{\tau}(\widehat{\lambda}^*)$ using the plugin estimator $\widehat{\lambda}^*$ converges asymptotically to the same distribution as the (infeasible) estimator $\widehat{\tau}(\lambda^*)$ that requires oracle knowledge of λ^* (see Supplementary Material 5.A.10 for the proof).

Theorem 5.2. *The estimator* $\widehat{\tau}(\widehat{\lambda}^*)$ *is asymptotically equivalent to* $\widehat{\tau}(\lambda^*)$*; that is,* $\sqrt{n}(\widehat{\tau}(\widehat{\lambda}^*)-\tau) \xrightarrow{d} N(0, v^2_{\lambda^*})$ *where* $\xrightarrow{d}$ *denotes convergence in distribution. In addition,* $\widehat{\tau}(\widehat{\lambda}^*)$ *is no less efficient than* $\widehat{\tau}(\widehat{g})$ *and* $\widehat{\tau}(\widehat{h}^*)$*; i.e.,* $\sigma^2_{\lambda^*} \leqslant min\{\sigma^2_g, \sigma^2_{h^*}\}$.

In sum, $\widehat{\tau}(\widehat{\lambda}^*)$ is robust in the sense that whether additional conditions hold or not, it is consistent for τ. Furthermore, its efficiency is no less than that of its component estimators.

5.6. Simulation studies

We evaluated our estimators' finite-sample performance against existing methods for integrating external control data in clinical trials via simulations. First, we considered a scenario favorable to all methods, where Conditions 5.1 to 5.3 and 5.5 hold, all parametric models are correctly specified, and there is no distribution shift in baseline characteristics. Second, we considered a more realistic scenario where Conditions 5.4 and 5.5 do not hold, parametric models are misspecified, and there is a distribution shift, such that $f(X \mid S = 1) \neq f(X \mid S = 0)$. Table 5.1 presents the results of a simulation study in which each experiment was repeated 5000 times. Detailed data-generating mechanism, simulation methods, additional results, and code are presented in Supplementary Material 5.D. In both scenarios, our estimators were more efficient than the trial-only estimator and remained nearly unbiased, while alternatives showed significant biases in the more realistic scenario. When comparing trial sizes of 50 or 200, the greatest variance reduction occurred with the smaller trial size; with larger trials, the variance improvements for all methods were less pronounced.

5.7. Augmenting a trial of treatments for schizophrenia using external controls

To illustrate the proposed methods, we used data from two independent placebo-controlled, double-blind trials (NCT00668837 and NCT00077714) that compared paliperidone ER tablets 6 mg versus placebo in patients with schizophrenia.[1] We designated one trial (Marder *et al.*, 2007) as the index trial (i.e., the trial that we aimed to augment using external data, denoted by $S = 1$), and used data from the placebo group in the second trial as external controls (denoted by $S = 0$) (Davidson *et al.*, 2007).

Positive and Negative Syndrome Scale (PANSS) total scores are used for rating the severity of schizophrenia symptoms. The outcome of interest in our analyses was the PANSS score at week 6 after randomization. We included patients assigned to either paliperidone ER or placebo and for whom PANSS scores were available at baseline and at week 6. From the

[1] This data can be obtained from the Yale University Open Data Access Project https://yoda.yale.edu/, subject to approval.

Table 5.1: A comparison of multiple estimators in different scenarios. We considered a small trial ($n_1 = 50$) and a large trial ($n_1 = 200$) where the number of external controls was kept fixed ($n_0 = 200$). We computed the mean absolute bias, variance and coverage rate from the estimated 95%-confidence intervals using 5000 repeated simulations.

	Scenario A						Scenario B					
	Small trial			Large trial			Small trial			Large trial		
Estimator	Bias	Var	Cov	Bias	Var	Cov	Bias	Var	Cov	Bias	Var	Cov
Unadjusted trial-only	0.02	0.92	0.97	0.01	0.23	0.97	0.00	0.91	0.97	0.00	0.23	0.97
AIPW	0.00	0.53	0.97	0.00	0.02	0.94	0.01	0.79	0.92	0.00	0.18	0.94
Optimized randomization-aware	0.00	0.29	0.96	0.00	0.02	0.94	0.01	0.73	0.93	0.00	0.18	0.94
Combined	0.00	0.31	0.95	0.00	0.02	0.94	0.02	0.75	0.92	0.00	0.18	0.94
Pooling	0.00	0.26	0.96	0.00	0.01	0.94	0.32	0.53	0.91	0.30	0.14	0.87
Test-then-Pool	0.00	0.28	0.96	0.00	0.02	0.94	0.27	0.64	0.89	0.28	0.16	0.87
Selective Borrowing	0.00	0.56	0.76	0.01	0.02	0.83	0.01	0.77	0.81	0.02	0.18	0.88
Dynamic Borrowing	0.02	0.71	0.88	0.01	0.18	0.90	0.36	0.74	0.86	0.25	0.21	0.85

index trial, we included 97 patients who were assigned to paliperidone ER tablets 6 mg, and 91 patients who were assigned to the placebo; the sample size of the external control data was 111. We used patient gender, age, race, and baseline PANSS score as covariates. Their summary statistics are given in Supplementary Material 5.E.1.

Table 5.2 summarizes results from the unadjusted trial-only, pooled (X. Li *et al.*, 2023), test-then-pool, selective borrowing (Gao *et al.*, 2023), and Bayesian dynamic borrowing (Viele *et al.*, 2014) estimators, alongside the novel estimators proposed in this paper. Results from the pooled and test-then-pool estimators were meaningfully different from the trial-only estimator, suggesting that the external data were not compatible with the index trial data. In contrast, the three proposed estimators produced results that were compatible with the trial-only estimator and more precise. The $\widehat{\lambda}^*$ for the combined estimator was 0.064, further suggesting that the external data were not compatible with the index trial data. The selective borrowing estimator produced results similar to our combined estimator and also compatible with the trial-only estimator; the Bayesian dynamic borrowing estimator produced results that were meaningfully different from the trial-only estimator.

We also repeated the analyses by randomly sampling with replacement a fraction ($\sim 75\%$, $\sim 50\%$, $\sim 25\%$) of the available control observations (see Supplementary Materials 5.E.2). The results showed that $\widehat{\tau}(\widehat{\lambda}^*)$ can generate reasonable estimates with smaller or similar standard errors (relative to its alternatives) as the control group sample size decreases.

5.8. Discussion

We proposed a novel approach for using external control data to improve inference in trials. Like earlier work, we show that efficiency gains are possible when the trial and external control populations are exchangeable. However, our optimized randomization-

Table 5.2: Estimates and standard errors of the different estimators.

	Unadjusted	Pooling	Test-pool	$\widehat{\tau}(\widehat{g})$	$\widehat{\tau}(\widehat{h}^*)$	$\widehat{\tau}(\widehat{\lambda}^*)$	Selective borrowing	Dynamic borrowing
Estimate	-7.965	-9.562	-9.562	-7.7726	-7.532	-7.714	-7.711	-10.038
Standard error	12.374	2.460	2.460	2.896	2.927	2.894	2.894	2.633

aware estimator explicitly avoids relying on these additional conditions for its consistency, and only uses them in an attempt to improve efficiency. Moreover, by combining the efficient trial-only estimator with our optimized randomization-aware estimator we provide a new estimator that is no less efficient than the most efficient of these two component estimators. The combined estimator may lead to further efficiency gains, but its main attraction is protection from performing worse than the efficient trial-only estimator in large samples. We acknowledge, however, that in more complex settings, potential efficiency gains may be offset by other challenges, and further investigation is warranted to assess the robustness and practical performance of the proposed method in such contexts.

Throughout, we used parametric M-estimation methods to jointly estimate all nuisance models and the trial-only and optimized randomization-aware estimators. This approach makes the logic of combining information transparent and ensures the joint normality of the two estimators under standard conditions. The majority of trial analyses use simple parametric models; therefore, our approach can be viewed as a relatively natural next step when trial data are to be combined with external control data. Extensions of our approach to use data-adaptive (e.g., machine learning) modeling strategies may further improve performance. One limitation of our approach may be unstable coverage when the sample size of the trial is very small. Nevertheless, even in this setting, our estimator had better coverage than common alternatives.

Appendices

5.A. Proofs

5.A.1. Proof of the robustness property of a special case of $\psi_a(h)$

Proof. We first observe that for any measurable function h, $\widehat{\psi}_a(h)$ converges in probability to the following quantity.

$$\begin{aligned}\widehat{\psi}_a(h) &= \left(\sum_{i=1}^{n} S_i\right)^{-1} \sum_{i=1}^{n} S_i \left[\frac{\mathbf{1}(A_i = a)}{e_a(X_i)}\{Y_i - h(X_i)\} + h(X_i)\right] \\ &\xrightarrow{p} \frac{1}{\Pr[S=1]}\mathbb{E}\left[S\frac{\mathbf{1}(A=a)}{e_a(X)}Y\right] + \frac{1}{\Pr[S=1]}\mathbb{E}\left[Sh(X)\left\{1 - \frac{\mathbf{1}(A=a)}{e_a(X)}\right\}\right].\end{aligned}$$

Further, the first term equals $\mathbb{E}[Y|A=a, S=1]$ and the second term is 0. Therefore, $\widehat{\psi}_a(h)$ is a consistent estimator for ψ_a regardless of the specification of h. □

5.A.2. Proof of Lemma 5.1

Proof. We have proved the estimator is consistent in 5.A.1; next, we prove it is asymptotically normal. Define $O_i = (Y_i, S_i, A_i, X_i)$, because $\{O_1, O_2, ..., O_n\}$ are independent and identically distributed random variables, by the Central limit theorem, when Y has finite mean and variance, the estimator is asymptotically normal. Now we derive its asymptotic variance.

Denote $\Pr[S=1]$ by q, $n^{-1}\sum_{i=1}^{n} I(S_i = 1)$ by $\widehat{q}$, and $\frac{A}{e_a(X)}(Y - h_{\text{fix}}(X)) + h_{\text{fix}}(X)$ by $T(X,Y)$.

We first prove that $\psi_a = \mathbb{E}[(S/q)T(X,Y)]$, which will be used in the later proofs.

We observe that

$$\begin{aligned}
\psi_a &= \mathbb{E}[\mathbb{E}[Y \mid X, A=a, S=1] \mid S=1] \\
&= \mathbb{E}\left[\frac{\Pr[A=a \mid X, S=1]\mathbb{E}[Y \mid X, A=a, S=1]}{e_a(X)} \middle| S=1 \right] \\
&= \mathbb{E}\left[\frac{E[AY \mid X, S=1]}{e_a(X)} \middle| S=1 \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{AY}{e_a(X)} \middle| X, S=1 \right] \middle| S=1 \right] \\
&= \mathbb{E}\left[\frac{AY}{e_a(X)} \middle| S=1 \right] \\
&= \mathbb{E}\left[\frac{AY}{e_a(X)} + \frac{e_a(X)-A}{e_a(X)} h_{\text{fix}}(X) \middle| S=1 \right] \\
&= \mathbb{E}\left[\frac{A}{e_a(X)} (Y - h_{\text{fix}}(X)) + h_{\text{fix}}(X) \middle| S=1 \right] \\
&= \mathbb{E}[T(X,Y) | S=1] \\
&= \mathbb{E}[(S/q)\, T(X,Y)].
\end{aligned}$$

The third to the fourth equation holds because

$$\mathbb{E}\left[\frac{e_a(X)-A}{e_a(X)} h_{\text{fix}}(X) \middle| S=1 \right] = \mathbb{E}\left[\mathbb{E}\left[\frac{e_a(X)-A}{e_a(X)} h_{\text{fix}}(X) \middle| X, S=1 \right] \middle| S=1 \right] = 0.$$

By definition of $\hat{\psi}_a(h_{\text{fix}}(X))$,

$$\mathbb{P}_n \left[\frac{S}{\hat{q}} \left[\frac{A}{e_a(X)} (Y - h_{\text{fix}}(X)) + h_{\text{fix}}(X) \right\} - \hat{\psi}_a(h_{\text{fix}}(X)) \right] = 0,$$

which implies

$$\sum_{i=1}^{n} \left[\frac{S_i}{\hat{q}} T(X,Y) - \hat{\psi}_a(h_{\text{fix}}(X)) \right] = 0.$$

By construction

$$n\{\hat{\psi}_a(h_{\text{fix}}(X)) - \psi_a\} = \sum_{i=1}^{n} \left[\frac{S_i}{\hat{q}} T(X,Y) - \psi_a \right].$$

By Taylor expansion,

$$
\begin{aligned}
& n\{\hat{\psi_a}(h_{\text{fix}}(X)) - \psi_a\} \\
\approx & \sum_{i=1}^{n} \left[\frac{S_i}{q} T(X,Y) - \psi_a \right] + \sum_{i=1}^{n} \frac{\partial}{\partial q} \left[\frac{S_i}{q} T(X,Y) - \psi_a \right] (\hat{q} - q) \\
= & \sum_{i=1}^{n} \left[\frac{S_i}{q} T(X,Y) - \psi_a \right] - \sum_{i=1}^{n} \frac{S_i T(X,Y)}{q^2} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(S_i - q \right) \\
\to & \sum_{i=1}^{n} \frac{S_i}{q} T(X,Y) - n\psi_a - \mathbb{E}\left[\frac{ST(X,Y)}{q^2} \right] \sum_{i=1}^{n} \left(S_i - q \right) \\
= & \sum_{i=1}^{n} \frac{S_i}{q} T(X,Y) - n\psi_a - \frac{1}{q}\psi_a \cdot \sum_{i=1}^{n} S_i + n\psi_a \\
= & \sum_{i=1}^{n} \left\{ \frac{S_i}{q} (T(X,Y) - \psi_a) \right\}.
\end{aligned}
$$

Therefore,

$$
\sqrt{n}\{\hat{\psi_a}(h_{\text{fix}}(X)) - \psi_a\} \to \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{S_i}{q} ((X,Y)T - \psi_a) \right],
$$

and

$$
\begin{aligned}
\text{AVar}(\hat{\psi_a}(h_{\text{fix}}(X))) &= \text{Var}\left[\frac{S}{q} (T(X,Y) - \psi_a) \right] \\
&= \mathbb{E}\left[\frac{S}{q^2} (T(X,Y) - \psi_a)^2 \right] - E^2 \left[\frac{S}{q} (T(X,Y) - \psi_a) \right] \\
&= \mathbb{E}\left[\frac{S}{q^2} (T(X,Y) - \psi_a)^2 \right] \\
&= \mathbb{E}\left[\frac{S}{q^2} T(X,Y)^2 \right] + \mathbb{E}\left[\frac{S}{q^2} \psi_a^2 \right] - 2\mathbb{E}\left[\frac{S}{q^2} T(X,Y)\psi_a \right] \\
&= \frac{1}{q} \mathbb{E}\left[T(X,Y)^2 \mid S = 1 \right] + \frac{\psi_a^2}{q} - 2\frac{\psi_a}{q} \cdot \psi_a \\
&= \frac{1}{q} \mathbb{E}\left[T(X,Y)^2 \mid S = 1 \right] - \frac{\psi_a^2}{q}.
\end{aligned}
$$

Now we take a closer look at $\mathbb{E}\left[T(X,Y)^2 \mid S=1\right]$.

$$
\begin{aligned}
&\mathbb{E}\left[T(X,Y)^2 \mid S=1\right]\\
=&\mathbb{E}\left[\left\{\frac{A}{e_a(X)}(Y-h_{\text{fix}}(X))+h_{\text{fix}}(X)\right\}^2 \,\middle|\, S=1\right]\\
=&\mathbb{E}\left[\frac{A}{e_a(X)^2}\left(Y^2+h_{\text{fix}}(X)^2-2Yh_{\text{fix}}(X)\right)+h_{\text{fix}}(X)^2+2\frac{A}{e_a(X)}(Y-h_{\text{fix}}(X))h_{\text{fix}}(X)\,\middle|\, S=1\right]\\
=&\mathbb{E}\left[A\cdot\frac{1-e_a(X)}{e_a(X)^2}\left(\frac{Y^2}{1-e_a(X)}+\frac{h_{\text{fix}}(X)^2}{1-e_a(X)}-2\frac{Yh_{\text{fix}}(X)}{1-e_a(X)}\right)+A\cdot\frac{1-e_a(X)}{e_a(X)^2}\cdot\frac{e_a(X)^2}{A(1-e)}h_{\text{fix}}(X)^2+\right.\\
&\left.2A\cdot\frac{1-e_a(X)}{e_a(X)^2}\frac{e_a(X)^2}{A(1-e_a(X))}\frac{A}{e_a(X)}(Y-h_{\text{fix}}(X))h_{\text{fix}}(X)\,\middle|\, S=1\right]\\
=&\mathbb{E}\left[A\cdot\frac{1-e_a(X)}{e_a(X)^2}\left\{\frac{Y^2}{1-e_a(X)}+\frac{h_{\text{fix}}(X)^2}{1-e_a(X)}-\frac{2}{1-e_a(X)}Yh_{\text{fix}}(X)+\frac{e_a(X)^2}{A(1-e_a(X))}h_{\text{fix}}(X)^2+\right.\right.\\
&\left.\left.\frac{2e_a(X)}{1-e_a(X)}Yh_{\text{fix}}(X)-\frac{2e_a(X)}{1-e_a(X)}h_{\text{fix}}(X)\right\}\,\middle|\, S=1\right]\\
=&\mathbb{E}\left[A\cdot\frac{1-e_a(X)}{e_a(X)^2}\left\{\frac{1}{1-e_a(X)}Y^2+\frac{(A-e_a(X))^2}{A(1-e_a(X))}h_{\text{fix}}(X)^2-Yh_{\text{fix}}(X)\right\}\,\middle|\, S=1\right]\\
=&\mathbb{E}\left[A\cdot\frac{1-e_a(X)}{e_a(X)^2}\left\{(Y-h_{\text{fix}}(X))^2+\frac{e_a(X)}{1-e_a(X)}Y^2+\frac{e_a(X)^2-Ae_a(X)}{A(1-e_a(X))}h_{\text{fix}}(X)^2\,\middle|\, S=1\right]\right.\\
=&\mathbb{E}\left[A\cdot\frac{1-e_a(X)}{e_a(X)^2}(Y-h_{\text{fix}}(X))^2\,\middle|\, S=1\right]+\mathbb{E}\left[\frac{A}{e_a(X)}Y^2\,\middle|\, S=1\right]+\\
&\mathbb{E}\left[\left(1-\frac{A}{e_a(X)}\right)h_{\text{fix}}(X)^2\,\middle|\, S=1\right]\\
=&\mathbb{E}\left[A\cdot\frac{1-e_a(X)}{e_a(X)^2}(Y-h_{\text{fix}}(X))^2\,\middle|\, S=1\right]+\mathbb{E}\left[\mathbb{E}\left[\frac{A}{e_a(X)}Y^2\,\middle|\, X,Y^a\cdot S=1\right]\,\middle|\, S=1\right]+\\
&\mathbb{E}\left[\mathbb{E}\left[\left(1-\frac{A}{e_a(X)}\right)h_{\text{fix}}(X)^2\,\middle|\, X,S=1\right]\,\middle|\, S=1\right]\\
=&\mathbb{E}\left[\Pr[A=a\mid S=1]\frac{1-e_a(X)}{e_a(X)^2}(Y-h_{\text{fix}}(X))^2\,\middle|\, S=1\right]+\\
&\mathbb{E}\left[\mathbb{E}\left[\frac{A}{e_a(X)}\left(Y^a\right)^2\,\middle|\, X,Y^a,S=1\right]\,\middle|\, S=1\right]\\
=&\mathbb{E}\left[\Pr[A=a\mid S=1]\frac{1-e_a(X)}{e_a(X)^2}(Y-h_{\text{fix}}(X))^2\,\middle|\, S=1\right]+\mathbb{E}\left[\left(Y^a\right)^2\mid S=1\right].
\end{aligned}
$$

The last term equals

$$\mathbb{E}\left[\left(Y^a\right)^2\mid S=1\right]=\text{Var}\left[Y^a\mid S=1\right]+\mathbb{E}^2\left[Y^a\mid S=1\right]=\text{Var}\left[Y^a\mid S=1\right]+\psi_a^2.$$

Therefore, $\text{AVar}(\hat{\psi}_a(h_{\text{fix}}(X)))=\frac{1}{q}\{\mathscr{C}_a+L_a(h_{\text{fix}}(X),Y)\}$, where $\mathscr{C}_a=\text{Var}\left[Y^a\mid S=1\right]$, and

$$L_a(h_{\text{fix}}(X),Y)=\mathbb{E}\left[\Pr[A=a\mid S=1]\frac{1-e_a(X)}{e_a(X)^2}(Y-h_{\text{fix}}(X))^2\,\middle|\, S=1\right].$$

This completes the proof.

We also note $\mathrm{AVar}(\hat{\psi}_a(h_{\mathrm{fix}}(X)))$ can be decomposed in other ways. For example,

$$
\begin{aligned}
&\mathbb{E}\left[T(X,Y)^2 \mid S=1\right] \\
=&\mathbb{E}\left[\left\{\frac{A}{e_a(X)}(Y-h_{\mathrm{fix}}(X))+h_{\mathrm{fix}}(X)\right\}^2 \middle| S=1\right] \\
=&\mathbb{E}\left[\frac{A^2}{e_a(X)^2}\left(Y^2+h_{\mathrm{fix}}(X)^2-2Yh_{\mathrm{fix}}(X)\right)+h_{\mathrm{fix}}(X)^2+2\frac{A}{e_a(X)}(Y-h_{\mathrm{fix}}(X))h_{\mathrm{fix}}(X)\middle| S=1\right] \\
=&\mathbb{E}\Bigg[\frac{1}{e_a(X)^2}\{AY^2+Ah_{\mathrm{fix}}(X)^2-2AYh_{\mathrm{fix}}(X)+e_a(X)^2h_{\mathrm{fix}}(X)^2+ \\
&2Ae_a(X)(Y-h_{\mathrm{fix}}(X))h_{\mathrm{fix}}(X)\}\middle| S=1\Bigg]. \\
=&\mathbb{E}\left[\frac{1}{e_a(X)^2}\left\{A^2Y^2+(A-e_a(X))^2h_{\mathrm{fix}}(X)^2-2A(1-e_a(X))Yh_{\mathrm{fix}}(X)\right\}\middle| S=1\right]. \\
=&\mathbb{E}\left[\frac{1}{e_a(X)^2}\{AY-(A-e_a(X))h_{\mathrm{fix}}(X)\}^2\middle| S=1\right]
\end{aligned}
$$

Therefore, $\mathrm{AVar}(\hat{\psi}_a(h_{\mathrm{fix}}(X))) = \frac{1}{q}\{\mathscr{C}'_a + L'_a(h_{\mathrm{fix}}(X),Y)\}$, where $\mathscr{C}'_a = -\psi_a^2$, and

$$
L'_a(h_{\mathrm{fix}}(X),Y) = \mathbb{E}\left[\frac{1}{e_a(X)^2}\{AY-(A-e_a(X))h_{\mathrm{fix}}(X)\}^2\middle| S=1\right].
$$

We opted for the $L_a(h_{\mathrm{fix}}(X),Y)$ because it is easier to be optimized. □

5.A.3. Regularity conditions for M-estimation

Denote the stack of estimating equations, and their derivative by $\boldsymbol{m}(O_i;\widetilde{\boldsymbol{\theta}})$ and $\boldsymbol{m'}(O_i;\widetilde{\boldsymbol{\theta}}) = \partial/\partial\widetilde{\boldsymbol{\theta}}\boldsymbol{m}(O_i;\widetilde{\boldsymbol{\theta}})$, respectively. Denote the set of possible parameter values of $\boldsymbol{\theta}$ by $\boldsymbol{\Theta}$. We list the regularity conditions for the M-estimators to be consistent and asymptotically normal (Newey & McFadden, 1994).

Condition 5.A.3.1. *Suppose $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ and let $\boldsymbol{m}_1(O_i;\boldsymbol{\theta}_1)$ be the estimating equation for $\boldsymbol{\theta}_1$. $1/n\sum_{i=1}^n \boldsymbol{m}_1(O_i;\boldsymbol{\theta}_1)$ and $1/n\sum_{i=1}^n \boldsymbol{m}(O_i;\widehat{\boldsymbol{\theta}}_1,\boldsymbol{\theta}_2)$ converge to 0 in probability for any partition of $\boldsymbol{\theta}$.*

Condition 5.A.3.2. *$\boldsymbol{\theta}$ is in the interior of $\boldsymbol{\Theta}$.*

Condition 5.A.3.3. *$\mathbb{E}[\boldsymbol{m}(O_i;\boldsymbol{\theta})]$ is continuous and $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}||1/n\sum_{i=1}^n \boldsymbol{m}(O_i;\boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{m}(O_i;\boldsymbol{\theta})]||$ converges to 0 in probability.*

Condition 5.A.3.4. *$\mathbb{E}[\boldsymbol{m}(O_i;\boldsymbol{\theta})] = 0$ and $\mathbb{E}[\boldsymbol{m}^2(O_i;\boldsymbol{\theta})] < \infty$.*

Condition 5.A.3.5. $\mathbb{E}[sup_{\boldsymbol{\theta}}||\boldsymbol{m'}(O_i;\boldsymbol{\theta})||] < \infty$.

Condition 5.A.3.6. *There is a neighborhood of* $\boldsymbol{\theta}$ *on which with probability one* $-\boldsymbol{m}(O_i;\widetilde{\boldsymbol{\theta}})$ *is continuously differentiable; and* $-\boldsymbol{m'}(O_i;\widetilde{\boldsymbol{\theta}})$ *converges uniformly to a non-stochastic limit which is non-singular at* $\widetilde{\boldsymbol{\theta}}$.

Condition 5.A.3.7. $B(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{m}(O_i;\boldsymbol{\theta})\boldsymbol{m}(O_i;\boldsymbol{\theta}^\top)]$ *is non-singular and* $\sqrt{n}\boldsymbol{m}(O_i;\boldsymbol{\theta})$ *converges in distribution to* $\mathcal{N}(0, B(\boldsymbol{\theta}))$.

5.A.4. Proof of Lemma 5.2

Before we present the proof of Lemma 5.2, we need to prove the following auxilliary lemma.

Lemma 5.A.4.1. *Under Conditions 5.1 to 5.5 we have that* $Y \perp\!\!\!\perp S \mid (X, A = 0)$.

Proof. First, due to the condition of no treatment variation in $\{S = 0\}$, we have $Y^a \perp\!\!\!\perp A \mid (X, S = 0)$ because A is a constant when $S = 0$ (alternatively, in case there is treatment variation in $\{S = 0\}$, we can replace Condition 5.5 by directly invoking $Y^a \perp\!\!\!\perp A \mid (X, S = 0)$ from Condition 5.6). Thus, combining this with that $Y^a \perp\!\!\!\perp A \mid (X, S = 1)$ (Condition 5.2), we have $Y^a \perp\!\!\!\perp A \mid (X, S)$. Combining this with Condition 5.4, $Y^a \perp\!\!\!\perp S \mid X$, we have $Y^a \perp\!\!\!\perp (A, S) \mid X$. This condition implies $Y^a \perp\!\!\!\perp S \mid (X, A)$, which follows from the weak union property of conditional independence. Finally, using Condition 5.1 (consistency), we have $Y \perp\!\!\!\perp S \mid (X, A = 0)$. □

Here follows the proof of Lemma 5.2.

Proof. We will express eq. (5.1) as a quantity that is not conditional on $\{S = 1\}$. We denote

$l(X,Y) = \dfrac{\Pr[A=0|S=1]e_1(X)}{e_0^2(X)}\{Y-h(X)\}^2$ and have

$$
\begin{aligned}
L_a(h(X),Y) &= \mathbb{E}\left[\, l(X,Y) \mid S=1, A=0\right] \\
&= \mathbb{E}\left[\frac{\mathbf{1}(S=1)}{\Pr[S=1 \mid A=0]}\, l(X,Y) \mid A=0\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}(S=1)}{\Pr[S=1 \mid A=0]}\, l(X,Y) \mid X, A=0\right] \mid A=0\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}(S=1)}{\Pr[S=1 \mid A=0]} \mid X, A=0\right] \mathbb{E}\left[\, l(X,Y) \mid X, A=0\right] \mid A=0\right] \\
&= \mathbb{E}\left[\frac{\Pr[S=1 \mid X, A=0]}{\Pr[S=1 \mid A=0]} \mathbb{E}\left[\, l(X,Y) \mid X, A=0\right] \mid A=0\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\Pr[S=1 \mid X, A=0]}{\Pr[S=1 \mid A=0]}\, l(X,Y) \mid X, A=0\right] \mid A=0\right] \\
&= \mathbb{E}\left[\frac{\Pr[S=1 \mid X, A=0]}{\Pr[S=1 \mid A=0]}\, l(X,Y) \mid A=0\right]
\end{aligned}
$$

where the fourth equation follows from that $Y \perp\!\!\!\perp S \mid (X, A=0)$ according to Lemma 5.A.4.1 and the fifth equation follows from that $\mathbb{E}[\mathbf{1}(S=1) \mid X, A=0] = \Pr[S=1 \mid X, A=0]$. □

5.A.5. Proof of Theorem 5.1

Proof. It is straightforward to verify that the estimating equations in eq. (5.3) satisfy regularity conditions (listed in Appendix 5.A.3), and thus $\widehat{\boldsymbol{\theta}}$ is asymptotically multivariate normal (Boos & Stefanski, 2013). The consistency and asymptotic normality of $\widehat{\psi}_0(\widehat{e}_0, \widehat{h}^*)$ are direct results from the M-estimation theories (Stefanski & Boos, 2002) and Lemmas 5.1 and 5.2. □

5.A.6. Obtain $\{\widehat{\psi}_0(\widehat{h}^*), \widehat{\psi}_1(\widehat{g}_1)\}$ via M-estimation

We already described how to obtain $\widehat{\psi}_0(\widehat{h}^*)$ by M-estimation. Here, we will describe how to obtain $\{\widehat{\psi}_0(\widehat{h}^*), \widehat{\psi}_1(\widehat{g}_1)\}$ by M-estimation. Compared to the estimation for ψ_0, estimating $\{\psi_0, \psi_1\}$ jointly requires two additional estimating equations, for estimating the parameters in $g_1(X)$, and for obtaining $\widehat{\psi}_0(\widehat{h}^*)$, respectively.

Denote $g_1(X)$ by $g_1(X;\zeta)$, and let $\boldsymbol{\theta'} = \{q, \beta, \gamma, \psi_0, \zeta, \psi_1\}$ be a vector of smooth finite dimensional target parameters. We propose to estimate $\boldsymbol{\theta'}$ by finding the $\widetilde{\boldsymbol{\theta}}'$ that solves the

following joint estimating equation

$$\begin{pmatrix} m_q(O_i;\widetilde{q}) \\ m_{\eta_0}(O_i;\widetilde{\beta}) \\ m_{h^*}(O_i;\widetilde{\beta},\widetilde{\gamma}) \\ m_{\psi_0}(O_i;\widetilde{q},,\widetilde{\gamma},\widetilde{\psi}_0) \\ m_{g_1}(O_i;\widetilde{\zeta}) \\ m_{\psi_1}(O_i;\widetilde{q},,\widetilde{\zeta},\widetilde{\psi}_1) \end{pmatrix} = \mathbf{0},$$

where the first four estimating equations are the same four estimating equations in eq. (5.3).

The estimating equation for $g_1(X;\widetilde{\zeta})$ depends on the outcome. For example, when the outcome Y is binary, $m_{g_1}(O_i;\widetilde{\zeta})$ is the score of the logistic regression models for $g_1(X;\zeta)$. To obtain $\widehat{\psi}_1(\widehat{g}_1)$, according to the construction of the estimator, we define

$$m_{\psi_1}(O_i;\widetilde{q},\widetilde{\zeta},\widetilde{\psi}_1) = \frac{1}{\widetilde{q}} S_i \left[\frac{\mathbf{1}(A_i=1)}{e_1(X_i)} \{Y_i - g_1(X_i,\widetilde{\zeta})\} + g_1(X_i,\widetilde{\zeta}) - \widetilde{\psi}_1 \right].$$

It is straightforward to verify that the above estimating equations satisfy regularity conditions (listed in Appendix 5.A.3).

5.A.7. Proof of Lemma 5.3

Proof. If $\widehat{\tau}(\widehat{g})$ and $\widehat{\tau}(\widehat{h}^*)$ are obtained via joint M-estimation (the details are given in Appendix 5.A.8) then they are are asymptotically bivariate normal (Boos & Stefanski, 2013). Because a linear combination of two bivariate normal distributed random variables is normally distributed, $\widehat{\tau}(\widehat{\lambda}^*)$ is asymptotically normal. As both $\widehat{\tau}(\widehat{g})$ and $\widehat{\tau}(\widehat{h}^*)$ are consistent estimators, it also follows that their linear combination is consistent. □

5.A.8. Obtain $\{\widehat{\tau}(\widehat{g}), \widehat{\tau}(\widehat{h}^*)\}$ via M-estimation

We already described the how to obtain $\{\widehat{\psi}_0(\widehat{h}^*), \widehat{\psi}_1(\widehat{g}_1)\}$ via M-estimation in Appendix 5.A.6. Here, we describe how to obtain $\{\widehat{\tau}(\widehat{g}), \widehat{\tau}(\widehat{h}^*)\}$ via M-estimation for a given $\widehat{\lambda}^*$. Observe that

$$\begin{aligned} \widehat{\tau}(\widehat{\lambda}^*) =& \widehat{\lambda}^* \widehat{\tau}(\widehat{g}) + (1-\widehat{\lambda}^*)\widehat{\tau}(\widehat{h}^*) \\ =& \widehat{\lambda}^* \{\widehat{\psi}_1(\widehat{g}_1) - \widehat{\psi}_0(\widehat{g}_0)\} + (1-\widehat{\lambda}^*)\{\widehat{\psi}_1(\widehat{g}_1) - \widehat{\psi}_0(\widehat{h}^*)\}. \end{aligned}$$

Compared to the M-estimation described in Appendix 5.A.6, obtaining $\{\widehat{\tau}(\widehat{g}), \widehat{\tau}(\widehat{h}^*)\}$ via M-estimation requires four additional estimating equations, for estimating the parameters in $\widehat{g}_0(X)$, and for obtaining $\widehat{\tau}(\widehat{h}^*)$, $\widehat{\psi}_0(\widehat{g}_0)$, and $\widehat{\tau}(\widehat{g})$, respectively.

Denote $g_0(X)$ by $g_0(X;\iota)$, and let $\boldsymbol{\theta''} = \{q,\beta,\gamma,\psi_0,\zeta,\psi_1,\tau(\widehat{h}^*),\iota,\tau(\widehat{g})\}$ be a vector of smooth finite dimensional targeted parameters parameters. We propose to estimate $\boldsymbol{\theta''}$ by finding

the $\widetilde{\theta}''$ that solves the following joint estimating equation

$$\begin{pmatrix} m_q(O_i;\widetilde{q}) \\ m_{\eta_0}(O_i;\widetilde{\beta}) \\ m_{h^*}(O_i;\widetilde{\beta},\widetilde{\gamma}) \\ m_{\psi_0}(O_i;\widetilde{q},\widetilde{\gamma},\widetilde{\psi}_0) \\ m_{g_1}(O_i;\widetilde{\zeta}) \\ m_{\psi_1}(O_i;\widetilde{q},\widetilde{\zeta},\widetilde{\psi}_1) \\ m_{\tau(h^*)}(O_i;\widetilde{\psi}_1,\widetilde{\psi}_0,\widetilde{\tau}(\widehat{h}^*)) \\ m_{g_0}(O_i;\widetilde{\iota}) \\ m_{\psi_0'}(O_i;\widetilde{q},\widetilde{\iota},\widetilde{\psi}_0') \\ m_{\tau(\widehat{g})}(O_i;\widetilde{\psi}_1,\widetilde{\psi}_0',\widetilde{\tau}(\widehat{g})) \end{pmatrix} = \mathbf{0},$$

where the first six estimating equations are the same six estimating equations in Appendix 5.A.6.

The estimating equation for $g_0(X;\iota)$ depends on the outcome. For example, when the outcome Y is binary $m_{g_0}(O_i;\widetilde{\iota})$ is the score of the logistic regression models for $g_0(X;\widehat{\iota})$. To obtain $\widehat{\psi}_0'$ (which is the estimate of ψ_0 using the trial-only estimator), $\widehat{\tau}(\widehat{g})$, and $\widehat{\tau}(\widehat{\lambda}^*)$, we define

$$\begin{aligned} &m_{\tau(h^*)}(O_i;\widetilde{\psi}_1,\widetilde{\psi}_0,\widetilde{\tau}(\widehat{h}^*)) = \widetilde{\psi}_1 - \widetilde{\psi}_0 - \widetilde{\tau}(\widehat{h}^*). \\ &m_{\psi_0'}(O_i;\widetilde{q},\widetilde{\iota},\widetilde{\psi}_0') = \frac{1}{\widetilde{q}} S_i \left[\frac{\mathbf{1}(A_i=0)}{1-e_1(X_i)} \{Y_i - \widetilde{g}_0(X_i,\widetilde{\iota})\} + \widetilde{g}_0(X_i,\widetilde{\iota}) - \widetilde{\psi}_0' \right] \\ &m_{\tau(\widehat{g})}(O_i;\widetilde{\psi}_1,\widetilde{\psi}_0',\widetilde{\tau}(\widehat{g})) = \widetilde{\psi}_1 - \widetilde{\psi}_0' - \widetilde{\tau}(\widehat{g}). \end{aligned}$$

It is straightforward to verify that all the above estimating functions satisfy regularity conditions (listed in Appendix 5.A.3).

5.A.9. Conditions for λ^* to be well-defined

When $\sigma_g^2 + \sigma_{h^*}^2 - \sigma_{g,h^*} = 0$, λ^* is not well defined. We will prove here that this pathological case only happens when $\widehat{g}$ and $\widehat{h}^*$ converge to the same limit.

Given that the vector $\{\widehat{\tau}(\widehat{g}), \widehat{\tau}(\widehat{h}^*)\}$ is jointly asymptotic normal, we can write

$$\sigma_g^2 + \sigma_{h^*}^2 - \sigma_{g,h^*} = 0 \iff \text{AVar}[\widehat{\tau}(\widehat{g}) - \widehat{\tau}(\widehat{h}^*)] = \text{AVar}[\widehat{\tau}(g^\dagger) - \widehat{\tau}(h^\dagger)] = 0,$$

where $g^\dagger = \{g_0^\dagger, g_1^\dagger\}$ and $h^\dagger$ are the limits of the (possibly misspecified) outcome functions, that is for $a \in \{0,1\}$ it holds that $||\widehat{g}_a - g_a^\dagger||_2 = o_P(1)$ and $||\widehat{h}^* - h^\dagger||_2 = o_P(1)$.

Now, we rewrite the difference of these two estimators as follows:

$$\widehat{\tau}(g^\dagger) - \widehat{\tau}(h^\dagger) = \frac{1}{n}\sum_{i=1}^n \Gamma_i, \text{ where } \Gamma_i = \frac{S_i}{P(S=1)} \left[\left(\frac{1-A_i}{e_0(X_i)} - 1 \right) \{h^\dagger(X_i) - g_0^\dagger(X_i)\} \right].$$

With the shorthand above we can rewrite the asymptotic variance of the difference as

$$\text{AVar}[\widehat{\tau}(g^\dagger) - \widehat{\tau}(h^\dagger)] = \text{Var}[\Gamma_i] = \text{Var}\left[\frac{S}{P(S=1)}\left(\frac{1-A}{e_0(X)} - 1\right)\{h^\dagger(X) - g_0^\dagger(X)\}\right].$$

Finally, we can conclude that

$$\begin{aligned}
\sigma_g^2 + \sigma_{h^*}^2 - \sigma_{g,h^*} = 0 &\iff \text{AVar}[\widehat{\tau}(g^\dagger) - \widehat{\tau}(h^\dagger)] = 0 \\
&\iff \text{Var}\left[\frac{S}{P(S=1)}\left(\frac{1-A}{e_0(X)} - 1\right)\{h^\dagger(X) - g_0^\dagger(X)\}\right] = 0 \\
&\iff \frac{S}{P(S=1)}\left(\frac{1-A}{e_0(X)} - 1\right)\{h^\dagger(X) - g_0^\dagger(X)\} = 0, \text{ almost surely} \\
&\iff h^\dagger(X) = g_0^\dagger(X), \text{ almost surely.}
\end{aligned}$$

Hence, λ^* is ill-defined if and only if $\widehat{h}^*$ and $\widehat{g}$ converge to the same limit (up to differences on some measure zero sets).

5.A.10. Proof of Theorem 5.2

In what follows we denote convergence in distribution and probability by $\xrightarrow{d}$ and $\xrightarrow{p}$, respectively.

Proof. If $\widehat{\tau}(\widehat{g})$ and $\widehat{\tau}(\widehat{h}^*)$ are obtained via joint M-estimation (the details are given in Appendix 5.A.8) then they are are asymptotically bivariate normal. More formally, we can write

$$\sqrt{n}\begin{pmatrix}\widehat{\tau}(\widehat{h}^*) - \tau \\ \widehat{\tau}(\widehat{g}) - \tau\end{pmatrix} \xrightarrow{d} Z, \text{ with } Z \sim \mathcal{N}(\mathbf{0}, \Sigma) \text{ and } \Sigma = \begin{pmatrix}\sigma_{h^*}^2 & \sigma_{g,h^*} \\ \sigma_{g,h^*} & \sigma_g^2\end{pmatrix}.$$

Further, for any $\lambda \in \mathbb{R}$, the combined estimators is equal to

$$\widehat{\tau}(\lambda) = \lambda\widehat{\tau}(\widehat{h}^*) + (1-\lambda)\widehat{\tau}(\widehat{g}) = (\lambda, 1-\lambda)^\top \begin{pmatrix}\widehat{\tau}(\widehat{h}^*) \\ \widehat{\tau}(\widehat{g})\end{pmatrix}.$$

Since by assumption $\widehat{\Sigma} \xrightarrow{p} \Sigma$, it follows from the continuous mapping theorem that $\widehat{\lambda}^* \xrightarrow{p} \lambda^*$. As a consequence, we have from Slutsky's theorem that

$$\sqrt{n}(\widehat{\tau}(\widehat{\lambda}^*) - \tau) \xrightarrow{d} (\lambda^*, 1-\lambda^*)^\top Z.$$

Therefore, we have asymptotic normality of the combined estimator and moreover, the asymptotic variance is given by:

$$\begin{aligned}
\text{AVar}[\widehat{\tau}(\widehat{\lambda}^*)] = \text{AVar}[\widehat{\tau}(\lambda^*)] &= (\lambda^*, 1-\lambda^*)^\top \Sigma (\lambda^*, 1-\lambda^*) \\
&= (\lambda^*)^2\sigma_{h^*}^2 + (1-\lambda^*)^2\sigma_g^2 + 2\lambda^*(1-\lambda^*)\sigma_{g,h^*}.
\end{aligned}$$

□

5.B. Discussion on study designs

Settings with data from multiple sources can often be categorized as either *nested trial designs* or *non-nested trial designs* (Dahabreh, Robertson *et al.*, 2020; Dahabreh *et al.*, 2021; F. Li *et al.*, 2022). For nested trial designs, the target population is well-defined with the trial population nested inside of it; often, the target population corresponds to a census from which trial-eligible individuals are selected. Those in the census who are not selected to participate in the trial correspond to the external population. Meanwhile, in non-nested trial designs the datasets are obtained separately. In this case, we do not know how the datasets were sampled from the target population.

Following the framework in Dahabreh *et al.*, 2021, the sampling mechanisms in both the nested and non-nested trial design can be formalized by introducing an indicator variable O. The variable O indicates whether an individual from the underlying target population is in the observed data: $\{O = 1\}$ for sampled individuals and $\{O = 0\}$ for non-sampled individuals. Using the diagram in Figure 5.2, we can illustrate how individuals from the underlying target population are first sampled into the actual population before they are divided into two sub-populations; in this case, either the trial population $\{S = 1\}$ or the external population $\{S = 0\}$. Here, we assume observations are simple random samples from each respective sub-population, meaning that is the sampling probabilities are determined by $\Pr[O = 1 \mid S = 1]$ and $\Pr[O = 1 \mid S = 0]$. Without any loss of generalization, we shall assume that all individuals in the trial sub-population are observed, i.e. $\Pr[O = 1 \mid S = 1] = 1$. We let $\Pr[O = 1 \mid S = 0] = u$ for some constant $0 < u \leq 1$, for nested trial designs is u known but for non-nested designs is u unknown.

Due to the simple random sampling, the observation indicator O is independent of the other variables conditioned on the sub-population S; that is $O \perp\!\!\!\perp (X, A, Y^1, , Y^0) \mid S$. This means $\mathbb{E}[Y^1 - Y^0 \mid S = s] = \mathbb{E}[Y^1 - Y^0 \mid S = s, O = 1]$, which implies that the average treatment effect in both sub-populations is still identifiable from the observed individuals. However, we see that $\mathbb{E}[Y^1 - Y^0] \neq \mathbb{E}[Y^1 - Y^0 \mid O = 1]$ can happen. The average treatment population on the target population is thus not identifiable unless we have a nested trial design, in which case we can get around this because u is known (Dahabreh *et al.*, 2021). Still, $\mathbb{E}[Y^1 - Y^0 \mid O = 1]$ can be interpreted as the average treatment effect on a mixture of the trial and external population that excludes all unobserved sub-populations (X. Li *et al.*, 2023).

We denote n as the number of observed individuals $(S, X, A, Y, O = 1)$ and N as the number of total individuals in the actual population (OS, OX, OA, OY, O). For obvious reasons, N is unknown to us. We assume that ratio $n_s/n \to q_s > 0$, for $S = 0, 1$, as $n \to \infty$.

5.B.1. Identifiability of eq. (5.2) in non-nested study designs

To minimize eq. (5.2), we need to estimate the study participation model $\eta_0(X) = P[S = 1 \mid X, A = 0]$. While η_0 is identifiable from the observed data in a nested design, this is not necessarily the case in a non-nested design. Whereas one could believe this will cause

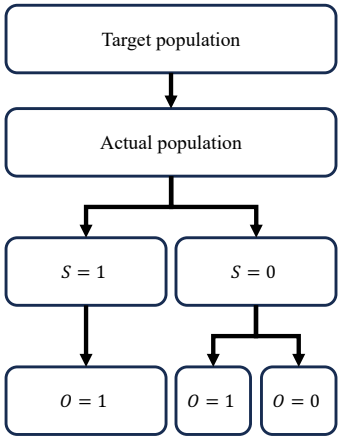


Figure 5.2: A diagram to conceptualize different study designs.

issues for minimizing eq. (5.2), we shall however show that this does not matter in the end and that eq. (5.2) still is identifiable from observations only; that is, even when everything is conditioned on $\{O=1\}$.

Letting $l(X,Y) = \dfrac{\Pr[A=0|S=1]e_1(X)}{e_0^2(X)}\{Y-h_{fix}(X)\}^2$, the claim of lemma 5.2 is that

$$\mathbb{E}[l(X,Y) \mid A=0,S=1] = \mathbb{E}\left[\frac{\Pr[S=1 \mid X,A=0]}{\Pr[S=1 \mid A=0]}\frac{\Pr[A=0|S=1]e_1(X)}{e_0^2(X)}\{Y-h_{\text{fix}}(X)\}^2 \middle| A=0\right].$$

However, due to simple random sampling of observations, we have

$$\mathbb{E}[l(X,Y) \mid A=0,S=1] = \mathbb{E}[l(X,Y) \mid A=0,S=1,O=1]\,.$$

Following the same proof as in the lemma, we can then show that

$$\begin{aligned}
&\mathbb{E}[\,l(X,Y) \mid S=1,A=0,O=1] = \\
&= \mathbb{E}\left[\frac{\mathbf{1}(S=1)}{\Pr[S=1 \mid A=0,O=1]}\,l(X,Y) \mid A=0,O=1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}(S=1)}{\Pr[S=1 \mid A=0,O=1]}\,l(X,Y) \mid X,A=0,O=1\right] \mid A=0,O=1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}(S=1)}{\Pr[S=1 \mid A=0,O=1]} \mid X,A=0,O=1\right]\mathbb{E}[\,l(X,Y) \mid X,A=0,O=1] \mid A=0,O=1\right] \\
&= \mathbb{E}\left[\frac{\Pr[S=1 \mid X,A=0,O=1]}{\Pr[S=1 \mid A=0,O=1]}\mathbb{E}[\,l(X,Y) \mid X,A=0,O=1] \mid A=0,O=1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\Pr[S=1 \mid X,A=0,O=1]}{\Pr[S=1 \mid A=0,O=1]}\,l(X,Y) \mid X,A=0,O=1\right] \mid A=0,O=1\right] \\
&= \mathbb{E}\left[\frac{\Pr[S=1 \mid X,A=0.O=1]}{\Pr[S=1 \mid A=0,O=1]}\,l(X,Y) \mid A=0,O=1\right]
\end{aligned}$$

where the fourth equation follows from that $Y \perp\!\!\!\perp S \mid (X, A = 0, O = 1)$. This follows from lemma 5.A.4.1 and that we have simple random sampling. It is easy to show that $O \perp\!\!\!\perp (X.A, Y^1, Y^0) \mid S \Rightarrow O \perp\!\!\!\perp Y \mid (S, X, A = 0)$ which combined with $Y \perp\!\!\!\perp S \mid (X, A = 0)$ implies that $Y \perp\!\!\!\perp S \mid (X, A = 0, O = 1)$. Thus, to conclude, we see in fact that $\mathbb{E}[l(X, Y) \mid A = 0, S = 1]$ can be expressed as an expectation over quantities of only observed data $\{O = 1\}$.

5.C. Discussion on using external control data under exchangeability of populations

In settings where data from multiple sources are combined, transportability conditions are often assumed to enable integration. Unlike our approach, existing methods typically construct estimators based on an identification strategy that depends on Conditions 5.4 and 5.5 (or Condition 5.6). While these additional conditions can be controversial, they are often justified alongside Conditions 5.1 to 5.3 through testable implications in the observed data. We discuss these issues below and further detail one estimator based on this strategy, which we use in our simulation study and application.

5.C.1. Identification assuming exchangeability of populations

Under Conditions 5.1 to 5.5, pooling of trial and external control data can be incorporated in analyses aiming to estimate the potential outcome mean in the population underlying the trial under intervention of the control treatment A to $a = 0$, that is, $\mathbb{E}[Y^0 \mid S = 1]$ (X. Li *et al.*, 2023; Valancius *et al.*, 2023). Specifically, we can identify $\mathbb{E}[Y^0 \mid S = 1]$ with

$$\zeta_0 = \mathbb{E}[\mathbb{E}[Y \mid X, A = 0] \mid S = 1].$$

Compared to the identification results using the trial data alone, we no longer condition on $S = 0$ in the inner expectation; instead, we pool the trial and external control data. The average treatment effect in the population underlying the trial, $\mathbb{E}[Y^1 - Y^0 \mid S = 1]$, is identified with $\psi_1 - \zeta_0$.

5.C.2. Testable implications of the additional identifiability conditions

Conditions 5.1 to 5.5 together have a testable implication in the law of observed data, namely that for each x with positive density in the population underlying the trial, $f(x, S = 1) \neq 0$,

$$H_0 : \mathbb{E}[Y \mid X = x, A = 0, S = 1] = \mathbb{E}[Y \mid X = x, A = 0, S = 0]. \tag{5.5}$$

To see this, we have from Condition 5.4 that

$$\mathbb{E}[Y^0 \mid X, S = 1] = \mathbb{E}[Y^0 \mid X, S = 0] \, .$$

Then, the above testable implication follows by noting that Conditions 5.1 to 5.3 allow us to re-write the left-hand as $\mathbb{E}[Y^0 \mid X, S=1] = \mathbb{E}[Y \mid X, A=0, S=1]$. Meanwhile, the right-hand side can be written as $\mathbb{E}[Y^0 \mid X, S=0] = \mathbb{E}[Y^0 \mid X, A=0, S=0] = \mathbb{E}[Y \mid X, A=0, S=0]$ where first equality follows from having no treatment variation in $\{S=0\}$ (Condition 5.5) and the second one from consistency (Condition 5.1).

The testable implication provides a way to evaluate whether Conditions 5.1 to 5.5 jointly hold; informally, it may be used to assess compatibility between the trial and external control data. Various methods exist for testing H_0, such as parametric likelihood-ratio tests or non-parametric alternatives (Luedtke *et al.*, 2019; Racine *et al.*, 2006). Complications related to doing a statistical test against H_0 and subsequently drawing statistical inferences using the same dataset can be addressed by sample-splitting or accounting for pre-testing when quantifying uncertainty (see, e.g., Rothenhäusler, 2024; Yang *et al.*, 2023).

5.C.3. Estimation under exchangeability of populations

We now present an existing estimator based on the identification strategy presented above, later we refer to it as the pooling estimator in our simulation study and application. X. Li *et al.*, 2023 proposed a doubly-robust estimator for ζ_0,

$$\widehat{\zeta}_0 = \left(\sum_{i=1}^{n} S_i\right)^{-1} \sum_{i=1}^{n} \left[\left(\frac{S_i(1-A_i) + (1-S_i)\widehat{r}(X_i)}{\widehat{\eta}(X_i)(1-\widehat{e}_0(X_i)) + (1-\widehat{\eta}(X_i))\widehat{r}(X_i)}\widehat{\eta}(X_i)\right)(Y_i - \widehat{g}_0(X_i)) + S_i\widehat{g}_0(X_i)\right],$$

where $\widehat{\eta}(X)$ is an estimator for the probability of participation in the trial $\Pr[S=1 \mid X]$, $\widehat{e}_0(X)$ is an estimator for the propensity score, and $\widehat{r}(X)$ is an estimator for the variance ratio $r(X) \equiv \mathrm{Var}[Y^0 \mid X, S=1]/\mathrm{Var}[Y^0 \mid X, S=0]$ comparing the trial population and the population underlying the external data. The estimator of the variance ratio $\widehat{r}(X)$ controls how much information to "borrow" from the external control data; this becomes evident by setting $r(X)=0$ in which case $\widehat{\zeta}_0 = \widehat{\phi}_0$.

Under Conditions 5.1 to 5.5, the estimator $\widehat{\zeta}_0$ is consistent if either the models for estimating $\eta(X)$ and $e_0(X)$ are correctly specified, or if the model for estimating $g_0(X)$ is correctly specified. Furthermore, if all working models are correctly specified, including for $r(X)$, $\widehat{\zeta}_0$ is the efficient estimator for the control outcome mean when trial and external control data are available (X. Li *et al.*, 2023). Also, under Conditions 5.4 and 5.5, one can replace $g_0(X)$ with an estimator for $\mathbb{E}[Y \mid X, A=0]$ using both the trial and external control data to borrow more information from the external population. However, if Conditions 5.4 and 5.5 do not hold (i.e. the external controls are not exchangeable) or the model for estimating $\eta(X)$ is misspecified, then $\widehat{\zeta}_0$ does not have these desirable properties whereas the trial-based estimator $\widehat{\phi}_0$ remains consistent and is the most efficient estimator that ignores the external control data.

5.D. Simulation studies

5.D.1. Data-generating process

For a given n_1 and n_0, we let $S_i = 1$ for $i = 1, \ldots, n_1$ and $S_i = 0$ for $i = n_1 + 1, \ldots, n_1 + n_0$. For observations with $S_i = 1$, we generated $A_i \sim \text{Bern}(1/2)$; for observations with $S_i = 0$, we let $A_i = 0$. The covariates X_i were sampled from a 10-dimensional multivariate Normal distribution $N(\mu_{S_i}, \Sigma)$ where μ_s depended on the scenario being considered; the diagonal elements of Σ were set to one and the off-diagonal elements were set to zero. We generated outcomes according to $Y_i = \sum_{j=1}^{5} \alpha_j X_{i,j} + \sum_{j=1}^{10} \beta_j X_{i,j}^2 + 5 \cdot A_i + \varepsilon_i$ with $\varepsilon_i \sim N(0,1)$, where we let $\alpha = (\frac{1}{2}, 1, -\frac{1}{2}, 1, -\frac{1}{2})$ and $\beta = (-\frac{1}{4}, -1, -\frac{1}{2}, -1, -\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$.

For scenario A, all parametric working models were correctly specified and we set $\mu_1 = \mu_0 = \mathbf{0}$ such that there was no distribution shift for the baseline covariates between the trial population and the population underlying the external control data. Meanwhile, for the more adversial scenario B, working models were misspecified by intentionally omitting the variables with $j = 5, \ldots, 10$ and all second-order terms, and we introduced distribution shift by setting $\mu_1 = \mathbf{0}$ and $\mu_0 = \frac{1}{2}\mathbf{1}$.

5.D.2. Additional results

To demonstrate how the relative sizes of the trial and external control groups influence the outcomes, we fixed the number of external controls at $n_0 = 200$ while varying the size of the trial group. Specifically, two settings were considered: (1) a small trial group with $n_1 = 50$ and (2) a large trial group with $n_1 = 200$.

We compare the following estimators: the unadjusted trial-only, the trial-only augmented inverse probability weighting (AIPW), an estimator that fully pools under exchangeability (X. Li *et al.*, 2023), test-then-pool, selective borrowing (Gao *et al.*, 2023), and Bayesian dynamic borrowing (Viele *et al.*, 2014), alongside the novel estimators proposed.

In scenario A for the small trial, all estimators had negligible bias and primarily differed in terms of variance. Among these, the pooling estimator achieved the lowest variance, followed by test-then-pool, optimized randomization-aware, combined, AIPW, selective borrowing, dynamic borrowing, and finally, the unadjusted trial-only estimator. Meanwhile, in scenario A for the large trial, the variance difference between AIPW (the best trial-only estimator) and estimators that incorporate external controls, excluding dynamic borrowing, became nearly negligible.

In scenario B, a significant increase in bias was observed for pooling, test-then-pool, and dynamic borrowing, regardless of trial size. This bias was slightly more pronounced in the small trial compared to the large trial. For the small trial, the variance ordering remained largely unchanged, except that dynamic borrowing performed comparably to optimized randomization-aware and combined estimators. The increase in AIPW variance in this scenario can be attributed to its sensitivity to misspecification, even though it does not

incorporate information from external controls.

Effect of removing weighting in the objective of optimized randomization-aware estimator In addition to the comparison between estimators, we investigated the importance of the weighting in the objective for $h^*(X)$ in the optimized randomization-aware estimator. This was done by repeating the same experiments as above but finding the minimizer that solves $\arg\min_{\tilde{h}} \mathbb{E}\left[\{Y - \tilde{h}(X)\}^2 \middle| A = 0\right]$. As shown in Table 5.4, the weighting had no impact in the scenario A. However, in the more adversarial scenario B, removing the weighting led to increased bias and variance. This outcome is unsurprising: weighting is necessary for improving inference on a targeted population, particularly when there is a distributional shift between populations and misspecification (see e.g. Shimodaira (2000)). The slight increase in bias, despite the robustness property of the randomization-aware estimators, could be due to the absence of weighting, which likely exacerbates the misspecification of the outcome model. This phenomenon has been observed to introduce non-negligible bias for doublY-robust estimators with a correctly specified propensity score in small sample sizes (Kang & Schafer, 2007).

Effect of replacing true propensity score with estimated propensity score We also explored a scenario where the true propensity score is replaced with an estimated propensity score in the optimized randomization-aware estimator. This investigation was motivated by the fact that, while the probability of treatment is known by design in a trial, estimating it is often preferred to improve efficiency (see, e.g., Williamson *et al.* (2014)). However, as shown in Table 5.4, using an estimated propensity score results in higher variance compared to using the true propensity score. Although this finding might seem to contradict recommendations in the literature, a plausible explanation is that estimating the propensity score introduces additional uncertainty, which destabilizes the minimization of the objective for $h^*(X)$.

Distribution of $\widehat{\lambda}$ in combined estimator To examine the behavior of the combined estimator, we analyzed the distribution of $\widehat{\lambda}$ in the small trial setting. Recall that when $\widehat{\lambda} = 0$ the combined estimator reduces to AIPW, and when $\widehat{\lambda} = 1$ it reduces to the optimized randomization-aware estimator. In scenario A, where the optimized randomization-aware estimator is expected to perform better, the distribution of $\widehat{\lambda}$ was centered closer to 1 (see Figure 5.3a). Conversely, in scenario B where the optimized randomization-aware is less favored, the distribution shifted closer to 0 (see Figure 5.3b). Interestingly, in both scenarios, $\widehat{\lambda}$ and in some instances, even outside this range. This suggests that the combined estimator frequently would leverage information from both estimators.

Computational costs with M-estimation We assessed the computational efficiency of our proposed methodology by measuring the execution time of the combined estimator as we increased either the sample size or the number of parameters to be estimated

while keeping all other factors constant. In both cases, we observed a linear increase in computational time. The results are shown in Figure 5.4.

5.D.3. Code

The code to reproduce the simulation studies is publicly available at https://github.com/RickardKarl/IntegratingExternalControls.

5.E. Data application details

5.E.1. Summary statistics of baseline covariates for different populations

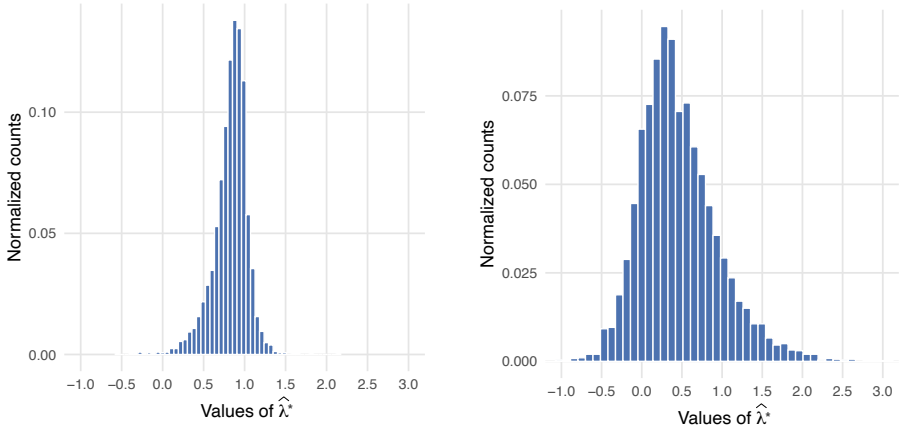
Table 5.3 summarises the baseline characteristics of the data we used from from the two independent, placebo-controlled, double-blind trials used in this paper.

Table 5.3: Baseline characteristics stratified by source trial and treatment

Characteristic	$S=1, A=1$ (n=97)	$S=1, A=0$ (n=91)	$S=0, A=0$ (n=111)
PANSS baseline score (mean (SD))	92.64 (11.76)	93.09 (11.53)	94.05 (12.65)
PANSS score at week 6 (mean (SD))	77.75 (20.34)	85.92 (23.12)	90.67 (24.95)
Age (mean (SD)	41.94 (10.49)	42.67 (10.85)	37.61 (10.92)
Gender (Female=1) (n (%))	31 (32.0)	21 (23.1)	35 (31.5)
Race (White=1) (n (%))	44 (45.4)	44 (48.4)	56 (50.5)

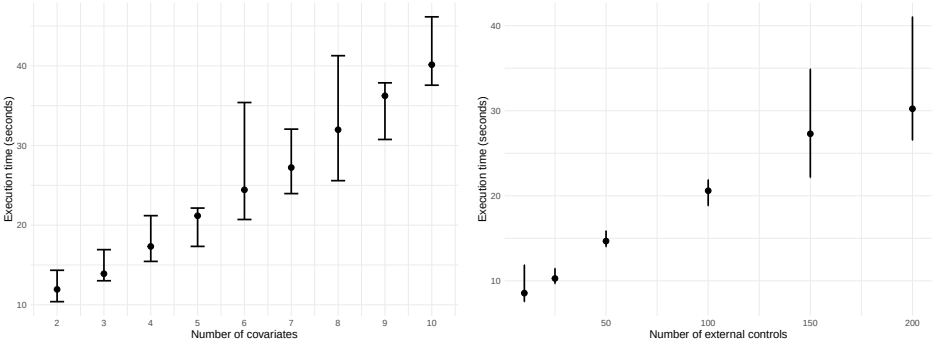
5.E.2. Additional analyses

Furthermore, to empirically examine the methods under different sample sizes of the index trial, we varied the sample size of the controls in the index trial by randomly sampling with replacement a fraction of the available observations. Specifically, of the 91 patients in the control group of the index trial; we sampled 68 ($\sim$ 75%), 46 ($\sim$ 50%), or 23 ($\sim$ 25%) individuals and repeated the analyses described above 100 times. The average of the results over the 100 analyses are shown in Table 5.5 (lines 2-4). With small sample sizes in the control group of the index trial (68, 46, and 23), the point estimates of the pooling estimator deviated from the points estimates of IPW and the AIPW estimator. This showed that the pooling estimator is biased in this data application, mostly because Conditions 5.4 and 5.5 did not hold. On the other hand, the point estimates of the proposed estimators (randomization-aware and combined estimators) were similar to the point estimates of IPW and AIPW estimators, indicating that the proposed estimators do not generate bias when augmenting the index trial using external controls even if Conditions 5.4 and 5.5 do not hold. In addition, the combined estimator's standard error is always smaller than the standard error of AIPW estimator; the efficiency improvement is larger with the decreasing sample size of the controls in the index trial.



(a) Scenario A (b) Scenario B

Figure 5.3: Distribution of $\hat{\lambda}^*$ for the combined estimator in the small trial case. Recall that when $\widehat{\lambda}^* = 0$ the combined estimator reduces to AIPW, and when $\widehat{\lambda}^* = 1$ it reduces to the optimized randomization-aware estimator. The histogram shows the normalized counts over 5000 repeated experiments.



(a) Varying number of parameters (b) Varying sample size

Figure 5.4: The computational time in seconds to for the proposed combined estimator when varying either the number of external controls or the number of observed covariates, while keeping everything else fixed. Each point corresponds to the median time over 10 repeated runs and the error bars show the min/max time.

Table 5.4: Simulation study to investigate effect on the optimized randomization-aware estimator when removing the weighting in its objective and another investigation of replacing the true propensity score with an estimated propensity score. We considered a small trial ($n_1 = 50$) and a large trial ($n_1 = 200$) where the number of external controls was kept fixed ($n_0 = 200$). We computed the mean absolute bias, variance and coverage rate from the estimated 95%-confidence intervals using 5000 repeated simulations.

	Scenario A						Scenario B					
	Small trial			Large trial			Small trial			Large trial		
Estimator	Bias	Var	Cov	Bias	Var	Cov	Bias	Var	Cov	Bias	Var	Cov
Optimized randomization-aware	0.00	0.29	0.96	0.00	0.02	0.94	0.01	0.73	0.93	0.00	0.18	0.94
Remove weighting in optimization	0.00	0.29	0.96	0.00	0.02	0.94	0.11	0.77	0.94	0.02	0.18	0.95
Estimate propensity score	0.00	0.31	0.97	0.00	0.02	0.94	0.02	0.80	0.91	0.00	0.18	0.94

Table 5.5: Estimates and standard errors of the different estimators with various sample sizes of the controls in the index trial (100%, 75%, 50%, 25% of the original sample size, 91). The results were the averages of 100 runs. The results of the test-then-pool estimator are not shown; they are the same as that of the pooling estimator.

	Unadjusted	Pooling	$\widehat{\tau}(\widehat{g})$	$\widehat{\tau}(\widehat{h}^*)$	$\widehat{\tau}(\widehat{\lambda}^*)$	Selective borrowing	Dynamic borrowing
68	-7.910 (13.483)	-9.534 (2.547)	-7.501 (3.164)	-7.241 (3.241)	-7.482 (3.159)	-7.487 (3.091)	-10.272 (2.707)
46	-8.134 (15.540)	-9.700 (2.663)	-7.667 (3.635)	-7.304 (3.709)	-7.657 (3.609)	-7.705 (3.592)	-10.698 (2.890)
23	-8.302 (20.623)	-9.888 (2.840)	-8.053 (4.680)	-7.001 (5.003)	-8.047 (4.544)	-8.301 (4.404)	-11.502 (3.090)

6

Estimating Heterogeneous Treatment Effects Leveraging External Data

Randomized trials are typically designed to detect average treatment effects but often lack the statistical power to uncover individual-level treatment effect heterogeneity, limiting their value for personalized decision-making. To address this, we propose the QR-learner, a model-agnostic learner that estimates conditional average treatment effects (CATE) within the trial population by leveraging external data from other trials or observational studies. The proposed method is robust: it can reduce the mean squared error relative to a trial-only CATE learner, and is guaranteed to recover the true CATE even when the external data are not aligned with the trial. Moreover, we introduce a procedure that combines the QR-learner with a trial-only CATE learner and show that it asymptotically matches or exceeds both component learners in terms of mean squared error. We examine the performance of our approach in simulation studies and apply the methods to a real-world dataset, demonstrating improvements in both CATE estimation and statistical power for detecting heterogeneous effects.

6.1. Introduction

By randomly assigning the interventions of interest, randomized trials are unique in their ability to estimate causal effects with little reliance on untestable assumptions. This strength has made trials the preferred approach for evaluating interventions across

This chapter appears as: Karlsson, R., De Bartolomeis, P., Dahabreh, I. J., & Krijthe, J. H. (2026). Robust estimation of heterogeneous treatment effects in randomized trials leveraging external data [Forthcoming]. *International Conference on Artificial Intelligence and Statistics*

many scientific domains. However, their high cost often limits sample size, which in turn limits the precision of statistical inferences that can be drawn from trial data. The problem is especially acute when the aim is not only to estimate an average treatment effect but also to characterize treatment effect heterogeneity (Lagakos *et al.*, 2006), a key step toward personalized decision-making for the population represented by the trial. A central quantity for this purpose is the conditional average treatment effect (CATE), which captures how treatment effects depend on individual-level covariates (Künzel *et al.*, 2019). However, the estimation of CATEs for different subgroups is more challenging than the estimation of average effects; therefore, the data from trials powered to detect average treatment effects are typically not adequate for the precise estimation of CATEs (Dahabreh *et al.*, 2016). As a result, accurately estimating CATEs within a trial population remains a difficult yet important challenge.

In recent years, there has been growing interest in augmenting trials with external data, mainly in the context of improving average treatment effect estimation (Jahanshahi *et al.*, 2021; van Rosmalen *et al.*, 2018). A key challenge in this setting is to properly account for differences between the trial population and the population underlying the external data (Ung *et al.*, 2024). These differences raise a fundamental concern: whether causal quantities such as the CATE remain stable across the two populations – a property known as transportability (Bareinboim & Pearl, 2016; Dahabreh & Hernán, 2019). In this paper, we investigate the analogous problem of using external data, such as from another trial or an observational study, to augment the estimation of CATEs in the trial population, in settings where the populations underlying the trial and external data may be misaligned. Here, we focus on the setting where transportability does not necessarily hold and the external data may be subject to unmeasured confounding. Our objective is to leverage external data for CATE estimation while ensuring that using the external data does not harm the estimation compared to if we had used trial data alone in cases of misalignment between the underlying populations.

Contributions We propose the QR-learner, a model-agnostic learner that improves estimation of the CATE in the population underlying a trial by leveraging external data from other trials or observational studies. We prove this learner is robust even when the external data are not aligned with the trial data: it recovers the true CATE even when external data come from a population which is not transportable with the trial population or are affected by uncontrolled confounding, while at the same time it can reduce the estimated CATE mean squared error compared to using trial data alone if the external data are sufficiently aligned (Sections 6.4.1-6.4.3). As another safeguard against potential harm from misaligned external data, we propose a procedure that combines the QR-learner with a trial-only CATE learner and prove that the combined learner asymptotically achieves a mean squared error that is no worse and potentially better than its component learners (Section 6.4.4). Using simulations and real-world data from the Student/Teacher Achievement Ratio (STAR) project (Krueger, 1999; Word *et al.*, 1990) we find that our method is robust when integrating external data that are not aligned with the trial data, and that it improves both CATE estimation mean squared error and the statistical power to detect treatment effect heterogeneity (Section 6.5 and 6.6).

6.2. Related works

Our work builds upon a rich and growing literature on CATE learners. Examples of approaches include adaptations of decision trees (Athey & Imbens, 2016), random forests (Wager & Athey, 2018), and neural networks (Shalit *et al.*, 2017). Our proposed learner is most closely related to model-agnostic "meta-learners" which allow for estimating the CATE and nuisance models using any supervised learning algorithm (Kennedy, 2023; Künzel *et al.*, 2019; Nie & Wager, 2021). However, most existing model-agnostic learners are tailored to settings where data are drawn from a single source, such as a single randomized trial or an observational study.

More recently, several CATE learners have been proposed for multi-source settings, often relying on the assumption that the CATE is transportable across the underlying populations (Hatt *et al.*, 2022; Shyr *et al.*, 2023; Wu & Yang, 2022, 2023). For example, Schweisthal *et al.* (2024) propose a learner that constructs bounds on the transportable CATE under unmeasured confounding across different populations, while Kallus *et al.* (2018) exploit the same assumption by estimating the CATE from a large, potentially confounded observational dataset and then apply a linear bias correction using data from a small randomized trial. Although some of these approaches might be used to estimate the trial-specific CATE under non-transportability, to our knowledge only Asiaee *et al.* (2023, 2025) explicitly address this setting; we discuss these methods in more detail later. Yang *et al.* (2023) studied the related problem of using multi-source data to perform valid statistical inference when estimating treatment effect heterogeneity under violations of transportability. However, their approach is restricted to a parametric linear working model for the CATE, whereas we focus specifically on optimizing the predictive performance (i.e., minimizing mean squared error) for a model-agnostic CATE learner which also allows for more flexible nonparametric working models.

Finally, our work draws on recent developments in the trial augmentation literature for average treatment effect estimation in trials using data from an external population. These developments have emphasized robustness to integrating external data misaligned with the trial data (De Bartolomeis *et al.*, 2025; M. Huang *et al.*, 2023; Karlsson, Wang *et al.*, 2026; Liao *et al.*, 2023; Schuler *et al.*, 2022). In particular, we will adapt ideas from the randomization-aware estimator framework proposed by Karlsson, Wang *et al.* (2026) to construct CATE learners that are also robust to misaligned external data.

6.3. Problem setting

Notation Let $X \in \mathscr{X}$ denote baseline (pre-treatment) covariates; S the binary indicator of data source ($S = 1$ for trial participants; $S = 0$ for individuals in the external data); A the binary indicator for treatment assignment ($A = 1$ for the experimental treatment; $A = 0$ denotes the control); and $Y \in \mathscr{Y}$ the outcome (continuous, binary, or count). Throughout, we use italic capital letters to denote random variables and lowercase letters for their specific values. We write $f(\cdot)$ to denote the density functions of random variables.

Study design and data structure We consider a non-nested trial design where the trial and external data are separately obtained and modeled as simple random samples from different populations, obtained with unknown and possibly unequal sampling fractions (Dahabreh *et al.*, 2021). For observation i with $S_i = s$, the data are modeled as i.i.d., conditional on study source, with the random tuple $O_i = (X_i, S_i = s, A_i, Y_i)$ for $i = 1, \ldots, n_s$, where n_s is the number of observations from source $S = s$. The composite dataset has total sample size $n = n_1 + n_0$, where the proportions of trial and external participants in the composite dataset may not reflect the size of their underlying populations. In the trial, treatment is randomly assigned according to the propensity score $e(X) = \Pr(A = 1 \mid X, S = 1)$ which is assumed to be known (Rosenbaum & Rubin, 1983b). As $n \to \infty$, we assume the ratios of the trial and external data sample sizes to the total sample size converge to some constants, i.e., $n_s/n \to q_s \in (0,1)$.

6.3.1. Identification of causal effects

To define the causal quantity of interest, we use potential outcomes (Rubin, 1974). For individual i and for $a \in \{0,1\}$, the potential outcome Y_i^a denotes the outcome under intervention to set treatment A to a, possibly contrary to fact. Our goal is to estimate the CATE in the population underlying the trial,

$$\tau(x) = \mathbb{E}[Y^1 - Y^0 \mid X = x, S = 1] \,. \tag{6.1}$$

Under standard conditions, the CATE $\tau(x)$ is identifiable from data in the trial population.

Condition 6.1 (Consistency)**.** *If $A_i = a$, then $Y_i^a = Y_i$ for every individual i and treatment $a \in \{0,1\}$.*

Condition 6.2 (Strong ignorability in the trial population)**.** *Positivity in trial: for each treatment $a \in \{0,1\}$, if $f(x, S = 1) \neq 0$, then $\Pr(A = a | X = x, S = 1) > 0$. Conditional exchangeability in trial: for each $a \in \{0,1\}$, $Y^a \perp\!\!\!\perp A | (X, S = 1)$.*

Conditions 6.1 and 6.2 are typically supported by a well-designed randomized trial and together suffice to identify the CATE as $\tau(x) = g_1(x) - g_0(x)$, where $g_a(x) = \mathbb{E}[Y \mid X = x, A = a, S = 1]$. However, it is common to assume additional conditions to enable identification and estimation of $\tau(x)$ using both the trial and external data.

Condition 6.3 (Strong ignorability in the external population)**.** *Positivity in external population: for each treatment $a \in \{0,1\}$, if $f(x, S = 0) \neq 0$, then $\Pr(A = a | X = x, S = 0) > 0$. Conditional exchangeability in external population: for each $a \in \{0,1\}$, $Y^a \perp\!\!\!\perp A | (X, S = 0)$.*

Condition 6.4 (Transportability)**.** *For each $a \in \{0,1\}$, $Y^a \perp\!\!\!\perp S \mid X$.*

The above two conditions can be controversial, especially when the external data originate from an observational study, because these conditions are uncertain and typically require

substantial domain expertise to justify. Notably, Conditions 6.1 to 6.4 together have testable implications that can be empirically assessed to falsify them, see e.g. Dahabreh *et al.* (2024), De Bartolomeis, Abad *et al.* (2024) and Z. Hussain *et al.* (2023). This can be used in particular to evaluate Conditions 6.3 and 6.4 because Condition 6.1 and 6.2 are supported by the trial's experimental design. Nonetheless, performing such falsification tests remains an inherently difficult task (Fawkes *et al.*, 2025).

6.4. Augmenting trials with external data

6.4.1. A class of robust pseudo-outcomes

Our goal is to learn a CATE function from a class of candidates $\mathscr{F}$ that minimizes the population risk relative to the true CATE function, namely $\arg\min_{\tilde{\tau}\in\mathscr{F}} R^*(\tilde{\tau})$ where $R^*(\tilde{\tau}) = \mathbb{E}[(\tau(X) - \tilde{\tau}(X))^2 \mid S = 1]$. However, as we cannot minimize $R^*(\tilde{\tau})$ directly because the true CATE $\tau(X)$ is unknown, we study the class of CATE learners obtained by minimizing a pseudo-risk (Foster & Syrgkanis, 2023),

$$\arg\min_{\tilde{\tau}\in\mathscr{F}} R(\tilde{\tau};\eta) \tag{6.2}$$

where $R(\tilde{\tau};\eta) = \mathbb{E}[(\psi(O;\eta) - \tilde{\tau}(X))^2 \mid S = 1]$. Here, we introduce an auxiliary random variable, sometimes referred to as a pseudo-outcome:

$$\begin{aligned}\psi(O_i;\eta) = &\frac{A_i - e(X_i)}{e(X_i)(1-e(X_i))}\big(Y_i - h_{A_i}(X_i)\big)\\ &+ h_1(X_i) - h_0(X_i)\end{aligned} \tag{6.3}$$

which is indexed by some nuisance models $\eta = \{h_1, h_0\}$, where $h_1 : \mathscr{X} \to \mathbb{R}$ and $h_0 : \mathscr{X} \to \mathbb{R}$ are real-valued functions defined on the covariate space $\mathscr{X}$. When clear from context, we omit the arguments and write $\psi_i = \psi(O_i;\eta)$ and $\widehat{\psi}_i = \psi(O_i;\hat{\eta})$ to denote the pseudo-outcomes computed using either the nuisance models η or replacing the nuisance models with their estimates $\hat{\eta}$.

Depending on our choice of η, we obtain different CATE learners when solving eq. (6.2). For instance, if $\eta = \{0, 0\}$, we obtain the (inverse) propensity weighted learner, refereed to as the PW-learner by Curth and Van der Schaar (2021). Meanwhile, if $\eta = \{g_1, g_0\}$, where $g_a = \mathbb{E}[Y \mid X, A = a, S = 1]$, we obtain the DR-learner (Kennedy, 2023). More generally, for any choice of η, we can prove an important robustness property of $R(\tilde{\tau};\eta)$ guaranteed by the trial's randomized design.

Theorem 6.1. *Under Conditions 6.1 and 6.2 where the propensity score $e(X)$ is known, for any fixed specification of the nuisance models η_{fixed}, the minimization problem in eq.* (6.2) *always yields the true CATE as its unique solution provided that $\tau \in \mathscr{F}$; that is, $\tau = \arg\min_{\tilde{\tau}\in\mathscr{F}} R(\tilde{\tau};\eta_{fixed})$.*

Although this result has appeared in the literature before (see, for example, Morzywolek *et al.* (2023) and references therein), we provide a derivation in Appendix 6.A.1 for completeness. We denote the nuisance models η_{fixed} with a fixed specification to emphasize that they must be chosen independently of the dataset used to compute the pseudo-outcome; this requirement can be satisfied using cross-fitting, which we describe in the following subsections.

Recognizing the central role of using the known propensity score in the above theorem, we refer to the pseudo-outcomes in eq. (6.3) constructed using the known propensity score as *randomization-aware* pseudo-outcomes (Karlsson, Wang *et al.*, 2026). This framing allows us to consider a broader class of pseudo-outcomes for robust CATE estimation, distinguished by varying specifications of the nuisance models η. While Theorem 6.1 guarantees that $R(\tilde{\tau};\eta)$ is a proper model selection criterion for the CATE regardless of the choice of η, as we will discuss next, the choice of η ultimately still plays a crucial role in estimating the CATE based on the observed data. This happens as we transition to the sample analog version of the minimization problem in eq. (6.2).

Remark 6.1. *While our main focus is on the pseudo-risk and pseudo-outcomes in eq.* (6.2) *and eq.* (6.3)*, another relevant pseudo-risk that may share similar robustness properties when having access to the true propensity score $e(X)$ is the one used by the R-learner (Nie & Wager, 2021). In Appendix 6.B, we outline potential connections between our results and the R-learner pseudo-risk, particularly in the special case where the propensity score $e(X)$ is constant.*

6.4.2. Using external data to improve CATE model selection in the trial population

To estimate the CATE from observed data, we must consider the sample analog of $R(\tilde{\tau};\eta)$, defined as $\widehat{R}(\tilde{\tau};\hat{\eta}) = \frac{1}{n_1}\sum_{i:S_i=1}\left(\psi\left(O_i;\hat{\eta}\right) - \tilde{\tau}\left(X_i\right)\right)^2$. We then obtain the CATE estimate by solving $\hat{\tau} = \arg\min_{\tilde{\tau}\in\mathscr{T}} \widehat{R}(\tilde{\tau};\hat{\eta})$.

Although the robustness property of randomization-aware pseudo-outcomes discussed earlier might suggest that the choice of nuisance models η is inconsequential, we will show that this is not the case. Because the model selection criterion $\widehat{R}(\tilde{\tau};\hat{\eta})$ is a sample average, in finite samples, this criterion can choose suboptimal CATE models and, importantly, this behavior is influenced by the choice of nuisance models η. To understand this, we first note that we can decompose the sample analog pseudo-risk as

$$\begin{aligned}\widehat{R}(\tilde{\tau};\hat{\eta}) = \frac{1}{n_1}\sum_{i:S_i=1}\Big[&(\tau(X_i)-\tilde{\tau}(X_i))^2\\ &-2(\tau(X_i)-\tilde{\tau}(X_i))(\widehat{\psi}_i-\tau(X_i))\\ &+(\widehat{\psi}_i-\tau(X_i))^2\Big].\end{aligned}$$

From the above decomposition, which we derive in Appendix 6.A.2, we see that $\widehat{R}(\tilde{\tau};\hat{\eta})$

consists of three parts: a first term that in expectation equals the population risk $R^*(\tilde{\tau}) = E[(\tau(X) - \tilde{\tau}(X))^2 \mid S = 1]$; a second term that introduces model selection uncertainty; and a third term which is independent of the candidate model $\tilde{\tau}$. Therefore, we may still end up selecting a suboptimal $\tilde{\tau}$, mainly due to the second term influencing $\widehat{R}(\tilde{\tau}; \hat{\eta})$.

To see why the second term is problematic for model selection, consider comparing two candidate models $\tilde{\tau}_1$ and $\tilde{\tau}_2$. If we observe $\widehat{\Delta}_{12} = \widehat{R}(\tilde{\tau}_1; \hat{\eta}) - \widehat{R}(\tilde{\tau}_2; \hat{\eta}) > 0$, we would conclude that $\tilde{\tau}_2$ is better than $\tilde{\tau}_1$; we would make the opposite decision when $\widehat{\Delta}_{12} < 0$, or remain inconclusive when $\widehat{\Delta}_{12} = 0$. However, note that only the third term from the decomposition cancels out in the risk difference $\widehat{\Delta}_{12}$. This means that in addition to the first term, which captures the true error relative to the true CATE, the second term can also influence our decision about which candidate model performs better.

To improve $\widehat{R}(\tilde{\tau}; \hat{\eta})$ as a model selection criterion, a natural strategy is to choose η to minimize the variance of the problematic second term. The next result provides insight into how this can be achieved (see proof in Appendix 6.A.3).

Lemma 6.1. *Under Conditions 6.1 and 6.2, assume the propensity score $e(X)$ is known and the nuisance functions $\hat{\eta} = \{\hat{h}_1, \hat{h}_0\}$ are estimated on a dataset independent of that used to compute the pseudo-outcomes $\widehat{\psi}$. Define $\epsilon := (\tau(X) - \tilde{\tau}(X))(\widehat{\psi} - \tau(X))$, then we have that $\mathbb{E}[\epsilon \mid S = 1] = 0$ and $\text{Var}(\epsilon \mid S = 1) \leq \tilde{C}\left[2\{L_1(\hat{h}_1) + L_0(\hat{h}_0)\} + \tilde{\sigma}^2\right]$ where*

$$
\begin{aligned}
L_a(\hat{h}_a) &= \mathbb{E}\left[w_a(X) \left(Y - \hat{h}_a(X)\right)^2 \mid A = a, S = 1 \right] \\
\tilde{\sigma}^2 &= \text{Var}(Y^1 - Y^0 \mid S = 1) - \text{Var}(\tau(X) \mid S = 1) \\
\tilde{C} &= \max_{x \in \mathscr{X}} (\tau(x) - \tilde{\tau}(x))^2
\end{aligned}
$$

with $w_a(X) = \left(\frac{1 - e(X)}{e(X)}\right)^{2a-1}$ for $a \in \{0, 1\}$.

The above result shows that a weighted mean squared error of the nuisance models $\hat{\eta} = \{\hat{h}_1, \hat{h}_0\}$ appears in an upper bound on the variance of the terms responsible for model selection uncertainty when using $\widehat{R}(\tilde{\tau}; \hat{\eta})$. This motivates our approach for selecting η: choose it to directly minimize both L_1 and L_0. Similar approaches to selecting η have appeared in prior work in single-source settings. Cao *et al.* (2009) addressed a related problem of mean estimation with missing data, while Saito and Yasui (2020) considered conditional average treatment effect estimation.

The tightness of the bound in the above lemma depends on the candidate model $\tilde{\tau}$ for the CATE. Specifically, it is proportional to the non-negative constant $\tilde{C} = \max_{x \in \mathscr{X}} (\tau(x) - \tilde{\tau}(x))^2$. Therefore, when the class of candidate models $\mathscr{T}$ contains good approximations of the true CATE τ and τ itself is bounded, $\tilde{C}$ is expected to be small, resulting in a relatively tight upper bound.

In the next result, we show how external data can be used to minimize the identified upper-bound in Lemma 6.1. To achieve this, we invoke the assumptions of strong ignorability in

the external data (Condition 6.3) and transportability (Condition 6.4). However, because these conditions are uncertain and may not hold in many settings, we recall that these conditions are not needed to ensure the robustness properties of the randomization-aware pseudo-outcomes. We later discuss a strategy for further improving robustness against these violations.

Theorem 6.2. *Under Conditions 6.3 and 6.4, in addition to those of Lemma 6.1, we can express*

$$L_a(\hat{h}_a) = \mathbb{E}\Big[\tilde{w}_a(X)\,(Y - \hat{h}_a(X))^2 \mid A = a\Big] \tag{6.4}$$

where $\tilde{w}_a(X) = \frac{\Pr(S=1|X,A=a)}{\Pr(S=1|A=a)}\left(\frac{1-e(X)}{e(X)}\right)^{2a-1}$.

The above theorem, proven in Appendix 6.A.4, shows that the function $L_a(\hat{h}_a)$, which appeared in the upper-bound of the variance $\mathrm{Var}(\epsilon \mid S = 1)$, can be rewritten in terms of all observed data from both the trial and external populations. Unlike the expression for $L_a(\hat{h}_a)$ in Lemma 6.1, we no longer condition on S. With this result, we can proceed with proposing a novel CATE learner that chooses the nuisance models η to minimize this upper-bound with the help of external data.

Remark 6.2. *The results in Theorem 6.2 are reminiscent of results from the transfer learning literature, where one express the expected mean squared error of a prediction function for data within some domain by reweighting data from another domain (Shimodaira, 2000; Weiss et al., 2016). To obtain this type of result in transfer learning, one typically assumes that the conditional distribution of labels remains constant across the two domains; this is analogous to how the transportability condition put certain requirements on the conditional distribution of the potential outcomes in our setting. However, unlike this literature, our goal is to be explicitly robust to violations of this condition, for which we propose a solution to this problem later in Section 6.4.4.*

6.4.3. The QR-learner algorithm

In this section, we introduce a novel method for estimating the CATE, which we call the *Quasi-optimized Randomization-aware* learner, or QR-learner. This method is model-agnostic, allowing it to use any supervised learning algorithm for estimating the CATE and nuisance models. It follows a two-stage procedure: In the first stage, it solves an optimization problem to select η using both trial and external data that minimizes the identified upper bound on the variance of the terms responsible for model selection uncertainty in finite samples. We call this step quasi-optimized because it targets an upper bound rather than the variance directly. In the second stage, the method regresses randomization-aware pseudo-outcomes on the covariates using only trial data to estimate the CATE. The first stage reduces finite-sample uncertainty for model selection, while the second stage leverages the robustness of the randomization-aware pseudo-outcomes to target the true CATE in the trial population.

To prevent overfitting and ensure that $\hat{\eta}$ is estimated independently of the data used for pseudo-outcomes, we employ a cross-fitting procedure that partitions the data into $\mathscr{D}^1 \cup \mathscr{D}^2$, stratified by treatment A and study population S. In the first stage, we use $\mathscr{D}^1$ to estimate $\hat{\eta}^* = \{\hat{h}_1^*, \hat{h}_0^*\}$ where each component $\hat{h}_a^*$ is obtained by solving the optimization problem

$$\min_{h_a \in \mathscr{H}} \sum_{i \in \mathscr{D}_a^1} \widehat{\pi}_a(X_i) \Big(\frac{1 - e(X_i)}{e(X_i)} \Big)^{2a-1} \big(Y_i - h_a(X_i) \big)^2 \tag{6.5}$$

where the sum is taken over $\mathscr{D}_a^1 = \{i \in D^1 : A_i = a\}$, $\mathscr{H}$ is the model class under consideration for the nuisance models $\eta = \{h_1, h_0\}$, and $\widehat{\pi}_a$ is an estimator from the model class $\mathscr{G}$ for the probability of trial participation $\Pr(S = 1 \mid X = x, A = a)$ which is also estimated using $\mathscr{D}_a^1$. Then, in the second stage, the pseudo-outcomes $\psi(O_i; \hat{\eta}^*)$ are computed on $\mathscr{D}^2$ using the estimated nuisance models $\hat{\eta}^*$. We estimate the CATE using $\mathscr{D}^2$ by solving $\hat{\tau} = \arg\min_{\tilde{\tau} \in \mathscr{F}} \sum_{i \in \mathscr{D}^2 : S_i = 1} (\psi(O_i; \hat{\eta}^*) - \tilde{\tau}(X_i))^2$. To efficiently use all available data, we reverse the roles of the splits to obtain a second CATE estimator, and then take the average of the predictions from the two resulting estimators; this procedure naturally extends to more than two data splits if desired. Pseudo-code for the QR-learner is also provided in Appendix 6.D.

We now discuss several noteworthy remarks about the QR-learner. First, we recommend using a linear logistic regression with a cross-validated ridge penalty for estimating $\widehat{\pi}_a$. The reason we use a linear model is that, although the external sample size is large, the trial sample size may be much smaller, so a more flexible model could still overfit or have poor calibration. It is also known that maximum likelihood estimation of logistic regression can be biased in small-sample settings or when the number of observations from one of the events is rare. Adding a shrinkage penalty, as we do here, can reduce this type of bias; see, for example, Firth (1993) and Leitgöb (2020) for more detailed discussion.

Second, while our primary focus is to predict the CATE well with respect to the risk $R^*(\hat{\tau}) = \mathbb{E}[(\tau(X) - \hat{\tau}(X))^2 \mid S = 1]$, we note that in some cases, it may also be possible to use the estimated function $\hat{\tau}(X)$ obtained from the QR-learner for inference on the CATE. Although a full theoretical treatment is beyond the scope of this work, we provide supporting arguments in Appendix 6.C to explain why this may be justified, and will also show in a simulation study that this is feasible.

Third, our motivation for proposing the QR-learner stems from Theorem 6.2, which shows that minimizing the objective in eq. (6.5) can improve CATE model selection. Here we provide another argument for why the QR-learner may outperform a trial-only learner, specifically the DR-learner. Both learners differ only in how their outcome nuisance models are estimated, and it is known that the DR-learner's convergence rate depends on errors in both estimating the CATE function in the second stage and the outcome nuisance models in the first stage (Kennedy, 2023). The QR-learner can improve upon the DR-learner by estimating these outcome nuisance models more accurately using both trial and external data when populations are well-aligned. With sufficiently large external data, outcome nuisance estimation errors become negligible, and we could expect the QR-learner to outperform the DR-learner, which is also what we observe in our

simulations.

Finally, the CFACE learner proposed by Asiaee *et al.* (2023) shares close similarities with the QR-learner. Their method differs from ours in the first stage by estimating nuisance models $\eta = \{m^*, m^*\}$, where $m^*(x) = e(x) \cdot \mu_0(x) + (1 - e(x)) \cdot \mu_1(x)$ with $\mu_a(x) = \mathbb{E}[Y \mid X = x, A = a, S = 0]$. This formulation comes from solving

$$\arg\min_m \mathrm{Var}\left(\psi(O; \eta = \{m, m\}) \mid X = x, S = 1\right) ,$$

with nuisance models estimated entirely from external data. Their approach works well when external data are abundant and aligned with the trial population, but can fail if the populations differ substantially or external data are limited. This pitfall was noted by the same authors in Asiaee *et al.* (2025), where they propose another learner called R-OSCAR. This learner differs more substantially from ours and instead resembles the bias correction approach of Kallus *et al.* (2018), but tailored specifically for estimating the CATE in the underlying population of a randomized trial. We further discuss the procedures of both CFACE and R-OSCAR in Appendix 6.E.

6.4.4. Combining CATE learners

The success of the optimization in eq. (6.5) relies on two conditions that are not necessary for the identification of the CATE in the trial population. First, it requires that the external data are aligned with the trial data: that is, Conditions 6.3 and 6.4 hold. Second, for each $a \in \{0, 1\}$, the estimator $\hat{\pi}_a(x)$ needs to be a correctly specified model of the probability $\Pr(S = 1 \mid X = x, A = a)$. As a result, while the QR-learner always targets the correct CATE in the trial population, it may perform worse in finite samples, in terms of mean squared error, than a CATE learner based on the trial data alone. To address this issue, we therefore propose a variant which combines the potential benefits from using external data via the QR-learner with the additional robustness from a trial-only learner. Specifically, we use the DR-learner (Kennedy, 2023) obtained by regressing randomization-aware pseudo-outcomes $\psi(O; \hat{\eta} = \{\hat{g}_1, \hat{g}_0\})$ on the covariates X where $\hat{g}_a$ estimates $\mathbb{E}[Y \mid X, A = a, S = 1]$ using only trial data.

Our proposed combined learner is defined as

$$\hat{\tau}(x; \lambda) := \lambda \cdot \hat{\tau}_{\mathrm{QR}}(x) + (1 - \lambda) \cdot \hat{\tau}_{\mathrm{DR}}(x),\ \lambda \in [0, 1] ,$$

where $\hat{\tau}_{\mathrm{QR}}$ is the estimator from the QR-learner using both trial and external data, and $\hat{\tau}_{\mathrm{DR}}$ is the estimator from the DR-learner using trial data alone; this combined learner can be viewed as an instance of stacked regression (Breiman, 1996) for CATE estimation. We recommend using the DR-learner as the trial-only learner here because it is also robust to misspecification of the outcome models, even though the theoretical guarantees in this section also hold if we replace it with another CATE learner fitted using only the trial data.

Recall the population risk introduced in Section 4.1, $R^*(\tilde{\tau}) = \mathbb{E}\big[(\tau(X) - \tilde{\tau}(X))^2 \mid S = 1\big]$. For the combined learner we write $R^*(\lambda) := R^*\big(\hat{\tau}(\cdot; \lambda)\big)$ with $R^*_{\mathrm{DR}} = R^*(0)$ and $R^*_{\mathrm{QR}} = R^*(1)$.

Our goal in this section is to choose λ such that $R^*(\lambda)$ is no larger than $\min\{R^*_{\text{DR}}, R^*_{\text{QR}}\}$. We first show that for the oracle weights the mean squared error of the combined estimator is no worse than the individual components (see proof in Appendix 6.A.5).

Lemma 6.2. *Let $\lambda^\star = \arg\min_{\lambda\in\Lambda} R^*(\lambda)$ be the oracle weight then $R^*(\lambda^\star) \leq \min\{R^*_{QR}, R^*_{DR}\}$. Furthermore, if $\mathbb{E}[(\hat{\tau}_{QR}(X) - \hat{\tau}_{DR}(X))^2 \mid S = 1] > 0$, then the oracle weight $\lambda^\star$ is unique.*

In practice, the population risk cannot be computed directly, so we minimize the sample analog pseudo-risk $\widehat{R}(\lambda) = \frac{1}{n_1}\sum_{i:S_i=1}(\widehat{\psi}_i - \hat{\tau}(X_i;\lambda))^2$. We show that an empirical estimate of the oracle weight λ^* can be obtained via cross-validation and achieves the same risk asymptotically as the oracle. Let the trial data be partitioned into K mutually exclusive folds. For each observation i, let $\hat{\tau}^{(-i)}_{\text{DR}}$ and $\hat{\tau}^{(-i)}_{\text{QR}}$ denote the learners trained on data excluding the fold containing i, and define $\hat{\tau}^{(-i)}(x;\lambda) := \lambda\cdot\hat{\tau}^{(-i)}_{QR} + (1-\lambda)\cdot\hat{\tau}^{(-i)}_{DR}$. We consider computing the pseudo-outcomes with $\eta = \{0,0\}$, yielding a model-free combination approach. The K-fold cross-validated pseudo-risk estimate is $\widehat{R}_{CV}(\lambda) = \frac{1}{n_1}\sum_{i=1}^{n_1}\big(\psi_i - \hat{\tau}^{(-i)}(x;\lambda)\big)^2$. The cross-validated weight is then obtained as $\hat{\lambda}^\star = \arg\min_{\lambda\in\Lambda}\widehat{R}_{CV}(\lambda)$.

Theorem 6.3. *Assume that the pseudo-outcome $\psi(O;\eta = \{0,0\})$ and the base learners $\hat{\tau}_{QR}(X)$ and $\hat{\tau}_{DR}(X)$ are uniformly bounded. Then, we have $R^*(\hat{\lambda}^*) \leq R^*(\lambda^\star) + o_p(1) \leq \min\{R_{DR}, R_{QR}\} + o_p(1)$ where $o_p(1)$ converges to 0 as $n\to\infty$.*

The above result describes the oracle excess risk of the combined learner and guarantees that selecting the parameter λ via cross-validation yields a combined learner whose risk differs from the best possible risk by a vanishing term. In other words, the cross-validated choice is asymptotically as good as the oracle-optimal choice and, in large samples, performs at least as well as the better of the two candidate learners. Note that our setting does not require the two component learners to converge to the same CATE function. If they use different model classes and converge to different functions, a linear combination may be strictly better than either component learner. In the special case where both learners converge to the same function such that $R_{DR} = R_{QR}$, the combined learner also converges to this function.

As a final remark, we note that an alternative approach to combine the QR- and DR-learner would be to take the pseudo-outcomes from their respective first stages and fit a CATE model to a linear combination of these pseudo-outcomes. However, this is a less general solution since our formulation allows the theoretical guarantees in this section to hold for any trial-only learner, including those that do not rely on learning the CATE using pseudo-outcomes such as the S- or T-learner (Künzel *et al.*, 2019).

6.5. Simulation study

We conduct a series of simulations to evaluate the performance of our proposed method against several baselines, focusing on CATE prediction mean squared error. We also

assess statistical power to detect treatment effect heterogeneity through interaction tests between the treatment and a hypothesized effect modifier, an analysis commonly performed in real-world trials. For this second evaluation, we consider only methods that support statistical inference. The code to reproduce our experiments is available in the public Github repository https://github.com/RickardKarl/robust-trial-CATE-augmentation.

Baselines In addition to our proposed QR-learner and combined learner, we apply the DR-learner with known propensity scores using only trial data (Kennedy, 2023); the T-learner, which computes CATE estimates as $\hat{g}_1(x) - \hat{g}_0(x)$ with the same nuisance models that are used in the DR-learner; a pooled variant of the T-learner obtained by estimating $\mathbb{E}[Y \mid X, A = 1] - \mathbb{E}[Y \mid X, A = 0]$ using both the trial and external data; the method proposed by Asiaee *et al.* (2025) called R-OSCAR, in addition to a method from a previous version of their manuscript called CFACE (see Asiaee *et al.* (2023); an extended discussion on their differences is found in Appendix 6.E); and the linear additive bias correction method of Kallus *et al.* (2018) which we refer to as KSP (after the authors' initials). Implementation details are provided in Appendix 6.F.2.

6.5.1. Influence of external data sample size and population misalignment

First, we assess the root mean squared error (RMSE) of CATE predictions in a setting with a fixed trial size while varying the size of the external dataset. We consider two scenarios: (i) an idealized setting in which both Conditions 6.3 and 6.4 hold – i.e., the external data are unconfounded and transportability holds – and (ii) a more realistic, challenging setting where these assumptions are violated. We use gradient boosting regressors to estimate the outcome nuisance components and a linear model to estimate both the trial participation probability and the CATE; this modeling choice aligns with the data-generating process, where the baseline outcome is a highly nonlinear function of the covariates, while the CATE is linear.

Table 6.1 shows that our proposed methods consistently achieve the lowest or near-lowest RMSE across all settings, with performance improving as the external sample size increases. The trial-only DR- and T-learners performs no better than simply predicting an estimated average treatment effect (ATE) for all individuals, indicating the difficulty of this task based on trial data alone. While all integrative methods perform best when Conditions 6.3 and 6.4 hold, it is noteworthy when these conditions are violated that our methods exhibit RMSEs comparable to that of the trial-only DR-learner. CFACE shows a similar pattern, except when the external dataset is small ($n_0 = 100$), where it performs worse than both the trial-only DR-learner and the ATE predictor. This suggests that their nuisance model fitting is somewhat less robust than ours, as it can underperform relative to the best trial-only CATE learner. Interestingly, R-OSCAR and KSP underperform in all settings; both of these methods rely on fitting separate functions to model the bias

between the trial and external population. The difficulty of this bias modeling when the trial size is small could likely explain why both these methods fail.

When examining average prediction bias (full results in Appendix 6.G), we find that all methods are largely unbiased except the pooled T-learner. This is expected since we fit linear CATE models that match the true CATE function. Thus, the observed RMSE gains among the integrative methods using external data primarily stem from variance reduction rather than bias reduction, provided the model is correctly specified.

6.5.2. Assessing statistical power to detect interaction effects

We next assess statistical power to detect treatment effect heterogeneity via interaction tests between the treatment and a hypothesized effect modifier when integrating external data while transportability (Condition 6.4) is violated. This analysis is conducted in a setting with five observed covariates X, where one is a potential effect modifier $Z \subseteq X$. We consider two scenarios: one in which the effect modifier is present, and another where it is absent. Methods based on the T-learner, R-OSCAR, KSP and the combined learner are excluded, as they do not support inference on the estimated CATE model $\hat{\tau}$. For the remaining methods, we regress $\hat{\tau}$ on Z and perform a two-sided test on the Z coefficient. As a baseline, we also include a simpler test aligned with standard practice in trial analyses, fitting a linear regression of Y on $(A, Z, A \cdot Z)$, testing the $A \cdot Z$ interaction using either only the trial data or the pooled data. We use a significance level $\alpha = 0.05$. Full implementation details of the tests are provided in Appendix 6.F.3.

The results in Figure 6.1 show that when varying the trial sample size with the external dataset size fixed at $n_0 = 1000$, all methods maintain nominal type I error in the absence of an effect modifier except for the pooled covariate adjustment which can be explained by transportability being violated. Among the methods with nominal type 1 error, the QR-learner and CFACE consistently improve the power by about 10 to 20 percentage points to detect the effect modifier when it is present.

6.6. Case study: STAR dataset

We use data from the Tennessee Student/Teacher Achievement Ratio (STAR) project (Krueger, 1999; Word *et al.*, 1990), a large-scale randomized trial on the effects of class size on student performance. The dataset is divided into two populations based on school location: rural and urban schools. Since both originate from a randomized study, strong ignorability is expected to hold within each population. However, outcome distributions differ across rural and urban schools, and by deliberately omitting school location from the observed covariates, we create a setting where transportability between the two populations is violated. Additional details on the data are provided in Appendix 6.F.4. Performance is measured using RMSE on a held-out test set from the target population, where the pseudo-outcome $\psi(O_i; \eta = 0, 0)$ serves as a proxy for the true CATE. Although

Table 6.1: Average root mean squared error reported over 500 repeated runs from the simulation study with a trial sample size $n_1 = 250$ under different scenarios. Lowest number for each scenario is marked with bold.

External sample size	100		1000		10000	
Condition 3 and 4 violated?	No	Yes	No	Yes	No	Yes
Predict ATE	0.31 (1e-3)	**0.31** (1e-3)	0.31 (1e-3)	0.31 (1e-3)	0.31 (1e-3)	0.31 (1e-3)
DR-learner	0.28 (4e-3)	0.32 (4e-3)	0.28 (4e-3)	0.32 (4e-3)	0.27 (4e-3)	0.32 (4e-3)
T-learner	0.55 (3e-3)	0.55 (3e-3)	0.55(3e-3)	0.55 (3e-3)	0.55 (3e-3)	0.55 (3e-3)
Pooled T-learner	0.52 (2e-3)	0.60 (3e-3)	0.47 (9e-4)	0.60 (1e-3)	0.33 (7e-4)	0.48 (1e-3)
CFACE (Asiaee *et al.*, 2023)	0.34 (5e-3)	0.36 (4e-3)	0.24 (4e-3)	0.30 (3e-3)	**0.19** (3e-3)	0.28 (3e-3)
R-OSCAR (Asiaee *et al.*, 2025)	0.60 (4e-03)	0.60 (4e-03)	0.52 (2e-03)	0.58 (2e-03)	0.37 (1e-03)	0.41 (2e-03)
KSP (Kallus *et al.*, 2018)	0.71 (1e-2)	0.76 (1e-2)	0.72 (1e-2)	0.77 (1e-2)	0.71 (1e-2)	0.77 (1e-2)
QR-learner (ours)	**0.28** (4e-3)	0.32 (4e-3)	**0.23** (3e-3)	**0.29** (3e-3)	**0.19** (3e-3)	**0.27** (3e-3)
Combined learner (ours)	0.29 (4e-3)	0.32 (4e-3)	**0.23** (3e-3)	**0.29** (4e-3)	**0.19** (3e-3)	**0.27** (3e-3)

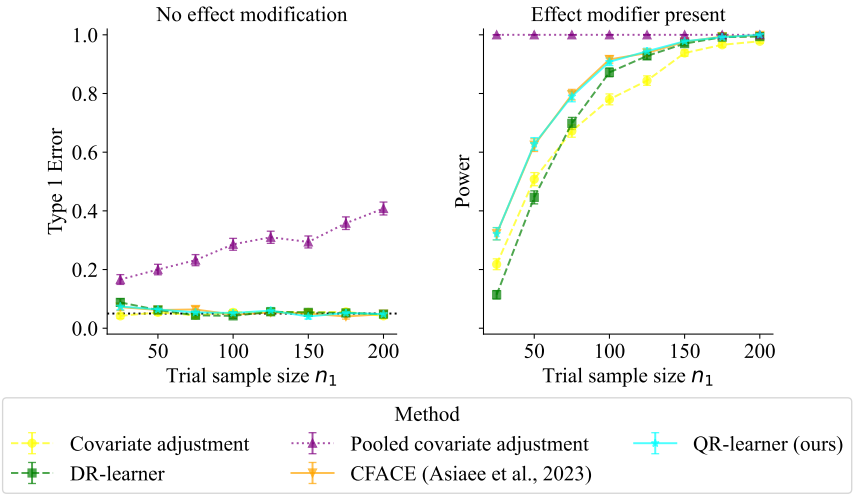


Figure 6.1: Evaluating type 1 error (lower better) and power (higher better) in the simulation study with the methods applicable for statistically testing for the presence of an effect modifier as sample size in trial increases, reported over 500 repeated runs.

this proxy has high variance, the use of the known trial propensity score ensures that the estimate aligns in expectation with the true population risk, as discussed in Section 6.4. Gradient boosting is used for estimating the outcome nuisance models, and ridge-penalized linear models to estimate both the trial participation probability and the CATE.

We evaluate the RMSE of each method as the trial sample size increases, while fixing the external data size at $n_0 = 1000$. In Figure 6.2, urban schools are treated as the target trial population ($S = 1$) and rural schools as the external population ($S = 0$); the flipped

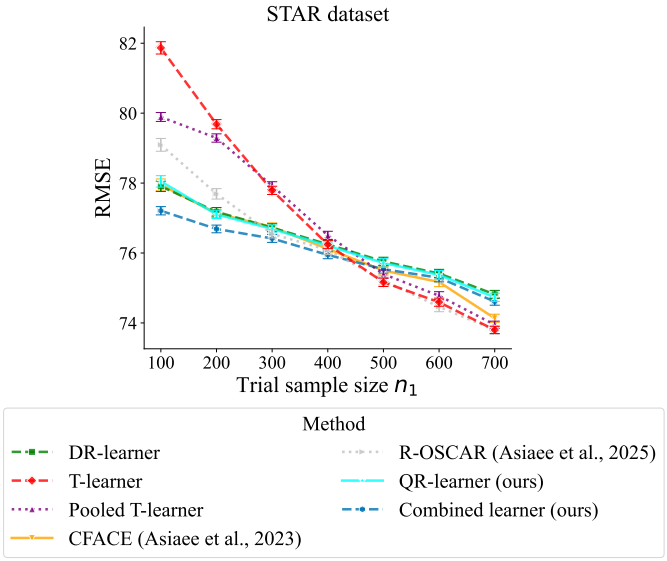


Figure 6.2: RMSE on the STAR dataset when increasing the trial sample size with a fixed external sample size of $n_0 = 1000$. Average RMSE and standard error are reported over 200 repeated runs.

setting is reported in Appendix 6.G. KSP performed significantly worse than the other methods and is therefore omitted from the main figure for ease of visualization, though it is included in the appendix. The integrative methods mostly benefit from external data when the trial sample is small, with gains diminishing as the trial grows. The combined learner consistently outperforms its component learners, the DR- and QR-learners, as predicted by our theory, and also outperforms CFACE and R-OSCAR at smaller sample sizes. The QR-learner offers no improvement over the DR-learner in this setting, though in the flipped setting we observe that it improve over the DR-learner. When the trial size becomes large, the trial-only and pooled T-learners, along with CFACE and R-OSCAR, achieve the lowest RMSE. The margin between these approaches and the DR-learner also grows, although this margin depends on whether the target population is urban or rural. This suggests that the advantage may reflect asymmetries between the two populations rather than a general property of the methods. Overall, the results show that our proposed CATE learners can effectively leverage external data to improve prediction accuracy.

6.7. Discussion

Our experimental findings demonstrate that the proposed learners for estimating the CATE using external data can effectively reduce the mean squared error of CATE estimates, while remaining robust in scenarios where the external data has unmeasured confounders or transportability is violated. Notably, in cases where the DR-learner failed to outperform

a simple baseline that predicts the average treatment effect – thus providing limited value – our proposed learners were still able to achieve better accuracy. This highlights the potential value of incorporating external data into analyses of heterogeneous treatment effects in randomized trials.

While our proposed methodology performs as intended, it is important to acknowledge its limitations, particularly given its potential impact on real-world decision-making in public health and policy. First, we observe an upper limit to the incremental gain of incorporating additional external data. This is because the external data primarily improves the first-stage estimation of the QR-learner, whereas the second stage relies solely on the randomized trial and is therefore constrained by the trial sample size. Second, although we observe that minimizing the upper bound in Lemma 6.1 improves CATE accuracy empirically and our method appears robust, it remains of interest to explore under what scenarios tighter bounds can be obtained or more direct strategies for improving CATE estimation can be developed.

Appendices

6.A. Proofs and derivations

6.A.1. Proof of Theorem 6.1

Proof. We write $\psi_{\text{fixed}} = \psi(O;\eta_{\text{fixed}})$.

First, we need to prove conditional unbiasedness, $\mathbb{E}[\psi_{\text{fixed}} \mid X = x, S = 1] = \tau(x)$, as follows:

$$\begin{aligned}\mathbb{E}[\psi_{\text{fixed}} \mid X = x] &= \mathbb{E}\left[\frac{A}{e(X)}(Y - h_1(X)) - \frac{1-A}{1-e(X)}(Y - h_0(X)) + h_1(X) - h_0(X) \mid X = x, S = 1\right] \\ &= \mathbb{E}\left[\frac{A}{e(X)}(Y^1 - h_1(X)) - \frac{1-A}{1-e(X)}(Y^0 - h_0(X)) \mid X = x, S = 1\right] + h_1(x) - h_0(x)\end{aligned}$$

where the second equality follows from consistency in Condition 6.1. Next, we inspect the first term inside the above expectation, which can be rewritten as follows

$$\begin{aligned}\mathbb{E}\left[\frac{A}{e(X)}(Y^1 - h_1(X)) \mid X = x, S = 1\right] &= \mathbb{E}\left[\frac{A}{e(X)} \mid X = x, S = 1\right]\left(\mathbb{E}\left[Y^1 \mid X = x, S = 1\right] - h_1(x)\right) \\ &= \frac{e(x)}{e(x)}\left(\mathbb{E}\left[Y^1 \mid X = x, S = 1\right] - h_1(X)\right) \\ &= \mathbb{E}\left[Y^1 \mid X = x, S = 1\right] - h_1(X)\end{aligned}$$

where the first equality follows from conditional exchangeability in the trial population, $Y^a \perp\!\!\!\perp A \mid X, S = 1$, in Condition 6.2 and the second equality follows from that $\mathbb{E}[A \mid X = x, S = 1] = e(x)$. Similarly, we can show that

$$\mathbb{E}\left[\frac{1-A}{1-e(X)}(Y^0 - h_0(X)) \mid X = x, S = 1\right] = \mathbb{E}\left[Y^0 \mid X = x, S = 1\right] - h_0(X)\,.$$

Putting all of the above together, we see that

$$\mathbb{E}[\psi_{\text{fixed}} \mid X = x] = \mathbb{E}\left[Y^1 - Y^0 \mid X = x, S = 1\right] = \tau(x)$$

Next, we show that $\tau = \arg\min_{\tilde{\tau}\in\mathscr{F}} R(\tilde{\tau};\eta_{\text{fixed}})$ when $\tau \in \mathscr{F}$. By adding and subtracting $\tau(X)$ inside $R(\tilde{\tau};\eta_{\text{fixed}})$, we can decompose it as

$$\underbrace{\mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))^2 \mid S = 1\right]}_{(a)} - \underbrace{\mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))\left(\psi_{\text{fixed}} - \tau(X)\right) \mid S = 1\right]}_{(b)} + \underbrace{\mathbb{E}\left[\left(\psi_{\text{fixed}} - \tau(X)\right)^2 \mid S = 1\right]}_{(c)}$$

First, we see that $(a) = R^*(\tilde{\tau})$. Next, we have that $(b) = 0$ because

$$\begin{aligned}&\mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))\left(\psi_{\text{fixed}} - \tau(X)\right) \mid S = 1\right] = \\ &= \mathbb{E}\left[\mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))\left(\psi_{\text{fixed}} - \tau(X)\right) \mid X, S = 1\right] \mid S = 1\right] \\ &= \mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))\,\mathbb{E}\left[\left(\psi_{\text{fixed}} - \tau(X)\right) \mid X, S = 1\right] \mid S = 1\right] \\ &= 0\end{aligned}$$

where the last equality follows from conditional unbiasedness such that

$$\mathbb{E}\left[\left(\psi_{\text{fixed}} - \tau(X)\right) \mid X, S=1\right] = 0.$$

Finally, $(c) = C$ is a real-valued constant $C \geq 0$ independent of $\tilde{\tau}$. Thus, we can write that

$$R(\tilde{\tau}; \eta_{\text{fixed}}) = R^*(\tilde{\tau}) + C$$

which implies that

$$\underset{\tilde{\tau}\in\mathscr{F}}{\arg\min}\, R(\tilde{\tau}; \eta_{\text{fixed}}) = \underset{\tilde{\tau}\in\mathscr{F}}{\arg\min}\{R^*(\tilde{\tau}) + C\} = \underset{\tilde{\tau}\in\mathscr{F}}{\arg\min}\, R^*(\tilde{\tau}) = \tau\ .$$

□

6.A.2. Decomposition of the sample analog pseudo-risk

We have that

$$\begin{aligned}
\widehat{R}(\tilde{\tau}; \hat{\eta}) &= \frac{1}{n_1} \sum_{i:S_i=1} \left(\widehat{\psi}_i - \tilde{\tau}(X_i)\right)^2 \\
&= \frac{1}{n_1} \sum_{i:S_i=1} \left(\widehat{\psi}_i - \tau(X_i) + \tau(X_i) - \tilde{\tau}(X_i)\right)^2 \\
&= \frac{1}{n_1} \sum_{i:S_i=1} \left\{ (\tau(X_i) - \tilde{\tau}(X_i))^2 - 2\left(\tau(X_i) - \tilde{\tau}(X_i)\right)\left(\widehat{\psi}_i - \tau(X_i)\right) + \left(\widehat{\psi}_i - \tau(X_i)\right)^2 \right\}.
\end{aligned}$$

6.A.3. Proof of Lemma 6.1

Proof. To make it more explicit that the estimated nuisance models $\hat{\eta}$ are obtained independently of the observations used to compute the pseudo-outcomes, we denote it as $\eta_{\text{fixed}} = \{h_1, h_0\}$. Moreover, we write $\psi_{\text{fixed}} = \psi(O; \eta_{\text{fixed}})$.

Defining $\epsilon := \left(\tau(X) - \tilde{\tau}(X)\right)\left(\psi_{\text{fixed}} - \tau(X)\right)$, we then have that $\mathbb{E}[\epsilon \mid S=1] = 0$ which we showed in the proof of Theorem 6.1. Next, we note that

$$\begin{aligned}
\text{Var}(\epsilon \mid S=1) &= \mathbb{E}\left[\epsilon^2 \mid S=1\right] \\
&= \mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))^2 \left(\psi_{\text{fixed}} - \tau(X)\right)^2 \mid S=1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))^2 \left(\psi_{\text{fixed}} - \tau(X)\right)^2 \mid X, S=1\right] \mid S=1\right] \\
&= \mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))^2 \mathbb{E}\left[\left(\psi_{\text{fixed}} - \tau(X)\right)^2 \mid X, S=1\right] \mid S=1\right] \\
&= \mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))^2 \text{Var}\left(\psi_{\text{fixed}} \mid X, S=1\right) \mid S=1\right]
\end{aligned}$$

where the first equality follows from that $\mathbb{E}[\epsilon \mid S=1] = 0$ and the last from the conditional unbiasedness of the pseudo-outcome, $\mathbb{E}[\psi_{\text{fixed}} \mid X = x, S=1] = \tau(x)$, which we derived in the proof of Theorem 6.1.

Next, we show how to upper-bound $\text{Var}(\epsilon \mid S=1)$ as follows:

$$\begin{aligned}\text{Var}(\epsilon \mid S=1) &= \mathbb{E}\left[(\tau(X)-\tilde{\tau}(X))^2 \text{Var}\left(\widehat{\psi} \mid X, S=1\right) \mid S=1\right] \\ &\leq \tilde{C} \cdot \mathbb{E}\left[\text{Var}\left(\psi_{\text{fixed}} \mid X, S=1\right) \mid S=1\right]\end{aligned}$$

where the inequality holds if define the constant $\tilde{C} = \max_{x \in \mathscr{X}} (\tau(x)-\tilde{\tau}(x))^2 \geq 0$. Next, we have from the law of total variance that

$$\begin{aligned}\mathbb{E}\left[\text{Var}\left(\psi_{\text{fixed}} \mid X, S=1\right) \mid S=1\right] &= \text{Var}(\psi_{\text{fixed}} \mid S=1) - \text{Var}\left(\mathbb{E}\left[\psi_{\text{fixed}} \mid X, S=1\right] \mid S=1\right) \\ &= \text{Var}(\psi_{\text{fixed}} \mid S=1) - \text{Var}\left(\tau(X) \mid S=1\right) \ .\end{aligned}$$

where the second equality follows from the conditional unbiasedness of the pseudo-outcomes. Thus, so far, we have $\text{Var}(\epsilon \mid S=1) \leq \tilde{C}\left[\text{Var}(\psi_{\text{fixed}} \mid S=1) - \text{Var}\left(\tau(X) \mid S=1\right)\right]$.

Next, we inspect the variance $\text{Var}(\psi_{\text{fixed}} \mid S=1)$, which we rewrite using the law of total variance:

$$\begin{aligned}\text{Var}(\psi_{\text{fixed}} \mid S=1) &= \mathbb{E}\left[\text{Var}\left(\psi_{\text{fixed}} \mid Y^1, Y^0, X, S=1\right) \mid S=1\right] \\ &\quad + \text{Var}\left(\mathbb{E}\left[\psi_{\text{fixed}} \mid Y^1, Y^0, X, S=1\right] \mid S=1\right) \\ &= \mathbb{E}\left[\text{Var}\left(\psi_{\text{fixed}} \mid Y^1, Y^0, X, S=1\right) \mid S=1\right] \\ &\quad + \text{Var}\left(Y^1 - Y^0 \mid S=1\right)\end{aligned}$$

where the second inequality follows from that

$$\mathbb{E}\left[\psi_{\text{fixed}} \mid Y^1, Y^0, X, S=1\right] = \mathbb{E}\left[Y^1 - Y^0 \mid Y^1, Y^0, X, S=1\right] = Y^1 - Y^0$$

where the first equality stems from the conditional unbiasedness of the pseudo-outcomes.

We next write $\psi_{\text{fixed}} = \psi_{1,\text{fixed}} - \psi_{0,\text{fixed}}$ where $\psi_{a,\text{fixed}} = \frac{\mathbf{1}(A=a)}{\mathbf{1}(A=1)e(X)+\mathbf{1}(A=0)(1-e(X))}(Y-h_a(X)) + h_a(X)$. This will help us simplify the expression for the above inner conditional variance as follows,

$$\begin{aligned}&\text{Var}\left(\psi_{\text{fixed}} \mid Y^1, Y^0, X, S=1\right) = \text{Var}\left(\psi_{1,\text{fixed}} - \psi_{0,\text{fixed}} \mid Y^1, Y^0, X, S=1\right) \\ &= \mathbb{E}\left[\left\{\psi_{1,\text{fixed}} - \psi_{0,\text{fixed}} - \underbrace{\mathbb{E}\left[\psi_{1,\text{fixed}} - \psi_{0,\text{fixed}} \mid Y^1, Y^0, X, S=1\right]}_{=Y^1-Y^0}\right\}^2 \mid Y^1, Y^0, X, S=1\right] \\ &= \mathbb{E}\left[\left\{\left(\psi_{1,\text{fixed}} - Y^1\right) - \left(\psi_{0,\text{fixed}} - Y^0\right)\right\}^2 \mid Y^1, Y^0, X, S=1\right] \\ &\leq 2\mathbb{E}\left[\left(\psi_{1,\text{fixed}} - Y^1\right)^2 + \left(\psi_{0,\text{fixed}} - Y^0\right)^2 \mid Y^1, Y^0, X, S=1\right]\end{aligned}$$

where the third inequality follows again from the conditional unbiasedness of the pseudo-outcomes and the final inequality from that $(a-b)^2 \leq 2(a^2+b^2)$ for any real numbers a

and b. At last, we note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\psi_{1,\text{fixed}} - Y^1\right)^2 \mid Y^1, Y^0, X, S=1\right] = \\
&= \mathbb{E}\left[\left(\frac{A}{e(X)}(Y - h_1(X)) + h_1(X) - Y^1\right)^2 \mid Y^1, Y^0, X, S=1\right] \\
&= \mathbb{E}\left[\left(\frac{A}{e(X)}(Y^1 - h_1(X)) + h_1(X) - Y^1\right)^2 \mid Y^1, Y^0, X, S=1\right] \\
&= \mathbb{E}\left[\left(\frac{A}{e(X)} - 1\right)^2 \mid X, S=1\right](Y^1 - h_1(X))^2 \\
&= \frac{1-e(X)}{e(X)}(Y^1 - h_1(X))^2
\end{aligned}$$

where the second equality follows from consistency (Condition 6.1) and the third equality from conditional exchangeability in the trial population (Condition 6.2). Similarly, we have that

$$\mathbb{E}\left[\left(\psi_{0,\text{fixed}} - Y^0\right)^2 \mid Y^1, Y^0, X, S=1\right] = \frac{e(X)}{1-e(X)}(Y^0 - h_0(X))^2 .$$

At last, plugging the above expressions back into our original expression for $\text{Var}(\psi_{\text{fixed}} \mid S=1)$, we obtain the inequality

$$\begin{aligned}
\text{Var}(\epsilon \mid S=1) \leq \tilde{C}\Big\{&2\mathbb{E}\left[\frac{1-e(X)}{e(X)}(Y^1 - h_1(X))^2 \mid S=1\right] + 2\mathbb{E}\left[\frac{e(X)}{1-e(X)}(Y^0 - h_0(X))^2 \mid S=1\right] \\
&+ \text{Var}\left(Y^1 - Y^0 \mid X, S=1\right) - \text{Var}\left(\tau(X) \mid S=1\right)\Big\} \\
= \tilde{C}\Big\{&2\mathbb{E}\left[\frac{1-e(X)}{e(X)}(Y - h_1(X))^2 \mid X, A=1, S=1\right] \\
&+ 2\mathbb{E}\left[\frac{e(X)}{1-e(X)}(Y - h_0(X))^2 \mid X, A=0, S=1\right] \\
&+ \text{Var}\left(Y^1 - Y^0 \mid S=1\right) - \text{Var}\left(\tau(X) \mid S=1\right)\Big\}
\end{aligned}$$

where the equality follows from consistency and conditional exchangeability in the trial population again (Condition 6.1 and 6.2). □

6.A.4. Proof of Theorem 6.2

Proof. We begin by writing

$$\begin{aligned}
L_a(h) &= \mathbb{E}\left[\left\{\frac{1-e(X)}{e(X)}\right\}^{2a-1}(Y-h(X))^2 \mid A=a, S=1\right] \\
&= \mathbb{E}\left[\frac{S}{\Pr(S=1\mid A=a)}\left\{\frac{1-e(X)}{e(X)}\right\}^{2a-1}(Y-h(X))^2 \mid A=a\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{S}{\Pr(S=1\mid A=a)}\left\{\frac{1-e(X)}{e(X)}\right\}^{2a-1}(Y-h(X))^2 \mid X, A=a\right] \mid A=a\right] \\
&= \mathbb{E}\left[\frac{1}{\Pr(S=1\mid A=a)}\left\{\frac{1-e(X)}{e(X)}\right\}^{2a-1}\mathbb{E}\left[S(Y-h(X))^2 \mid X, A=a\right] \mid A=a\right].
\end{aligned}$$

We inspect the inner conditional expectation, $\mathbb{E}\left[S(Y-h(X))^2 \mid X, A=a\right]$, and note that

$$\begin{aligned}
\mathbb{E}\left[S(Y-h(X))^2 \mid X, A=a\right] &= \mathbb{E}[S \mid X, A=a]\mathbb{E}[(Y-h(X))^2 \mid X, A=a] \\
&= \Pr(S=1 \mid X, A=a)\mathbb{E}[(Y-h(X))^2 \mid X, A=a]
\end{aligned}$$

where the first equality follows from that $Y \perp\!\!\!\perp S \mid (X, A)$ holds under Conditions 6.1-6.4. To show this, we note that the conditional independencies $Y^a \perp\!\!\!\perp A \mid (X, S=1)$ (Condition 6.2) and $Y^a \perp\!\!\!\perp A \mid (X, S=0)$ (Condition 6.3) jointly imply that $Y^a \perp\!\!\!\perp A \mid (X, S)$. Combining this conditional independence statement with $Y^a \perp\!\!\!\perp S \mid X$ from Condition 6.4, they together imply $Y^a \perp\!\!\!\perp (A, S) \mid X$. Thus, from the weak union of conditional independence, we have that $Y^a \perp\!\!\!\perp S \mid (X, A) \Rightarrow Y \perp\!\!\!\perp S \mid (X, A)$ where the final implication follows from consistency (Condition 6.1).

Combining all of the above, we finally obtain the following expression,

$$L_a(h) = \mathbb{E}\left[\frac{\Pr(S=1 \mid X, A=a)}{\Pr(S=1 \mid A=a)}\left\{\frac{1-e(X)}{e(X)}\right\}^{2a-1}(Y-h(X))^2 \mid A=a\right]$$

□

6.A.5. Proof of Lemma 6.2

Proof. The inequality $R^*(\lambda^\star) \leq \min\{R^*_{QR}, R^*_{DR}\}$ follows directly from the definition of $\lambda^\star$ as the minimizer of $R^*(\lambda)$ over the set Λ. Since both $\lambda = 0$ and $\lambda = 1$ are in this set, the minimum value $R^*(\lambda^\star)$ cannot be greater than the value at either endpoint: $R^*(\lambda^\star) \leq R^*(0) = R^*_{DR}$ and $R^*(\lambda^\star) \leq R^*(1) = R^*_{QR}$. Therefore, $R^*(\lambda^\star) \leq \min\{R^*_{DR}, R^*_{QR}\}$.

For uniqueness, we note that the risk $R^*(\lambda)$ can be expanded as

$$\begin{aligned}
R^*(\lambda) &= \mathbb{E}\left[(V+\lambda(U-V))^2 \mid S=1\right] \\
&= \mathbb{E}[(U-V)^2]\lambda^2 + 2\mathbb{E}[V(U-V)]\lambda + \mathbb{E}[V^2] \\
&= A\lambda^2 + B\lambda + C,
\end{aligned}$$

where $U = \hat{\tau}_{QR}(X) - \tau(X)$, $V = \hat{\tau}_{DR}(X) - \tau(X)$, $A = \mathbb{E}[(U-V)^2]$, $B = 2\mathbb{E}[V(U-V)]$, and $C = \mathbb{E}[V^2]$. If $A > 0$, the risk function $R^*(\lambda)$ is a strictly convex quadratic function. A strictly convex function has a unique minimum over any convex set, including the set Λ. Therefore, if $A > 0$, the minimizing weight $\lambda^\star$ is unique. □

6.A.6. Proof of Theorem 6.3

Proof. Let $\widetilde{R}(\lambda)$ be the expectation of the cross-validated loss term. Since the squared loss is bounded in $[0, B]$, $\Lambda := [0,1]$ is a compact subset of $\mathbb{R}$, the uniform law of large numbers gives

$$\sup_{\lambda \in \Lambda} \left| \widehat{R}_{CV}(\lambda) - \widetilde{R}(\lambda) \right| \;=\; o_p(1).$$

Further, by definition $\widehat{R}_{CV}(\hat{\lambda}^\star) \le \widehat{R}_{CV}(\lambda^\star)$, hence

$$\widetilde{R}(\hat{\lambda}^\star) = \widehat{R}_{CV}(\hat{\lambda}^\star) + o_p(1) \;\le\; \widehat{R}_{CV}(\lambda^\star) + o_p(1) = \widetilde{R}(\lambda^\star) + o_p(1).$$

Finally, it follows from the proof in Lemma 6.1 that $\widetilde{R}(\lambda) = R^*(\lambda) + C$ for some constant C

$$R^*(\hat{\lambda}^\star) + C \;\le\; R^*(\lambda^\star) + C + o_p(1) \quad \Longrightarrow \quad R^*(\hat{\lambda}^\star) \;\le\; R^*(\lambda^\star) + o_p(1).$$

Finally, Lemma 6.2 yields

$$R^*(\lambda^\star) \;\le\; \min\{R^*_{DR}, R^*_{QR}\},$$

completing the chain:

$$R^*(\hat{\lambda}^\star) \;\le\; R^*(\lambda^\star) \;+\; o_p(1) \;\le\; \min\{R^*_{DR}, R^*_{QR}\} \;+\; o_p(1).$$

□

6.B. Connections to R-learner

It can be shown that the DR-learner (Kennedy, 2023) and the R-learner (Nie & Wager, 2021) both minimize two closely related loss functions when estimating the CATE. Below, we outline this connection and discuss its implications for connections between our proposed QR-learner and the R-learner.

Morzywolek *et al.* (2023) introduced a general formulation of a loss function,

$$R_o(\tilde{\tau}; \eta, \lambda) = \frac{1}{\mathbb{E}[\lambda\{e(X)\}]} \mathbb{E}\left[\rho\{A, e(X); \lambda\}\{\phi(O; \eta, \lambda) - \tilde{\tau}\}^2 \mid S = 1\right] \tag{6.6}$$

which, when minimized, yield a broad class of CATE learners. Here, the function $\rho\{A, e(X); \lambda\}$ and pseudo-outcome $\phi(O; \eta, \lambda)$ are defined as

$$\begin{aligned} \rho\{A, e(X); \lambda\} &= \{A - e(X)\}\lambda'\{e(X)\} + \lambda\{e(X)\} \\ \phi(O; \eta, \lambda) &= \frac{\lambda\{e(X)\}}{\rho\{A, e(X)\}} \frac{A - e(X)}{e(X)(1 - e(X))} (Y - g_A) + g_1(X) - g_0(X) \end{aligned}$$

where the nuisance models are $\eta = \{e(X), g_1, g_0\}$ and $g_a = \mathbb{E}[Y \mid X, A = a, S = 1]$ for $a \in 0, 1$.

The loss $R_o(\tilde{\tau}; \eta, \lambda)$ is indexed by a weighting function $\lambda : [0, 1] \to \mathbb{R}$ and different CATE learners are obtained depending on this weighting function. For example, the DR-learner and R-learner can both be recovered as special cases with different functions λ.

If we let $\lambda\{e(X)\} \equiv c$ for some constant $c \in \mathbb{R}$, we obtain $\rho\{A, e(X); \lambda\} = c$ and $\phi(O; \eta, \lambda) = \frac{A - e(X)}{e(X)(1-e(X))}(Y - g_A) + g_1(X) - g_0(X)$ which is equivalent to the pseudo-outcome $\psi(O; \eta = \{g_1, g_0\})$ from eq. (6.3). Consequently, the loss function $R_0(\tilde{\tau}; \eta, \lambda)$ then equals the pseudo-risk $R(\tilde{\tau}; \eta = \{g_1, g_0\})$ of the DR-learner.

Meanwhile, when letting $\lambda\{e(X)\} = e(X)(1 - e(X))$ and following some algebraic steps, we arrive at another expression for the loss function $R_0(\tilde{\tau}; \eta, \lambda)$, given by

$$\frac{1}{\mathbb{E}[e(X)(1 - e(X)) \mid S = 1]} \mathbb{E}\left[\left(\{Y - q(X)\} - \{A - e(X)\} \tilde{\tau}(X) \right)^2 \mid S = 1 \right],$$

where $q(X) = \mathbb{E}[Y \mid X, S = 1] = e(X) \cdot g_1(X) + (1 - e(X)) \cdot g_0(X)$. This expression can be recognized as the loss function of the R-learner (Nie & Wager, 2021).

The above observations highlight an interesting point: the difference between the DR-learner and R-learner depends on the propensity score $e(X)$. If $e(X)$ is constant, then $\lambda\{e(X)\} = e(X)(1 - e(X))$ is also constant, and the two loss functions coincide. Meanwhile, if $e(X)$ is non-constant, the DR-learner and R-learner lead to different CATE learners.

This connection has direct implications for our proposed QR-learner. As discussed in the main paper, the QR-learner has natural links to the DR-learner, since it uses the same loss function but fits the nuisance models differently with the help of external data. When the propensity score $e(X)$ is constant, a scenario realistic in randomized trials which often have fixed treatment probabilities, the R-learner and DR-learner coincide. Hence, in these cases, we argue that the QR-learner can be viewed as minimizing an analogous loss function to either the DR-learner or the R-learner. On the other hand, if the treatment probabilities are covariate-dependent and not fixed, the QR-learner does not minimize the same loss as the R-learner, but rather that of the DR-learner.

6.C. Statistical inference with QR-learner

In this section, we discuss when statistical inference is feasible for the estimated CATE function $\hat{\tau}(x)$ produced by the QR-learner. Although a full theoretical analysis is beyond the scope of this work, we outline key conditions under which inference is expected to be valid.

Under standard regularity conditions and with the use of cross-fitting, inference on $\hat{\tau}(x)$ becomes possible when the regression of pseudo-outcomes $\widehat{\psi}$ on covariates X is performed using a low-dimensional model (Chernozhukov *et al.*, 2024, Chapter 14). This is enabled by the Neyman orthogonality of the pseudo-outcomes in eq. (6.3), which

ensures that the impact of nuisance estimation errors on $\hat{\tau}(x)$ is second-order (Foster & Syrgkanis, 2023).

Moreover, Kennedy (2023) show that if a "stable regressor" is used to estimate $\tau(x)$ in the DR-learner (we refer to their paper for the formal definition), statistical inference may also be guaranteed. Examples of such regressors include linear regression, smoothing splines, and kernel ridge regression. Importantly, this remains feasible even if the nuisance components $\hat{\eta}$ converge at slower rates than $\hat{\tau}(x)$, since their influence appears as a product of estimation errors – specifically, those from $\hat{\eta}$ and the estimated propensity score (if one were to estimate it). In randomized trials, where the true propensity score is known, we could expect this product to vanish rapidly, making the impact of nuisance estimation asymptotically negligible on the final CATE estimate $\hat{\tau}(x)$. In this case, asymptotically valid inference for $\hat{\tau}(x)$ may be warranted as long as the regressor used for the final stage regressor has inferential guarantees (Kennedy, 2023).

6.D. Pseudo-code for QR-learner

Algorithm 2: QR-learner: Quasi-optimized Randomization-aware Learner

Input: Data $\mathscr{D} = \{(X_i, S_i, A_i, Y_i)\}_{i=1}^n$, treatment propensity score $e(X)$, model classes $\mathscr{H}$ (for outcome models), $\mathscr{G}$ (for treatment participation probability) and $\mathscr{F}$ (for CATE), number of folds $K = 2$

Partition $\mathscr{D}$ into K folds $\{\mathscr{D}^k\}_{k=1}^K$, stratified by treatment A and study indicator S;
for $k = 1$ **to** K **do**
  Stage 1: Estimate nuisance models;
  for $a \in \{0, 1\}$ **do**
    Define $\mathscr{D}_a^k = \{i \in \mathscr{D}^k : A_i = a\}$;
    Estimate $\Pr(S = 1 \mid X, A = a)$ using $\mathscr{D}_a^k$ using an estimator $\hat{\pi}_a(X)$ from model class $\mathscr{G}$;
    Solve optimization problem in eq. (6.5);
  Set $\hat{\eta}^{(k)} = \{\hat{h}_1^{(k)}, \hat{h}_0^{(k)}\}$;
  Stage 2: Estimate CATE model;
  Let $\mathscr{D}^{-k} = \mathscr{D} \setminus \mathscr{D}^k$;
  Compute pseudo-outcome $\psi(O_i; \hat{\eta}^{(k)})$ for each $i \in \mathscr{D}^{-k}$ according to eq. (6.3);
  Estimate CATE on trial data by solving
  $\hat{\tau}^{(k)} = \arg\min_{\tilde{\tau} \in \mathscr{F}} \sum_{i \in \mathscr{D}^{-k}: S_i = 1} \left(\psi(O_i; \hat{\eta}^{(k)}) - \tilde{\tau}(X_i)\right)^2$;

Output: $\hat{\tau}(X) = \frac{1}{K} \sum_{k=1}^K \hat{\tau}^{(k)}(X)$

6.E. Discussion on methods from Asiaee et al.

In this section, we explain the method R-OSCAR (Robust Observational Studies for CMO-Augmented RCT) from Asiaee *et al.* (2025) as well as another variant proposed in an earlier version of their manuscript (see Asiaee *et al.* (2023)), refered to as CFACE (CounterFactual Average Covariate Effect). We start with explaining CFACE since it came out first.

CFACE The central idea of CFACE is to estimate a trial-specific CATE learner leveraging pseudo-outcomes which fits into our randomization-aware framework, but using a different procedure for estimating the outcome nuisance models. Specifically, we instead consider randomization-aware pseudo-outcomes on the restricted form:

$$\psi(O;\eta=\{m,m\}) = \frac{A-e(X)}{e(X)(1-e(X))}(Y-m(X)), \tag{6.7}$$

where the propensity score $e(X)$ is known and the nuisance models are constrained to be identical, $h_1 = h_0 = m$. Then, Asiaee *et al.* (2023) show that m can be chosen to minimize the conditional variance,

$$\begin{aligned} m^*(x) &= \arg\min_m \mathrm{Var}\left(\psi(O;\eta=\{m,m\}) \mid X=x, S=1\right) \\ &= e(x)\cdot\mu_0^{S=1}(x) + (1-e(x))\cdot\mu_1^{S=1}(x), \end{aligned}$$

with $\mu_a^{S=1}(x) = \mathbb{E}[Y \mid X=x, A=a, S=1]$. Under transportability (Condition 4), $\mu_a^{S=1}(x)$ also equals $\mu_a^{S=0}(x) = \mathbb{E}[Y \mid X=x, A=a, S=0]$ and can thus be estimated from external data. Thus, we can fit $\mu_a^{S=0}(x)$ on only the external data for $a \in \{0,1\}$ and then compute pseudo-outcomes using m^* following a two-stage procedure similar to the DR- and QR-learners. However, in their updated manuscript, the authors acknowledge that this procedure is possible but caution that it may fail when trial and external populations are misaligned (i.e., when Condition 4 does not hold). For this reason, they now recommend against this and instead focus on a new CATE learner, called R-OSCAR.

R-OSCAR R-OSCAR departs from CFACE by proposing to fit $\mu_a^{S=0}(x)$ on external data and correct for possible transportability violations by modeling the difference

$$\delta_a(x) = \mathbb{E}[Y \mid X=x, A=a, S=1] - \mathbb{E}[Y \mid X=x, A=a, S=0]. \tag{6.8}$$

Since their focus is on the CATE itself, they further propose modeling the direct discrepancy in the CATE between the populations:

$$\delta(x) = \mathbb{E}[Y^1 - Y^0 \mid X=x, S=1] - \mathbb{E}[Y^1 - Y^0 \mid X=x, S=0]. \tag{6.9}$$

Using the optimality of m^*, computed using estimates of $\mu_a^{S=0}$ and δ_a, and the unbiasedness of $\psi(O;\eta=\{m^*,m^*\})$ for the true CATE, they derive a procedure to also estimate $\delta(x)$. Their final CATE learner, R-OSCAR, is then

$$\hat{\tau}_{\text{R-OSCAR}}(x) = (\hat{\mu}_1^{S=0}(x) + \hat{\delta}_1(x)) - (\hat{\mu}_0^{S=0}(x) + \hat{\delta}_0(x)) + \hat{\delta}(x), \tag{6.10}$$

where $\hat{\mu}_a^{S=0}$, $\hat{\delta}_a$, and $\hat{\delta}$ are estimated using trial and external data. Modeling the differences δ_a and δ resembles the additive bias correction of Kallus *et al.* (2018) more than our approach. In fact, R-OSCAR could be seen as a T-learner with additive bias correction, and diverges substantially from the two-stage procedures of the DR- and QR-learners. By contrast, the implementation of CFACE follows a two-stage design and aligns more closely with our framework.

6.F. Experimental details

6.F.1. Data-generating process in simulation studies

We simulate data as follows: We set $S_i = 1$ for $i = 1, \dots, n_1$, and for $S_i = 0$ for $i = n_1 + 1, \dots, n_1 + n_0$, and sampled a Normal d-dimensional covariate according to $X_i \sim N(\mu_{S_i}, \frac{1}{\sqrt{d}}\Sigma)$ with the mean $\mu_1 = \mathbf{0}$ or $\mu_0 = 0.2 \cdot \mathbf{1}$ and the covariance matrix Σ of shape $d \times d$ had its diagonal elements set to 1 and its off-diagonal elements set to 0.1. Thereafter, we sampled the treatment $T_i \sim \text{Bern}(e(X_i, S_i))$ according to the Bernoulli probability

$$e(X_i, S_i) = \begin{cases} 0.5, & \text{if } S_i = 1 \\ \frac{1}{1+\exp\{-(\alpha_0 + \alpha^\top X_i)\}}, & \text{otherwise} \end{cases}$$

Finally, we computed outcomes $Y_i = b(X_i) + A_i \cdot \tau(X_i) + \varepsilon_i$ where the noise variables was sampled according to $\varepsilon_i \sim N(0, \sigma^2 = 1/4)$.

For the experiment in Section 6.5.1, we modeled a highly non-linear baseline risk together with a linear CATE, which as done by defining:

$$b(X_i) = \sum_{j=1}^{d} \frac{3}{d} \cos\left(\frac{3}{2} X_{ij}\right) + \sum_{j=1}^{d} \sum_{j'=1}^{d} \frac{1}{d} X_{ij} X_{ij'},$$

$$\tau(X_i) = \sum_{j=1}^{d} \frac{1}{d} X_{ij}.$$

In the scenario where Conditions 6.3 and 6.4 held, we set the covariate dimension to $d = 5$. To simulate violations of these assumptions, we increased the covariate dimension to $d = 7$ but masked the last two dimensions, so that only 5 covariates remained observed.

For the experiment in Section 6.5.2, we considered a setting with a linear baseline outcome and a sparse linear CATE that depended only on a single covariate. To violate transportability (Condition 6.4), this time we encoded different functions $\tau(X)$ for the trial and external population. Specifically, we defined:

$$b(X_i) = \sum_{j=1}^{d} \frac{1}{d} X_{ij},$$

$$\tau(X_i) = \begin{cases} \beta \cdot d \cdot X_{i1}, & \text{if } S_i = 1 \\ \left(\beta + \frac{1}{20}\right) \cdot d \cdot X_{i1}, & \text{otherwise} \end{cases}$$

where β was a tunable parameter that controlled the size of the interaction effect between the treatment A_i and the first covariate X_{i1}. As before, we set $d = 5$.

6.F.2. Implementation details for CATE learners

For the experiments in Section 6.5.1 and 6.6, we implemented the CATE learners as follows.

For the estimators used inside the CATE learners we used implementations from the *scikit-learn* Python package (Pedregosa *et al.*, 2011) . For the DR-learner, T-learner, pooled T-learner, CFACE, and QR-learner, we used histogram-based gradient boosting regression tree using default hyperparameters. As the final CATE regressor in the two-stage CATE learners (all of the above except the T-learner variants), we used a linear regression model. For estimating $\pi_a(X) = \Pr(S = 1 \mid X, A = a)$, we used a cross-validated logistic regression with ridge penalty. For KSP, we fitted the DR-learner on the external dataset and then used a linear regression to estimate the bias model. For R-OSCAR, we also used histogram-based gradient boosting regression for the outcome nuisance models and, similar to KSP, a cross-validated logistic regression with ridge penalty for estimating both its bias functions. We applied cross-fitting to all two-stage CATE learners using two folds consistently. For cross-fold validation in the combined learner, we used three folds.

To predict with the average treatment effect (ATE), we used the difference-in-means estimate

$$\hat{\tau}_{DM} = \frac{\sum_{i=1}^{n_1} A_i Y_i}{\sum_{i=1}^{n_1} A_i} - \frac{\sum_{i=1}^{n_1} (1 - A_i) Y_i}{\sum_{i=1}^{n_1} 1 - A_i}$$

as a constant CATE prediction $\hat{\tau}(x) = \hat{\tau}_{DM}$ for all x.

For the case study Tennessee STAR dataset, we made a few changes to the implementation of all learners. First, due to having a large number of features relative to the sample size, we use ridge-penalized linear regression for the CATE model, and further changed from the default hyperparameters of the histogram-based gradient boosting regression to a max depth of 3 for the decision trees (default is unconstrained depth) and a minimum sample size per leaf of 5 (default is 20). Finally, we used 10 folds for the cross-validation in the combined learner, to prevent the training splits from becoming very small in size.

6.F.3. Statistical tests for treatment effect heterogeneity

Below we describe our implementations for the statistical tests used to detect treatment effect modification in the experiment in Section 6.5.2.

For covariate adjustment, we fit a linear regression model of Y on the covariates $(A, X_1, A \cdot X_1)$ using the trial data. We then estimated 95% confidence intervals for the coefficient of $A \cdot X_1$ to assess whether it was significantly different from zero. For the pooled covariate

adjustment, we followed the same approach but fit the linear regression model on the combined dataset consisting of both the trial data and the external data.

For the DR-learner, QR-learner, and CFACE, we split the data into two folds (stratified by the treatment A and the study population S). We used the first fold to estimate the nuisance components in η via linear regression using their respective strategies as outlined in the main paper, then computed pseudo-outcomes on the second fold. These pseudo-outcomes were regressed on the feature X_1 using another linear regression model, resulting in the first CATE estimate $\hat{\tau}(X_1) = \hat{\alpha}^{(1)} \cdot X_1$. To utilize the entire dataset efficiently, we repeated the process by swapping the folds, obtaining a second CATE estimate $\hat{\tau}(X_1) = \hat{\alpha}^{(2)} \cdot X_1$. We then computed 95% confidence intervals to test the null hypothesis that $\alpha = 0$ as follows:

$$[\bar{\alpha} - 1.96 \cdot \text{se}(\bar{\alpha}),\ \bar{\alpha} + 1.96 \cdot \text{se}(\bar{\alpha})]\,,$$

where $\bar{\alpha} = \frac{1}{2}(\hat{\alpha}^{(1)} + \hat{\alpha}^{(2)})$ and using the individual standard errors $\text{se}(\hat{\alpha}^{(1)})$ and $\text{se}(\hat{\alpha}^{(2)})$ to compute

$$\text{se}(\bar{\alpha}) = \sqrt{\frac{1}{4}\left(\text{se}(\hat{\alpha}^{(1)})^2 + \text{se}(\hat{\alpha}^{(2)})^2\right)},$$

assuming normality and that the covariance between $\hat{\alpha}^{(1)}$ and $\hat{\alpha}^{(2)}$ to be negligible.

6.F.4. Tennessee STAR dataset

The data from the STAR study consists of 4218 students: 2811 from rural schools and 1407 from urban schools. The treatment is class size: small ($A = 0$) versus regular ($A = 1$), and the continuous outcome Y is the student's average test score. Our observed covariates include gender, race, birth date, teacher ID, and free lunch eligibility. After applying one-hot encoding to these variables, which are either binary or categorical, we obtain a 310-dimensional feature vector.

To construct two datasets from different populations, we split the dataset into rural and urban school. One school location is assigned as the target trial population and the other as the external population. In the experiment, we sample without replacement half of the target trial population as a held-out test set, and then sample without replacement n_1 and n_0 observations from the target trial and external populations, respectively. We kept $n_0 = 1000$ fixed while varying n_1 from 100 to 700 (about half of the number of urban schools).

We further evaluate whether transportability (Condition 6.4) is violated between the two populations in this dataset. As shown in Figure 6.3a, we can see that the distribution of outcomes (average test scores) differs between rural and urban schools. This may indicate that transportability also is violated if the school location is omitted from the observed covariates X. We performed a conditional independence test of $Y \perp\!\!\!\perp S \mid (X, A)$, which is a testable implication for Conditions 6.1 to 6.4 jointly (Dahabreh *et al.*, 2024). We applied both a partial linear correlation test and the randomized conditional correlation test (RCoT) (Strobl *et al.*, 2019), implemented in the Python package *pybnesian* (Atienza

et al., 2022). Both tests indicate that the conditional independence is violated at the 5% significance level, suggesting that one of Conditions 6.1 to 6.4 may not hold. Since consistency (Condition 6.1) and strong ignorability in both populations (Conditions 6.2 and 6.3) are expected to hold, it is most likely that transportability (Condition 6.4) is violated.

We further used t-SNE for dimensionality reduction to visualize the distribution of covariates across the two populations. These visualizations, shown in Figure 6.3b, reveal limited overlap in certain regions between rural and urban schools.

6.F.5. Compute resources used for experiments

All experiments were run on a CPU machine. The simulations in Sections 6.5 and 6.6 each completed in under 72 hours on a laptop with a 2 GHz Quad-Core Intel Core i5 processor and 16GB of RAM. Including preliminary experiments not shown in the paper, no single run exceeded this runtime.

6.G. Additional experimental results

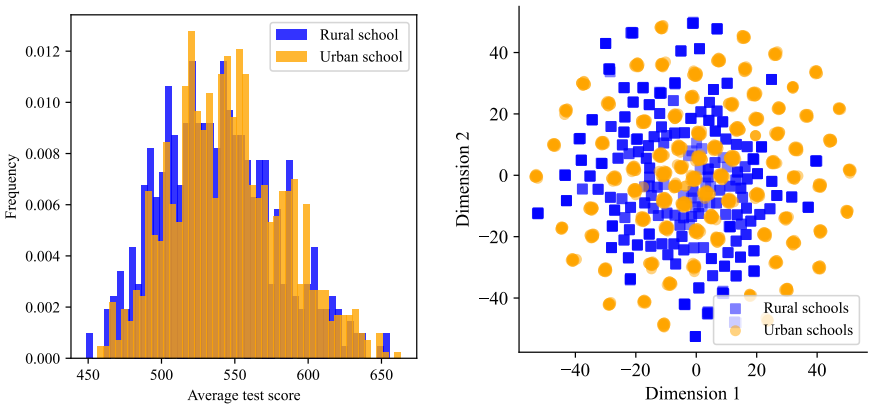
6.G.1. Analysis of prediction bias

As a complement to Table 6.1, which reports the average root mean squared error of each method in our simulation study, Table 6.2 presents the corresponding average prediction bias, defined as $\text{bias}(\hat{\tau}) = \mathbb{E}[\hat{\tau}(X) - \tau(X) \mid S = 1]$. We find that all methods, except the pooled T-learner, were largely unbiased. This aligns with expectations, as we fitted linear models for the CATE which matches with the true underlying linear CATE function. The pooled T-learner, however, exhibited increasing bias when conditions 3 and 4 were violated, which is consistent with its reliance on these conditions for identification of the CATE.

6.G.2. Tennessee STAR dataset

We include both settings from the Tennessee STAR dataset case study. In the first setting, presented in the main paper, we treat the urban schools as the target trial population and the rural schools as the external population. We then flipped the populations, treating the rural schools as the target trial population. Overall, we observe similar trends in both cases, but we highlight below the differences that arise when flipping the two populations. The results from both settings are shown in Figure 6.4.

First, we observe that the trial-only and pooled T-learner improve faster when the urban schools are the target population compared to when the rural schools are the target. Second, CFACE and R-OSCAR perform better when the rural schools are the target population. Finally, the QR-learner shows little improvement over the DR-learner when the urban



(a) (b)

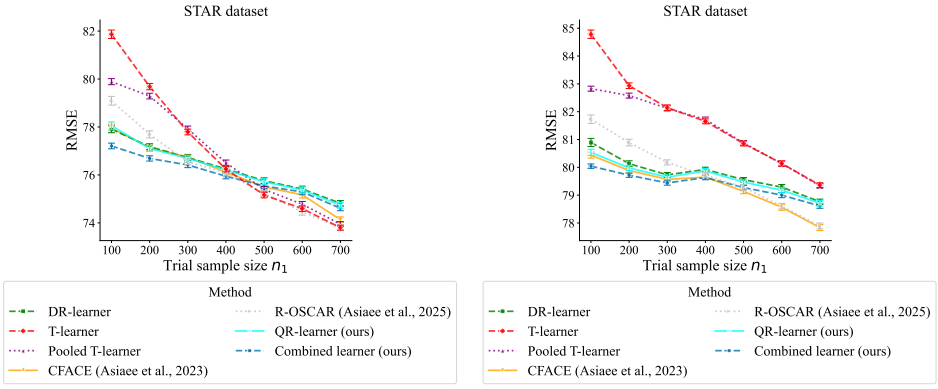
Figure 6.3: **(a)**: Distributions of the outcome (average test scores) for rural and urban schools in the STAR dataset. We observe a slight shift in the mean between the two groups, suggesting potential violations of transportability, as the primary difference between the trial and the external population lies in school location. **(b)**: t-SNE plot over features colored by study population. We observe some lack of overlap between populations.

Table 6.2: Average prediction bias reported over 500 repeated runs from the simulation study with a trial sample size $n_1 = 250$ under different scenarios.

External sample size	100		1000		10000	
Condition 3 and 4 violated?	No	Yes	No	Yes	No	Yes
DR-learner	-0.00 (5e-03)	-0.00 (5e-03)	-0.00 (5e-03)	-0.01 (5e-03)	0.00 (4e-03)	0.01 (5e-03)
T-learner	0.01 (4e-03)	0.01 (5e-03)	0.01 (5e-03)	0.00 (5e-03)	0.01 (4e-03)	0.00 (5e-03)
Pooled T-learner	0.06 (3e-03)	0.12 (4e-03)	0.05 (2e-03)	0.20 (3e-03)	-0.03 (1e-03)	0.11 (1e-03)
CFACE (Asiaee *et al.*, 2023)	0.01 (5e-03)	0.00 (5e-03)	0.01 (4e-03)	-0.00 (5e-03)	0.00 (3e-03)	0.00 (4e-03)
R-OSCAR (Asiaee *et al.*, 2025)	0.00 (6e-03)	0.00 (5e-03)	0.00 (4e-03)	0.01 (5e-03)	-0.00 (3e-03)	0.00 (4e-03)
KSP (Kallus *et al.*, 2018)	0.01 (5e-03)	-0.01 (5e-03)	-0.01 (6e-03)	-0.01 (6e-03)	-0.01 (5e-03)	0.01 (6e-03)
QR-learner (ours)	0.01 (5e-03)	-0.00 (5e-03)	0.01 (4e-03)	-0.00 (4e-03)	0.00 (3e-03)	0.00 (4e-03)
Combined learner (ours)	0.00 (5e-03)	0.00 (5e-03)	0.00 (4e-03)	0.00 (4e-03)	-0.00 (3e-03)	0.00 (4e-03)

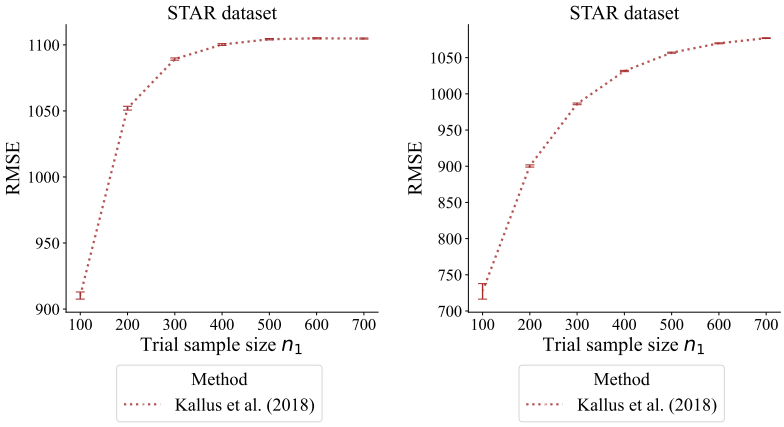
schools are the target, but it exhibits noticeable improvement when the rural schools are the target.

We also include the performance of KSP, the additive bias correction method of Kallus *et al.* (2018), in Figure 6.5. This method exhibits significantly higher RMSE than the other methods and, unusually, its performance worsens as the number of trial samples increases. While we cannot fully explain this behavior, one possible reason is that, unlike the other learners, this approach is not specifically designed for estimating a trial-specific CATE.



(a) Urban school as target population (b) Rural schools as target population

Figure 6.4: We evaluate the RMSE on the STAR dataset when increasing the trial sample size with a fixed external sample size of $n_0 = 1000$. We report the average RMSE and standard error over 200 repeated runs.



(a) Urban school as target population (b) Rural schools as target population

Figure 6.5: Separate plot for additive bias correction method of Kallus *et al.* (2018) because its large RMSE values make it difficult to display alongside the other methods. We evaluate the RMSE on the STAR dataset when increasing the trial sample size with a fixed external sample size of $n_0 = 1000$. We report the average RMSE and standard error over 200 repeated runs.

Part Three

Policy Evaluation

7

Qini Curve Estimation Under Clustered Network Interference

Qini curves are a widely used tool for assessing treatment policies under allocation constraints as they visualize the incremental gain of a new treatment policy versus the cost of its implementation. Standard Qini curve estimation assumes no interference between units: that is, that treating one unit does not influence the outcome of any other unit. In many real-life applications such as public policy or marketing, however, the presence of interference is common. Ignoring interference in these scenarios can lead to systematically biased Qini curves that over- or under-estimate a treatment policy's cost-effectiveness. In this paper, we address the problem of Qini curve estimation under clustered network interference, where interfering units form independent clusters. We propose a formal description of the problem setting with an experimental study design under which we can account for clustered network interference. Within this framework, we describe three estimation strategies, each suited to different conditions, and provide guidance for selecting the most appropriate approach by highlighting the inherent bias-variance trade-offs. To complement our theoretical analysis, we introduce a marketplace simulator that replicates clustered network interference in a typical e-commerce environment, allowing us to evaluate and compare the proposed strategies in practice.

7.1. Introduction

Understanding treatment effect heterogeneity – the variation in individual responses to the same treatment within a population – is central in shaping individualized treatment

This chapter appears as: Karlsson, R., Akker, B. v. d., Moraes, F., Proença, H. M., & Krijthe, J. H. (2025). Qini curve estimation under clustered network interference. *arXiv preprint arXiv:2502.20097*. RK and BvdA contributed equally to this work.

policies across various domains, including personalized medicine (Kravitz *et al.*, 2004), uplift modeling in marketing and e-commerce (Goldenberg *et al.*, 2020), and targeted subgroup interventions in public policy (Brand & Davis, 2011). In these scenarios, the same questions recurs: *Who should we treat?* Sometimes, it is sufficient to identify individuals who respond positively to a treatment. However, when treatments involve monetary or practical costs, the challenge is to devise a cost-effective policy that targets those who benefit the most from the treatment while staying within a given budget for treatment allocation.

First introduced by Radcliffe (2007) in the marketing literature, Qini curves have become a widely used method for evaluating the cost-effectiveness of treatment policies. A Qini curve plots the incremental gain by treating units prioritized by a given treatment rule under varying allocation budgets. By comparing the Qini curves of different prioritization rules, practitioners can determine which rule most effectively identifies who responds well to treatment. However, reliable estimation of Qini curves depends on some key assumptions being met, one of which is the Stable Unit Treatment Value Assumption (Rubin, 1980). This assumption implies that there is no treatment interference, meaning that treating one unit has no influence on the outcome of any other unit.

Interference arises in a variety of contexts, from peer effects in social networks (Manski, 2013; Ogburn *et al.*, 2020) to cannibalization effects on marketplace platforms (Holtz *et al.*, 2025). One of the most common settings is so-called *clustered network interference* where interfering units form independent clusters; here units interfere within, rather than between, clusters. While there exists an extensive body of literature on estimating treatment effects under clustered network interference, e.g. Hudgens and Halloran (2008) and Sobel (2006), little attention has been given to the problem of estimating Qini curves in this setting. As we demonstrate in Figure 7.1, traditional methods for estimating Qini curves become significantly biased when interference is present. Since biased Qini curves lead to incorrect assessments of the cost-effectiveness of treatment policies, this is an important yet unaddressed problem. For example, Imai and Li (2023) emphasize the need for methods that evaluate individualized treatment strategies under interference. Motivated by this gap, the central question we aim to answer in this paper is: *How can we account for interference when estimating Qini curves, and how does this adjustment affect decision-making using Qini curves?*

Contributions To address our research question, we first formulate the experimental study design and necessary identification conditions for estimating Qini curves under clustered network interference. Next, we describe three separate estimation strategies based on different assumptions of the underlying interference. Our theoretical analysis demonstrates that stronger assumptions on the interference can yield more efficient estimators, though at the potential cost of increased bias when those assumptions are violated. We explore these trade-offs by empirically comparing all methods using a simulated data-generating process designed to mimic a marketplace with different types of interference in the form of cannibalization among different vendors. Finally, based on our findings, we offer practical recommendations for estimating Qini curves in settings

with clustered network interference.

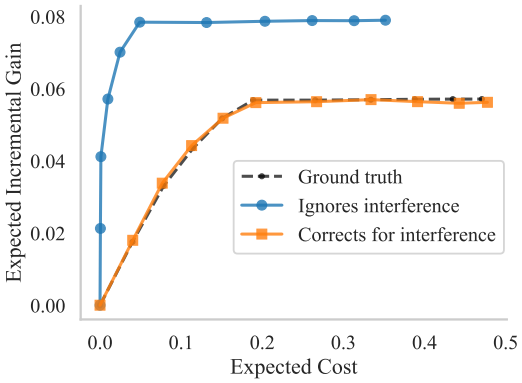


Figure 7.1: An illustrative simulation study with Qini curve estimation under clustered network interference. The black dashed line represents the true underlying Qini curve, while the solid lines depict two estimation approaches: one based on a traditional method that assumes no interference, and the other representing a proposed strategy in this paper that adjusts for interference using inverse probability weighting. More details on the simulation used to generate this figure can be found in Appendix 7.D.

7.2. Related works

The task of estimating treatment effects becomes considerably more complex in the presence of interference. Hence, despite early influential works in causal inference such as Rubin (1974), only recently has a substantial body of literature emerged to address interference. One of the most commonly studied settings is clustered network interference, where treatment units form independent clusters (Hudgens & Halloran, 2008; Sobel, 2006; Tchetgen & VanderWeele, 2012), a condition also known as partial interference. Interference also naturally arises in network data, where some units are related to other units by being neighbors in e.g. a social network (Eckles *et al.*, 2017; Ugander *et al.*, 2013). In some cases, an experimental study design can be constructed in a way to detect and reduce bias from interference, for instance, through a two-stage randomization design (Hudgens & Halloran, 2008) or by stratified randomization across different blocks (Bajari *et al.*, 2023). Prior work has tackled specific tasks under interference, such as heterogeneous treatment effect estimation (Zhao *et al.*, 2024) and policy evaluation/learning (Y. Zhang & Imai, 2023). To our knowledge, however, no prior work has considered the problem of estimating Qini curves in the presence of interference.

Evaluating treatment prioritization rules using Qini curves in settings without interference has gained more attention in recent years (Bokelmann & Lessmann, 2024; Radcliffe, 2007; Rößler & Schoder, 2022). The development of estimations procedures with better

statistical inference guarantees has enabled the use of Qini curves in this context (Yadlowsky *et al.*, 2024). While none of these works consider interference, Sverdrup *et al.* (2025) considers the related problem of estimating Qini curves for combinatorial multi-armed treatments.

Combinatorial treatment problems and clustered network interference are inherently connected, as assigning treatments within a cluster can be seen as a combinatorial treatment decision. While causal inference methods exist for such settings, see e.g. Dasgupta *et al.* (2014) and Goplerud *et al.* (2025), they typically assume a fixed treatment dimension – an assumption too restrictive when cluster sizes also may vary. However, the connection between these high-dimensional treatment settings and our setting underscores the challenge of estimating Qini curves under clustered network interference: as cluster size grows the combinatorial space of possible treatments expands exponentially, leading to a corresponding increase in interactions among units within the cluster. To address this challenge, we propose estimation strategies designed to more accurately estimate Qini curves, even as the cluster size differs or grows.

7.3. Data structure & assumptions

Notation We assume access to observations from a distribution P. We have clusters $i = 1, \dots, N$ and each cluster contains the units $j = 1, \dots, M_i$. A unit can be referred to by the tuple (i, j). For each cluster i, we observe pre-treatment covariates X_i in $\mathscr{X} \subseteq \mathbb{R}^{d_x}$. For each unit, we observe pre-treatment covariates Z_{ij} in $\mathscr{Z} \subseteq \mathbb{R}^{d_z}$, a binary treatment $W_{ij} \in \{0, 1\}$, and an outcome of interest Y_{ij} in $\mathscr{Y} \subseteq \mathbb{R}$. The outcome may be binary or continuous. In addition, we also observe a non-negative cost of treatment C_{ij} in $\mathscr{C} \subseteq [0, \infty)$. The cost C_{ij} depends on both the treatment and outcome, and specifically we assume there to be no cost $C_{ij} = 0$ when no treatment is given $W_{ij} = 0$. We consider cluster-level outcomes and costs which we define as $\tilde{Y}_i = \sum_{j=1}^{M_i} Y_{ij}$ and $\tilde{C}_i = \sum_{j=1}^{M_i} C_{ij}$. We also define the cluster-level treatment which is a binary vector $\mathbf{W}_i = [W_{i1}, W_{i2}, \dots, W_{iM_i}] \in \{0, 1\}^{M_i}$. At last, random variables are denoted by capital letters, while their instantiated values use lowercase. Probability densities are represented as $f(\cdot)$.

Clustered network interference In the setting of clustered network interference, we assume observations can be divided in independent clusters. Treating one unit belonging to cluster i may influence the outcomes of other units from that same cluster. However, treating units from a different cluster i' will not influence the outcomes of the units in cluster i. To define causal effects in this setting, we posit potential (counterfactual) outcomes $Y_{ij}(\mathbf{w})$ corresponding to the outcomes we would observe for an unit (i, j) if the treatment vector $\mathbf{W}_i$ would be set to $\mathbf{w}$ (Tchetgen & VanderWeele, 2012). Analogously, we define the counterfactual cost $C_{ij}(\mathbf{w})$ if $\mathbf{W}_i$ would be set to $\mathbf{w}$.

Study design Throughout this paper, we consider an experimental study design where the unit-level treatments W_{ij} are independently and randomly assigned. The treatment probability is determined by $e_w(x) = \Pr(W = w \mid X = x)$ which is known and the same for all units within a cluster. Following standard convention, we refer to this probability as the propensity score (Rosenbaum & Rubin, 1983b). We assume the following conditions are fulfilled by our experimental study design.

Assumption 7.1. *Consistency: if* $\mathbf{W}_i = \mathbf{w}$ *then* $Y_{ij}(\mathbf{W}_i) = Y_{ij}$ *and* $C_{ij}(\mathbf{w}) = C_{ij}$*, for all units* (i, j) *and treatments* $\mathbf{w} \in \{0,1\}^{M_i}$*. Conditional exchangeability: for each treatment* $\mathbf{w} \in \{0,1\}^{M_i}$*,* $\big(Y_{ij}(\mathbf{w}), C_{ij}(\mathbf{w})\big) \perp\!\!\!\perp \mathbf{W}_i \mid X_i$*. Positivity: for each treatment* $\mathbf{w} \in \{0,1\}^{M_i}$*, if* $f(x) \neq 0$ *then* $\Pr(\mathbf{W}_i = \mathbf{w} \mid X_i = x) > 0$.

Consistency is met when the intervention is unambiguously defined, meaning that no undisclosed variants of the treatment exist. *Conditional exchangeability* corresponds to assuming no unmeasured confounding; specifically, the characteristics captured by cluster-level covariates are sufficient to control for any confounding between treatment assignment and outcome/cost. *Positivity* necessitates that all clusters have a non-zero probability of receiving any of combination of available treatments among its units. In the context of our experimental study design, we emphasize that conditional exchangeability and positivity can be guaranteed by (conditional) randomization.

7.4. Assessing treatment policies using Qini curves under clustered network interference

We are interested in assessing treatment policies based on some treatment prioritization rule $S : \mathscr{X} \times \mathscr{Z} \to \mathbb{R}$ that attempts to rank all units across the clusters based on who responds best to the treatment. A larger $S(X_i, Z_{ij})$ should here be interpreted as that the unit (i, j) is expected to have a larger treatment effect. Given a treatment prioritization rule S and a fixed treatment threshold $R \in \mathbb{R}$, we will evaluate decisions by treatment policies defined as

$$\pi_{S,R}(x, z) = \begin{cases} 1, & S(x, z) \geq R \\ 0, & S(x, z) < R \end{cases}. \tag{7.1}$$

Importantly, throughout this paper, we will assume that all treatment prioritization rules are derived independently of the data we will use to evaluate them on. For instance, S could be the estimated model of the conditional average treatment effect, see e.g. Künzel *et al.* (2019), trained on a separate dataset or a formal prioritization rule developed by experts using domain knowledge. To simplify notation, we will omit the subscripts in $\pi_{S,R}$ when possible and simply write π to denote a policy.

For clarity and consistency with prior work on Qini curve estimation, we present scoring rules that depend only on unit- and cluster-level covariates, ignoring interference effects. In principle, however, scoring rules could incorporate treatments and covariates of other

units within the same cluster. Notably, the identifiability results at the end of this section remain valid under such generalized scoring rules, though policies based on them may be difficult to implement in practice.

7.4.1. Definition of the Qini curve

A common approach to assess how well a treatment prioritization rule identifies those who best respond to a treatment is by using Qini curves (Radcliffe, 2007; Sverdrup *et al.*, 2025). We denote the decisions made by the policy π on an evaluation dataset as $\pi_{ij} = \pi(X_i, Z_{ij}) \in \{0,1\}$ for unit (i,j) and $\boldsymbol{\pi}_i = [\pi(X_i, Z_{i1}), \ldots, \pi(X_i, Z_{iM_i})] \in \{0,1\}^{M_i}$ for the collective treatment decision on cluster i. Then, we first define the policy value in terms of average cluster-level outcomes under decisions made by the policy π as

$$V(\pi) = \mathbb{E}\left[\sum_{j=1}^{M_i} Y_{ij}(\boldsymbol{\pi}_i)\right], \tag{7.2}$$

and the policy cost of applying π as

$$C(\pi) = \mathbb{E}\left[\sum_{j=1}^{M_i} C_{ij}(\boldsymbol{\pi}_i)\right]. \tag{7.3}$$

Let R_B denote the treatment threshold such that $C(\pi_{S,R_B}) = B$, then we can define the Qini curve for a treatment prioritization rule S as follows:

$$Q_S(B) = V(\pi_{S,R_B}) - V(\pi_0), \;\; B \in [0, B_{\max}], \tag{7.4}$$

where $\pi_0 \equiv 0$ is a reference policy that treats none and $B_{\max} > 0$ is the maximal allowed cost by the treatment strategy under consideration.

The above definition is more general than the one by Radcliffe (2007), who assumes uniform cost across all units. In their approach, one plots $Q_S(B)$ on the y-axis against the fraction of treated units on the x-axis. The definition presented here can be applied to this case by plotting $B/B_{\max}$ on the x-axis, where $B_{\max}$ represents the total cost of treating all units. We refer to this as the uniform cost case.

To understand how interference introduces challenges in the estimation of Qini curves, consider a scenario where two units from the same cluster fall on opposite sides of the treatment threshold R – one above (indicating it should be treated) and one below (indicating it should not). Normally, these units would be considered independent, but in the presence of interference, spillover effects may occur between them. If these effects are not accounted for, we might over- or underestimate the policy value and cost which also affects the Qini curve estimation.

For the remainder of this paper, we will demonstrate how to address this issue and provide a methodology for estimating Qini curves that appropriately accounts for interference within a clustered network setting. However, before we propose multiple estimation strategies, we also establish a necessary identifiability result that enables the estimation.

7.4.2. Identifiability of policy value and policy cost

To estimate $Q_S(B)$ in eq. (7.4), $V(\pi)$ and $C(\pi)$ must be identifiable from the observed data. Under a study design that satisfies Assumption 7.1, this identifiability is guaranteed. More specifically, recall that $\tilde{Y}_i = \sum_{j=1}^{M_i} Y_{ij}$ and $\tilde{C}_i = \sum_{j=1}^{M_i} C_{ij}$, then we define

$$\phi(\pi) = \mathbb{E}\big[\mathbb{E}\big[\tilde{Y} \mid \mathbf{W} = \boldsymbol{\pi}, X\big]\big] \quad \text{and} \quad \psi(\pi) = \mathbb{E}\big[\mathbb{E}\big[\tilde{C} \mid \mathbf{W} = \boldsymbol{\pi}, X\big]\big]\,. \tag{7.5}$$

With these definitions, we obtain the following important identifiability result (see proof in Appendix 7.A.1).

Lemma 7.1. *Under Assumption 7.1, we have* $V(\pi) = \phi(\pi)$ *and* $C(\pi) = \psi(\pi)$.

With the established identification results, we can now outline the general procedure for Qini curve estimation in our experimental study design. In practice, to estimate the Qini curve for a treatment prioritization rule S, one typically estimates the policy value $\phi(\pi)$ and cost $\psi(\pi)$ over a range of pre-specified thresholds R. For the uniform cost case, it is sufficient to only estimate the policy value. We outline the full procedure in Algorithm 3. So far, we assumed ranking according to S leads to no ties, but if there are ties one could add tiebreakers by, e.g., injecting a small amount of random noise. To implement this algorithm we would need estimators for $\phi(\pi)$ and $\psi(\pi)$; for this reason, next we describe three estimation strategies designed for different scenarios.

7.5. Estimation strategies

In this section we consider three strategies for estimating Qini curves under clustered network inference. In particular, we will focus on weighting estimators that require the propensity score, $e_w(x) = \Pr(W = w \mid X = x)$. Since we assumed the propensity score to be known in our design, we avoid the need to estimate any other nuisance models. Doing so avoids the risk of introducing additional biases due to misspecification when having to estimate nuisance models. Although weighting estimators are the main focus of this paper, we also present in Appendix 7.B some preliminary investigations into techniques for augmenting weighted estimators by incorporating nuisance models that predict the outcome.

Throughout this section, due to the similarity of the estimands for the policy value $\phi(\pi)$ and cost $\psi(\pi)$, we only present strategies for estimating the policy value $\phi(\pi)$ which analogously can be applied for estimating the policy cost $\psi(\pi)$ as well.

Algorithm 3: Qini curve estimation

Input: Dataset $D = \{X_i, \{Z_{ij}, W_{ij}, Y_{ij}, C_{ij}\}_{j=1}^{M_i}\}_{i=1}^{N}$; treatment prioritization rule S; number of percentiles K; max budget $B_{\max}$; estimators $\widehat{\phi}$ and $\widehat{\psi}$; Boolean flag `uniformCost`

$\widehat{V}_0 \leftarrow \widehat{\phi}(\pi_0 \equiv 0)$;
$(\widehat{B}_0, \widehat{Q}_0) \leftarrow (0,0)$;
Let $S_{\text{sorted}}(i)$ be the score for the ith unit when sorted by S in descending order;
for $k \leftarrow 1$ **to** K **do**
  $i_k \leftarrow \text{round}\left(\frac{k}{K} \cdot |D|\right)$;
  $R_B \leftarrow S_{\text{sorted}}(i_k)$;
  if *`uniformCost` is true* **then**
    $\widehat{B}_k \leftarrow \frac{k}{K} \cdot B_{\max}$;
  end
  else
    $\widehat{B}_k \leftarrow \widehat{\psi}(\pi_{S,R_B})$;
  end
  $\widehat{Q}_k \leftarrow \widehat{\phi}(\pi_{S,R_B}) - \widehat{V}_0$;
end
Return: $\{(\widehat{B}_k, \widehat{Q}_k)\}_{k=0}^{K}$;

7.5.1. Cluster-level inverse probability weighting

The simplest estimator for $\phi(\pi)$ in our setting is

$$\widehat{\phi}^{\,\text{IPW}}(\pi) = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \tilde{Y}_i \tag{7.6}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. This estimator is a natural extension of the traditional inverse probability weighting (IPW) estimator (Robins *et al.*, 1994) to settings with clustered network interference, see e.g. Tchetgen and VanderWeele (2012). For this reason, we will refer to $\widehat{\phi}^{\,\text{IPW}}(\pi)$ as the standard IPW estimator. We can show the following important properties of the standard IPW estimator (see proof in Appendix 7.A.2).

Theorem 7.1. *Under Assumption 7.1, the standard IPW estimator $\widehat{\phi}^{\,IPW}(\pi)$ is an unbiased estimator for the policy value, $\mathbb{E}[\widehat{\phi}^{\,IPW}(\pi)] = V(\pi)$, and its sampling variance can be written as*

$$\text{Var}\left(\widehat{\phi}^{\,IPW}(\pi)\right) = \frac{1}{N^2}\sum_{i=1}^{N}\left\{\mathbb{E}\left[\omega(X_i)\left[\tilde{Y}_i(\boldsymbol{\pi}_i)\right]^2\right] + \text{Var}\left(\tilde{Y}_i(\boldsymbol{\pi}_i)\right)\right\} \tag{7.7}$$

$$\omega(X_i) = \left[\prod_{j=1}^{M_i}\left(\frac{e_1(X_i)e_0(X_i)}{e_{\pi_{ij}}(X_i)^2} + 1\right) - 1\right] \tag{7.8}$$

From the above theorem, we observe that while the standard IPW estimator is unbiased, its efficiency is poor which becomes evident from inspecting its sampling variance, Since $e_1(X_i) \neq 0$ and $e_0(X_i) \neq 0$ due to Assumption 7.1, the factor $\omega(X_i)$ increases exponentially with the cluster size M_i. Consequently, its variance scales exponentially with the cluster size M_i which makes it prohibitively difficult to use the standard IPW estimator for Qini curve estimation in scenarios where the cluster size M_i is large.

For this reason, we explore other weighting estimators that introduce additional conditions on the structure of the underlying interference. It is important to emphasize here that these additional conditions are not required for identification of the policy value $V(\pi)$, but invoked for more efficient estimation. As we will see, in the cases where these additional conditions do not hold, their respective estimators may introduce additional biases. This results in an inherent bias-variance trade-off for estimating Qini curves in the presence of interference.

7.5.2. Interference under a fractional exposure mapping

One strategy to deal with interference is by defining exposure mappings (Aronow & Samii, 2017). An exposure mapping is a function $d_{ij} : \{0,1\}^{M_i} \to \mathscr{D}$ between all possible treatment configurations for unit (i,j) and a representation (or, embedding) of the treatment configurations. In essence, we want to map similar treatment configurations to the same "effective treatment" (Manski, 2013). If the space $\mathscr{D}$ has smaller cardinality than the original space $\{0,1\}^{M_i}$, which has cardinality 2^{M_i}, we have the possibility for more efficient estimation. Any exposure mapping, however, must fulfill the following condition.

Assumption 7.2. *The potential outcomes of a unit* (i,j) *can be grouped by* d_{ij}*, meaning that* $d_{ij}(\mathbf{w}) = d_{ij}(\mathbf{w}')$ *implies* $Y_{ij}(\mathbf{w}) = Y_{ij}(\mathbf{w}')$ *for all* $\mathbf{w}, \mathbf{w}' \in \{0,1\}^{M_i}$.

Here, we consider Qini curve estimation using one of the most common ways to define an exposure map. Namely, assuming that the potential outcome $Y_{ij}(\mathbf{w})$ for a unit (i,j) is only a function of both its own treatment status W_{ij} and the fraction of treated units within the same cluster (Bajari *et al.*, 2023; Ugander *et al.*, 2013). This corresponds to the exposure mapping $d_{ij}(\mathbf{W}_i) = [W_{ij}, \overline{W}_i]$ where $\overline{W}_i = M_i^{-1} \sum_{j=1}^{M_i} W_{ij}$.

Denoting the fraction of treated in cluster i by policy π as $\overline{\pi}_i = M_i^{-1} \sum_{j=1}^{M_i} \pi_{ij}$, we define the fractional IPW estimator as follows:

$$\widehat{\phi}^{\text{fracIPW}}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{\mathbf{1}\left(W_{ij} = \pi_{ij}, \overline{W}_i = \overline{\pi}_i\right)}{q_{ij}(\boldsymbol{\pi}_i, X_i)} Y_{ij} \tag{7.9}$$

where $q_{ij}(\boldsymbol{\pi}_i, X_i) = \Pr(W_{ij} = \pi_{ij}, \overline{W}_i = \overline{\pi}_i \mid X_i)$. As we show in Appendix 7.A.3, the probability $q_{ij}(\boldsymbol{\pi}_i, X_i)$ can be expressed in terms of the known propensity score. We can now show that the fractional IPW estimator is unbiased in the setting that the fractional exposure mapping is correctly specified (see proof in Appendix 7.A.4).

Theorem 7.2. *Under Assumptions 7.1 and 7.2, we have that the fractional IPW estimator* $\widehat{\phi}^{fracIPW}(\pi)$ *is an unbiased estimator for the policy value,* $\mathbb{E}\left[\widehat{\phi}^{fracIPW}(\pi)\right] = V(\pi)$.

Notably, the fractional IPW estimator requires additional assumptions compared to the standard IPW estimator to be unbiased, but its variance scales more favorably with the cluster size M_i because the fractional exposure mapping reduces the cardinality of the treatment space from 2^{M_i} to $2(M_i+1)$ per cluster. This reduction makes estimation more feasible in settings with large clusters. However, this does not imply that the variance of $\widehat{\phi}^{\text{fracIPW}}$ grows linearly with M_i. The estimator remains inversely proportional to the probability $q_{ij}(\pi_i, X_i)$. As M_i increases, the number of possible treatment fractions grows, making it less likely to observe a specific fraction. Consequently, $q_{ij}(\pi_i, X_i)$ approaches zero as M_i increases, which amplifies the variance, though at a slower rate than the standard IPW estimator.

7.5.3. Interference under β-additive model

Next, we consider another strategy that can reduce variance compared to the standard IPW estimator. Specifically, we use the following polynomial model to describe the interference, as proposed by Y. Zhang and Imai (2023).

Assumption 7.3. *The potential outcome model satisfies* $\mathbb{E}[Y_{ij}(\mathbf{W}_i) \mid X_i] = \mathbf{g}_j(X_i)^\top \gamma(\mathbf{W}_i)$ *where* $\mathbf{g}_j(X_i) = [g_j^{(0)}, \ldots, g_j^{(m)}]^\top$ *is an unknown vector of functions* $g_j^{(\cdot)} : \mathcal{X} \to \mathbb{R}$ *that may vary across units in the same cluster. Furthermore, we have the an augmented treatment vector* $\gamma(\mathbf{W}_i) = \left[1, \mathbf{W}_i, \mathbf{W}_i^{(2)}, \ldots, \mathbf{W}_i^{(\beta)}\right]^\top$ *with* $\mathbf{W}_i^{(k)} = \left\{\prod_{m=1}^k W_{ij_m} \;\middle|\; j_1 < \cdots < j_k\right\}$ *that contains interactions up to the order of* β *between treatment of different units in the same cluster. Here,* β *is upper bounded by the largest possible* M_i.

This assumption states that each unit's conditional mean potential outcome is a linear function of the augmented treatment vector $\gamma(\mathbf{W}_i)$, which includes interaction terms up to order β between treatments within the same cluster. We will refer to the above assumption as the β-additive assumption.

Denoting $\mathscr{I}_i^\beta$ as the power set of $\{1, \ldots, M_i\}$ with cardinality at most β, we define the β-additive IPW estimator

$$\widehat{\phi}^{\,\beta\text{-IPW}}(\pi; \beta) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{\mathscr{U} \in \mathscr{I}_i^\beta}\prod_{j \in \mathscr{U}}\left(\frac{\mathbf{1}(W_{ij} = \pi_{ij})}{e_{\pi_{ij}}(X_i)} - 1\right)\right]\tilde{Y}_i\,. \tag{7.10}$$

To use the estimator $\widehat{\phi}^{\,\beta\text{-IPW}}(\pi; \beta)$, we must specify β. When this parameter is chosen to satisfy Assumption 7.3 alongside Assumption 7.1, Y. Zhang and Imai (2023) proved the following result.

Theorem 7.3. *Under Assumptions 7.1 and 7.3, we have that the β-IPW estimator $\widehat{\phi}^{\beta\text{-}IPW}(\pi)$ is an unbiased estimator for the policy value, $\mathbb{E}\left[\widehat{\phi}^{\beta\text{-}IPW}(\pi;\beta)\right] = V(\pi)$.*

The choice of β dictates the strength of the β-additive assumption, which becomes less restrictive as β increases. Setting $\beta = \max_i M_i$ imposes no additional constraints on the interference structure since, in this case, we have that $\widehat{\phi}^{\beta\text{-IPW}} = \widehat{\phi}^{\text{IPW}}$ (Y. Zhang & Imai, 2023). Thus, from a practical point of view, the greatest variance reduction can be achieved by using a smaller β.

To highlight the best variance reduction we can possibly achieve with the β-IPW estimator, we consider the special case of $\widehat{\phi}^{\text{addIPW}}(\pi) = \widehat{\phi}^{\beta\text{-IPW}}(\pi;\beta = 1)$, which we refer to as the additive IPW estimator because there are no interactions between multiple treatments within the same cluster. Here, we can establish the following claim (see Appendix 7.A.5 for proof).

Lemma 7.2. *We have that* $\text{Var}\left(\widehat{\phi}^{addIPW}\right) \propto \max_i M_i^2$.

Thus, for the additive IPW estimator, we observe that its variance will scale quadratically with the cluster size. This is a notable improvement over the exponential scaling of the standard IPW estimator.

7.6. Experiments

We aim to evaluate the performance of our proposed strategies for estimating Qini curves under clustered network interference. To balance realistic structures of interference with the benefit of having a ground-truth in synthetic data, we designed a simulator that mimics an e-commerce marketplace where interference arises through cannibalization among product items sold by different vendors. Notably, we implemented various interference structures in which Assumptions 7.2 and 7.3 are also violated. We present our simulator as a framework that can be reused for future research on the topic of interference. We provide all code in the GitHub repository: https://github.com/bookingcom/uplift-interference-simulator.

In our experiments, we compare five strategies for estimating Qini curves. First, we implement a strategy that ignores all interference, which we refer to as the naive estimator (more details are found in Appendix 7.C). Next, we implement the estimators discussed in this paper: the standard IPW estimator, the fractional IPW estimator and the β-IPW estimator with $\beta = 1$ (additive IPW) or $\beta = 2$.

Evaluation criteria Our experiments focus on two key aspects of using Qini curves for decision-making. The first aspect is calibration: how accurately the estimates $\{\widehat{Q}_k\}_{k=0}^K$ reflect the ground truth values $\{Q_k\}_{k=0}^K$. Depending on the experiment, we assess this

using bias $K^{-1}\sum_{k=1}^{K}\mathbb{E}[\widehat{Q}_k - Q_k]$, variance $K^{-1}\sum_{k=1}^{K}\mathrm{Var}(\widehat{Q}_k)$, and mean squared error $K^{-1}\sum_{k=1}^{K}\mathbb{E}[(\widehat{Q}_k - Q_k)^2]$. The second aspect is discrimination: the ability to determine which policy is better. For this, we rank policies based on the estimated area under the Qini curve (higher is better) and use Kendall rank correlation to measure how well each estimator ranks policies compared to the ground truth ranking. By default, since our focus lies on evaluating treatment policies rather than the policies themselves, we implement a simple baseline policy with access to the underlying data-generating process. We then progressively degrade its performance by adding increasing levels of noise to its scoring rule. More details on this policy are provided in Appendix 7.D.1. To simplify evaluation, we perform experiments in the uniform cost case.

7.6.1. Simulating an e-commerce marketplace with clustered network interference

In this section, we describe a data-generating process where clusters correspond to potential buyers searching for some item, while treatment units are the items shown. In this marketplace, the treatment of an item W_{ij} could correspond to e.g. discounts or promotions, and the outcome Y_{ij} is whether the item (i, j) was purchased by the buyer i. Treatment effects manifest as an incremental change in the probability of a purchase to occur due to the treatment. Each buyer can make at most one purchase, causing cannibalization as treatments may shift purchases between items rather than increasing total purchases.

To construct the dataset, we first sample the covariates (X_i, Z_{ij}) and treatment W_{ij}. Next, to introduce heterogeneous treatment effects, we compute an item attractiveness score matrix $\mathbf{A}$, where each element $A_{ij} \in [0, 1]$ represents buyer i's interest in purchasing item (i, j). The elements in $\mathbf{A}$ depend on the covariates and assigned treatment. Details on this sampling and computation are provided in Appendix 7.D.1. For simplicity, we assume all buyers observe the same number of items, denoted by M.

Next, before sampling the outcome Y_{ij}, which indicates whether item (i, j) is purchased, we first determine whether buyer i makes any purchase at all, denoted with the binary variable $\tilde{Y}_i$. Conditional on a purchase occurring, we then sample which specific item the buyer purchases. We sample $\tilde{Y}_i$ according to the Bernoulli probability $\eta(\mathbf{A}_i) = P(\tilde{Y}_i = 1 \mid \mathbf{A}_i)$. The structure of the interference in this dataset is largely determined by the choice of η.

We consider three alternatives for η. The simplest is $\eta_{\max}(\mathbf{A}_i) = \max_j A_{ij}$ where only the most attractive item contributes the probability of a purchase by buyer i. We refer to $\eta_{\max}$ as the max function. Next, we consider the product function $\eta_{\text{product}}(\mathbf{A}_i) = 1 - \prod_{j=1}^{M}(1 - A_{ij})$. This function assumes that items contribute independently to a purchase such that $P(\tilde{Y}_i = 1) = 1 - P(\tilde{Y}_i = 0) = 1 - \prod_{j=1}^{M} P(Y_{ij} = 0)$. Lastly, the third function is inspired by position bias, commonly found in ranking systems used in e-commerce platforms (Joachims *et al.*, 2005). We refer to this as the exponential decay function, defined as $\eta_{\text{exp-decay}}(\mathbf{A}_i) = \sum_{j=1}^{M}(\frac{1}{2})^{\text{rank}(A_{ij})} A_{ij}$, where $\text{rank}(\cdot)$ returns the rank of the attractiveness

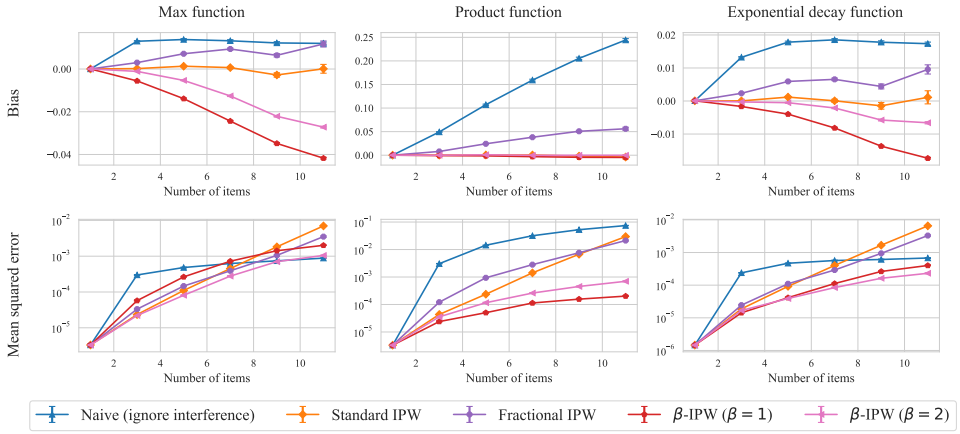


Figure 7.2: Comparison of bias and mean squared error for each strategy under different interference structures. We let $N = 100.000$ and $M = 11$. Averages and standard errors are reported over 150 repetitions.

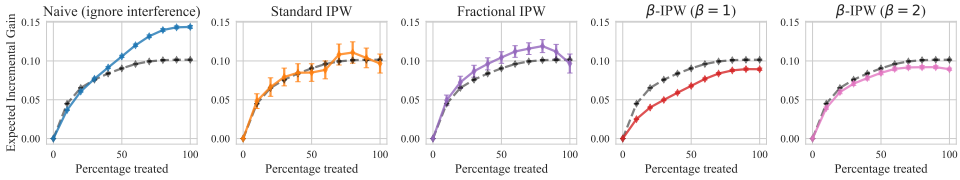


Figure 7.3: Qini curves of each estimation strategy with $N = 100.000$ and $M = 11$ using with the exponential decay function $\eta_{\text{exp-decay}}$. The average Qini curve and standard error are reported over 150 repetitions. The dashed black line corresponds to the true underlying Qini curve.

scores for buyer i in descending order. Each successive item contributes half as much as the preceding one to the probability of a purchase by buyer i.

For the final step, conditional on that we sample $\tilde{Y}_i = 1$, we determine which item (i, j) is purchased by sampling according to the probabilities given by the softmax function $P(Y_{ij} = 1 \mid \tilde{Y}_i = 1, \mathbf{A}_i) = e^{A_{ij}/\lambda} / \sum_{j=1}^{M} e^{A_{ij}/\lambda}$ where we set the temperature parameter $\lambda = 0.1$.

7.6.2. How does interference affect the estimation error?

In the first experiment, we evaluated all estimation strategies under different interference structures by varying η as described in the previous subsection. In addition, we fixed the number of buyers (i.e., clusters) while varying the number of items M (i.e., units), since spillover effects due to interference are largely expected to depend on cluster size; when

the cluster size is one, there is no interference. We compared the bias and mean squared error (MSE) of each estimation strategy, as shown in Figure 7.2.

Starting with the bias, we observed that the naive estimator is significantly biased for all interference structures. The bias, which increased with number of items, was generally smaller for the fractional IPW and β-IPW estimators with $\beta = 2$. Meanwhile, the standard IPW estimator appeared unbiased in all cases. In all cases, the β-IPW estimator with $\beta = 2$ is observed to have lower bias than the variant with $\beta = 1$ and, in some cases, for larger number of items (e.g., more than 10) the variant with $\beta = 1$ had a similar or larger absolute bias as the naive estimator.

In terms of MSE, the standard IPW and fractional IPW estimators performed worst, with MSE increasing exponentially with cluster size. The β-IPW estimators performed the best; in particular, for the product function, $\beta = 1$ yielded the best MSE. The naive estimator performed the worst for small item counts but its increase in MSE appeared to slow down for larger number of items.

Finally, by visually examining the average Qini curve for a fixed number of items $M = 11$ with the exponential decay function, as in Figure 7.3, we can obtain a more qualitative assessment of the bias of each estimation strategy. Starting from the left, we observe that the naive estimator vastly overestimates the Qini curve; the standard IPW is centered around the true curve; fractional IPW also overestimates, but to a lesser degree than the naive; and the β-IPW estimators underestimate the Qini curves, where $\beta = 2$ has less bias than $\beta = 1$ as expected. We further examined the average difference in the AUC of the estimated Qini curves and the ground-truth ones across all interference structures, and we observe similar trends for the other interference structure (max and product); the full results are provided in Appendix 7.D.5.

7.6.3. Which estimation strategy is most efficient?

In the next experiment, we evaluated the efficiency of each estimation strategy by reporting the variance as we varied the number the buyers N (i.e., clusters) or items M (i.e., units). While varying one, the other was kept fixed to $N = 20.000$ and $K = 11$. We present results only using $\eta_{\text{exp-decay}}$ as we observed no difference when changing this function. The results are shown in Figure 7.4 in the appendix, and we summarize our main findings below.

The results indicate that the most efficient estimators, ranked from lowest to highest variance, are: naive, β-IPW ($\beta = 1$), β-IPW ($\beta = 2$), fractional IPW, and standard IPW. While the naive has the lowest variance, we recall that it also has the largest bias since it fails to take into account the interference. As expected, the variance of all estimators improved with more buyers (i.e., clusters) at a similar rate. Meanwhile, the variance of the standard and fractional IPW estimators increased exponentially with the number of items (i.e., cluster size), whereas the others scaled sub-exponentially.

7.6.4. How well can we rank policies under clustered network interference?

In the final experiment, we evaluated each estimation strategy's ability to rank policies based on the estimated area under the Qini curve. To do so, we degraded the baseline treatment policy used for evaluation by adding progressively larger noise to its treatment prioritization rule, generating seven policies with decreasing performance. This experiment was repeated 150 times with $N = 20.000$ and $K = 11$ for each η, and we report the average Kendall rank correlation coefficient between each strategy's ranking and the ground truth.

The results, shown in Table 7.1, indicate that β-IPW with $\beta = 1$ performed best overall, achieving perfect rankings in all cases (rank correlation = 1). The other estimators achieved similar or only slightly lower correlations, except for η_{product}, where the naive estimator, standard IPW, and fractional IPW had correlations between 0.80 and 0.85.

7.7. Discussion

Our findings indicate that, while clustered network interference can cause severe bias in Qini curve estimates when the interference is ignored, it is possible to get accurate estimates using different estimation strategies that take this interference into account. However, the best estimation strategy will depend on several application-specific factors, including cluster size, the number of observations, and prior beliefs about interference and the intended use of the Qini curve.

For small cluster sizes (e.g., fewer than 5), the choice of estimation strategy had a limited impact on estimation error, as IPW, fractional IPW, and β-IPW all performed comparably in terms of bias and mean squared error. For larger cluster sizes, however, we observed a trade-off between using an unbiased, high-variance estimator and a possibly biased, low-variance estimator. For unbiased estimation, the standard IPW estimator is preferred as it relies only on the weakest conditions regarding the interference, though it requires a large number of observations to be reliable. When the number of observations is limited, low-variance unbiased estimation might still be feasible if strong domain expertise about the interference structure can justify the use of either the fractional IPW or the β-IPW estimator. However, if some bias is tolerated, then our results suggest that β-IPW is a strong choice, as it has the lowest variance among strategies that account for interference.

The acceptable level of bias depends on the decision-making context. For model selection – i.e., discriminating between good and bad policies – we observed the β-IPW estimator performing best. Interestingly, the naive estimator that ignores interference also ranked policies correctly in some cases, suggesting that bias due to interference may have a limited effect on this type of decision-making. However, if the goal is to determine a suitable threshold for a treatment prioritization rule, a well-calibrated Qini curve becomes more critical, making an estimator with low bias preferable.

Table 7.1: Comparison of ability to rank policies by each estimation strategy. We used $N = 20.000$ number of buyers (i.e., clusters) and $K = 11$ items (i.e., units) per buyer. We report the average Kendall rank correlation with respect to the ground truth ranking over 150 repetitions. Higher is better, where 1 corresponds to a perfect rank correlation.

Estimation strategy	Max	Product	Exponential decay
Naive (ignore interference)	1.000	0.808	1.000
Standard IPW	0.928	0.845	0.945
Fractional IPW	0.991	0.806	0.995
β-IPW ($\beta = 1$)	1.000	1.000	1.000
β-IPW ($\beta = 2$)	1.000	0.995	1.000

Additional considerations when using β-IPW estimator Since Assumptions 7.2 and 7.3 are untestable, it is difficult to choose between the fractional and β-IPW estimators without domain expertise. However, for the β-IPW in particular, an alternative could be to adopt a data-adaptive strategy to find an estimator optimal with respect to the MSE. We can leverage the fact that larger values of β make the β-additivity assumption less restrictive. Recall that as β increases, the β-IPW estimator converges to the standard IPW estimator. This relationship can be exploited to estimate the bias for a given β by comparing the β-IPW estimate to that of the unbiased standard IPW estimator. Combining this bias estimate with an estimate of the sampling variance of the β-IPW estimator obtained via bootstrapping allows us to compute an MSE estimate for each β, from which we can select the value that minimizes the estimated MSE. A potential limitation of this approach, however, is that the MSE estimate itself may have high variance, owing to the uncertainty propagated from the standard IPW estimator used as a reference for estimating the bias. Despite this practical challenge, we view this approach as a promising direction for future work toward developing more data-adaptive methods for choosing among the estimators proposed here.

7.8. Conclusion

To summarize, we have introduced a framework for estimating Qini curves in experimental study designs with clustered network interference, along with multiple estimation strategies. Our results demonstrate that properly accounting for interference leads to more accurate Qini curve estimation, though the best estimation strategy depends on the specific context. While these methods are not a universal solution for all types of interference and should be applied with care, especially in high-stakes settings, we provide practical recommendations based on both theoretical insights and empirical evidence. These guidelines are intended to help practitioners more effectively assess the cost-effectiveness of treatment policies in complex settings where interference is present.

Appendices

7.A. Proofs and derivations

7.A.1. Proof of Lemma 7.1

Proof. Under Assumption 7.1, we can show that $\mathbb{E}[\phi(\pi)] = V(\pi)$ by rewriting the expectation as follows,

$$\begin{aligned}
\phi(\pi) &= \mathbb{E}\left[\mathbb{E}[\tilde{Y} \mid \mathbf{W} = \boldsymbol{\pi}, X]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{M_i} Y_{ij}(\boldsymbol{\pi}_i) \mid \mathbf{W} = \boldsymbol{\pi}_i, X_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{M_i} Y_{ij}(\boldsymbol{\pi}_i) \mid X_i\right]\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{M_i} Y_{ij}(\boldsymbol{\pi}_i)\right]
\end{aligned}$$

where the second equality from that we defined $\tilde{Y}_i = \sum_{j=1}^{M_i} Y_{ij}$ and then $Y_{ij} = Y_{ij}(\boldsymbol{\pi}_i)$ due to consistency in Assumption 7.1, and finally the third equality from conditional exchangeability $Y_{ij}(\boldsymbol{\pi}_i) \perp\!\!\!\perp \mathbf{W}_i \mid X_i$ in Assumption 7.1. We can prove analogously using the same arguments that $\mathbb{E}[\psi(\pi)] = C(\pi)$. □

7.A.2. Proof of Theorem 7.1

Proof. The unbiasedness of the standard IPW estimator $\widehat{\phi}^{\,\mathrm{IPW}}(\pi)$ follows from the same arguments as deriving the inverse probability weighting estimator in settings with no interference (Hernan & Robins, 2023). We can show that

$$\begin{aligned}
\mathbb{E}\left[\widehat{\phi}^{\,\mathrm{IPW}}(\pi)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \tilde{Y}_i\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\Pr(\mathbf{W}_i = \boldsymbol{\pi}_i \mid X_i)} \tilde{Y}_i\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\Pr(\mathbf{W}_i = \boldsymbol{\pi}_i \mid X_i)} \tilde{Y}_i \mid X_i\right]\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\tilde{Y}_i \mid \mathbf{W}_i = \boldsymbol{\pi}_i, X_i\right]\right] \\
&= \mathbb{E}\left[\phi(\pi)\right]
\end{aligned}$$

The second equality follows from the independent treatment assignments where

$$\Pr(\mathbf{W}_i = \boldsymbol{\pi}_i \mid X_i) = \prod_{j=1}^{M_i} \Pr(W_{ij} = \pi_{ij} \mid X_i) = \prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i),$$

The unbiasedness of $\widehat{\phi}^{\,\mathrm{IPW}}(\pi)$ then follows from $\mathbb{E}\left[\phi(\pi)\right] = V(\pi)$ under Assumption 7.1.

Next, we derive the expression for $\mathrm{Var}\left(\widehat{\phi}^{\,\mathrm{IPW}}(\pi)\right)$. Due to independence of clusters, we first note that

$$\mathrm{Var}\left(\widehat{\phi}^{\,\mathrm{IPW}}(\pi)\right) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \tilde{Y}_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}\left(\frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \tilde{Y}_i\right).$$

By using the law of total variance, we can rewrite

$$\begin{aligned}\mathrm{Var}\left(\frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \tilde{Y}_i\right) = \mathbb{E}\Bigg[\underbrace{\mathrm{Var}\left(\frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \tilde{Y}_i \mid \tilde{Y}_i(\boldsymbol{\pi}_i), X_i\right)}_{(a)}\Bigg] \\ + \mathrm{Var}\left(\underbrace{\mathbb{E}\left[\frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \tilde{Y}_i \mid \tilde{Y}_i(\boldsymbol{\pi}_i), X_i\right]}_{(b)}\right).\end{aligned}$$

Inspecting (b) first, we note that

$$\begin{aligned}(b) &= \mathbb{E}\left[\frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \tilde{Y}_i(\boldsymbol{\pi}_i) \mid \tilde{Y}_i(\boldsymbol{\pi}_i), X_i\right] \\ &= \tilde{Y}_i(\boldsymbol{\pi}_i)\mathbb{E}\left[\frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \mid \tilde{Y}_i(\boldsymbol{\pi}_i), X_i\right] \\ &= \tilde{Y}_i(\boldsymbol{\pi}_i).\end{aligned}$$

where it follows from conditional exchangeability in Assumption 7.1 that

$$\mathbb{E}\left[\frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \mid \tilde{Y}_i(\boldsymbol{\pi}_i), X_i\right] = \mathbb{E}\left[\frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} \mid X_i\right] = 1.$$

Similarly, we can show that

$$\begin{aligned}(a) &= \text{Var}\left(\frac{\mathbf{1}\left(\mathbf{W}_i = \boldsymbol{\pi}_i\right)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)}\,\tilde{Y}_i(\boldsymbol{\pi}_i) \mid \tilde{Y}_i(\boldsymbol{\pi}_i), X_i\right) \\ &= \left[\frac{\tilde{Y}_i(\boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)}\right]^2 \text{Var}\left(\mathbf{1}\left(\mathbf{W}_i = \boldsymbol{\pi}_i\right) \mid \tilde{Y}_i(\boldsymbol{\pi}_i), X_i\right) \\ &= \left[\frac{\tilde{Y}_i(\boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)}\right]^2 \text{Var}\left(\mathbf{1}\left(\mathbf{W}_i = \boldsymbol{\pi}_i\right) \mid X_i\right)\end{aligned}$$

where the last equality follows from conditional exchangeability again. Next, using that treatment assignments are independent, we can further rewrite

$$\begin{aligned}\text{Var}\left(\mathbf{1}\left(\mathbf{W}_i = \boldsymbol{\pi}_i\right) \mid X_i\right) &= \text{Var}\left(\prod_{j=1}^{M_i} \mathbf{1}\left(W_{ij} = \pi_{ij}\right) \mid X_i\right) \\ &= \prod_{j=1}^{M_i} \left\{\text{Var}\left(\mathbf{1}\left(W_{ij} = \pi_{ij}\right) \mid X_i\right) + \mathbb{E}\left[\mathbf{1}\left(W_{ij} = \pi_{ij}\right) \mid X_i\right]^2\right\} \\ &\quad - \prod_{j=1}^{M_i} \mathbb{E}\left[\mathbf{1}\left(W_{ij} = \pi_{ij}\right) \mid X_i\right]^2 \\ &= \prod_{j=1}^{M_i} \left\{e_1(X_i)e_0(X_i) + e_{\pi_{ij}}(X_i)^2\right\} - \prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)^2\end{aligned}$$

Plugging the above expression back into (a), we get

$$\begin{aligned}(a) &= \left[\frac{\tilde{Y}_i(\boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)}\right]^2 \left[\prod_{j=1}^{M_i} \left\{e_1(X_i)e_0(X_i) + e_{\pi_{ij}}(X_i)^2\right\} - \prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)^2\right] \\ &= \left[\tilde{Y}_i(\boldsymbol{\pi}_i)\right]^2 \left[\prod_{j=1}^{M_i} \left\{\frac{e_1(X_i)e_0(X_i)}{e_{\pi_{ij}}(X_i)^2} + 1\right\} - 1\right]\end{aligned}$$

At last, plugging our expression of (a) and (b) back into where we started, we obtain the final expression for the variance of the standard IPW estimator,

$$\text{Var}\left(\widehat{\phi}^{\,\text{IPW}}(\pi)\right) = \frac{1}{N^2}\sum_{i=1}^{M_i}\left\{\mathbb{E}\left[\left[\tilde{Y}_i(\boldsymbol{\pi}_i)\right]^2\left[\prod_{j=1}^{M_i}\left(\frac{e_1(X_i)e_0(X_i)}{e_{\pi_{ij}}(X_i)^2} + 1\right) - 1\right]\right] + \text{Var}\left(\tilde{Y}_i(\boldsymbol{\pi}_i)\right)\right\}$$

□

7.A.3. Expressing q_{ij} in terms of the propensity score

We have

$$\begin{aligned} q_{ij}(\boldsymbol{\pi}_i, X_i) &= \Pr(W_{ij} = \pi_{ij}, \overline{W}_i = \overline{\pi}_i \mid X_i) \\ &= \Pr(W_{ij} = \pi_{ij} \mid \overline{W}_i = \overline{\pi}_i X_i) \Pr(\overline{W}_i = \overline{\pi}_i \mid X_i) \,. \end{aligned}$$

As the propensity score $e_w(X_i) = \Pr(W_{ij} = w \mid X_i)$ is the same for every $j = 1, \ldots, M_i$, we have that all units in a cluster have the same probability of being treated. Therefore, once conditioning on the fraction $\overline{W}_i$ of treated in a cluster, the fraction equals to the probability that a unit has been treated in that cluster. We can thus write $\Pr(W_{ij} = \pi_{ij} \mid \overline{W}_i = \overline{\pi}_i, X_i) = \pi_{ij} \cdot \overline{\pi}_i + (1 - \pi_{ij}) \cdot (1 - \overline{\pi}_i)$.

Next, for the second probability $\Pr(\overline{W}_i = \overline{\pi}_i \mid X_i)$, we note that $\overline{W}_i = M_i^{-1} \sum_{j=1}^{M_i} W_{ij}$ can be seen a Binomial random variables scaled by M_i^{-1}. This means $M_i^{-1} \sum_{j=1}^{M_i} W_{ij} \sim \mathrm{B}(M_i, e_1(X_i))$ and thus we have $\Pr(\overline{W}_i = \overline{\pi}_i \mid X_i) = \binom{M_i}{\overline{\pi}_i \cdot M_i} [e_1(X_i)]^{\overline{\pi}_i \cdot M_i} [1 - e_1(X_i)]^{(1-\boldsymbol{\pi}_i) \cdot M_i}$.

Combining both expressions from above, we get

$$q_{ij}(\pi_i, X_i) = \left[\pi_{ij} \cdot \overline{\pi}_i + (1 - \pi_{ij}) \cdot (1 - \overline{\pi}_i)\right] \times \binom{M_i}{\overline{\pi}_i \cdot M_i} [e_1(X_i)]^{\overline{\pi}_i \cdot M_i} [e_0(X_i)]^{(1-\overline{\pi}_i) \cdot M_i} \,.$$

7.A.4. Proof of Theorem 7.2

Proof. We can show that

$$\begin{aligned} \mathbb{E}\left[\widehat{\phi}^{\text{fracIPW}}\right] &= \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{1(W_{ij} = \pi_{ij}, \overline{W}_i = \overline{\pi}_i)}{q_{ij}(\boldsymbol{\pi}_i, X_i)} Y_{ij}\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{M_i} \frac{1(W_{ij} = \pi_{ij}, \overline{W}_i = \overline{\pi}_i)}{q_{ij}(\boldsymbol{\pi}_i, X_i)} Y_{ij} \mid X_i\right]\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{M_i} Y_{ij} \mid d_{ij}(\mathbf{W}_i) = [\pi_{ij}, \overline{\pi}_i], X_i\right]\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{M_i} Y_{ij}(\mathbf{W}_i) \mid d_{ij}(\mathbf{W}_i) = [\pi_{ij}, \overline{\pi}_i], X_i\right]\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{M_i} Y_{ij}(\mathbf{W}_i) \mid X_i\right]\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\sum_{j=1}^{M_i} Y_{ij}(\mathbf{W}_i)\right] \\ &= \frac{1}{N} \sum_{i=1}^{N} V(\pi) = V(\pi) \end{aligned}$$

where the second equality follows from linearity of expectations and law of iterated expectations, the fourth equality follows consistency in Assumption 7.1 and that the exposure mapping fulfills Assumption 7.2, and finally the fifth equality from conditional exchangeability in Assumption 7.1 because $Y_{ij}(\mathbf{w}) \perp\!\!\!\perp \mathbf{W_i} \mid X_i \Rightarrow Y_{ij}(\mathbf{w}) \perp\!\!\!\perp d(\mathbf{W_i}) \mid X_i$ for all $\mathbf{w} \in \{0,1\}^{M_i}$. □

7.A.5. Proof of Lemma 7.2

Proof. We have defined $\widehat{\phi}^{\text{addIPW}}(\pi) = \widehat{\phi}^{\beta\text{-IPW}}(\pi; \beta = 1)$ which has a simpler form

$$\widehat{\phi}^{\text{addIPW}}(\pi) = \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{j=1}^{M_i}\frac{\mathbf{1}(W_{ij}=\pi_{ij})}{e_1(X_i)} - (M_i - 1)\right\}\tilde{Y}_i \,.$$

As clusters are independent, we can write

$$\text{Var}\left(\widehat{\phi}^{\text{addIPW}}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\text{Var}\left(\left\{\sum_{j=1}^{M_i}\frac{\mathbf{1}(W_{ij}=\pi_{ij})}{e_1(X_i)} - (M_i - 1)\right\}\tilde{Y}_i\right)$$

where the variance terms inside the sum can be decomposed as

$$\underbrace{\mathbb{E}\left[\left(\left\{\sum_{j=1}^{M_i}\frac{\mathbf{1}(W_{ij}=\pi_{ij})}{e_1(X_i)} - (M_i - 1)\right\}\tilde{Y}_i\right)^2\right]}_{(a)} - \underbrace{\mathbb{E}\left[\left\{\sum_{j=1}^{M_i}\frac{\mathbf{1}(W_{ij}=\pi_{ij})}{e_1(X_i)} - (M_i - 1)\right\}\tilde{Y}_i\right]^2}_{(b)} .$$

When $\widehat{\phi}^{\text{addIPW}}$ is an unbiased estimator, we have that $(b) = V(\pi)^2$. For (a), we note that the sum $\sum_{j=1}^{M_i}\frac{\mathbf{1}(W_{ij}=\pi_{ij})}{e_1(X_i)}$ is linear with respect to M_i. Thus, inspecting the full expression for the variance,

$$\text{Var}\left(\widehat{\phi}^{\text{addIPW}}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\left(\mathbb{E}\left[\left\{\sum_{j=1}^{M_i}\frac{\mathbf{1}(W_{ij}=\pi_{ij})}{e_1(X_i)} - (M_i - 1)\right\}^2\tilde{Y}_i^2\right]\right) - V(\pi)^2 \,,$$

we can see that the variance will scale quadratically with M_i. □

7.B. Variance reduction through augmented weighted estimators

In this work, we have considered weighting estimators that require only a single nuisance model, namely the propensity score $e_w(x) = \Pr(W = w \mid X = x)$. Because the propensity score is assumed to be known in our setting, these estimators are appealing as they eliminate the need to estimate any additional nuisance models. This, in turn, helps avoid potential biases that could arise from model misspecification. However, a key limitation of weighting estimators is that they can suffer from high variance. In this appendix, we briefly

discuss a technique that introduces an additional nuisance model without risking bias, while offering the potential for variance reduction if the added model is well specified. We also present experimental results showing that this approach can indeed reduce variance without introducing bias, although the magnitude of the reduction may vary depending on the setting.

We start with a weighting estimator of the form $\widehat{\phi}(\pi;\omega) = \frac{1}{N}\sum_{i=1}^{N} \omega(\boldsymbol{W}_i, \boldsymbol{\pi}_i, X_i)\tilde{Y}_i$, where ω is a pre-specified weighting function. Both the standard IPW estimator and the β-IPW estimator can be expressed in this form using different choices of weighting functions:

$$\omega^{\text{IPW}}(\boldsymbol{W}_i, \boldsymbol{\pi}_i, X_i) = \frac{\mathbf{1}(\mathbf{W}_i = \boldsymbol{\pi}_i)}{\prod_{j=1}^{M_i} e_{\pi_{ij}}(X_i)} ,$$

$$\omega^{\beta\text{-IPW}}(\boldsymbol{W}_i, \boldsymbol{\pi}_i, X_i) = \sum_{\mathscr{U}\in\mathscr{J}_i^{\beta}} \prod_{j\in\mathscr{U}} \left(\frac{\mathbf{1}(W_{ij} = \pi_{ij})}{e_{\pi_{ij}}(X_i)} - 1 \right) .$$

This class of weighting estimators does not include the fractional IPW estimator, since it performs weighting directly on unit-level outcomes, as can be seen from inspecting eq. (7.9).

We shall focus on an "augmented" variant of these weighted estimators which we define as

$$\widehat{\phi}^{\text{augmented}}(\pi;\omega) = \frac{1}{N}\sum_{i=1}^{N} \omega(\boldsymbol{W}_i, \boldsymbol{\pi}_i, X_i)(\tilde{Y}_i - \widehat{g}(X_i)) + \widehat{g}(X_i) , \tag{7.11}$$

where $\widehat{g}(x)$ is an estimator of the cluster-level conditional expectation of the outcome given the pre-treatment covariates, $\mathbb{E}[\tilde{Y} \mid X = x]$. The cluster-level treatment information is excluded to avoid modeling the potentially high-dimensional structure of the cluster-level treatment.

The augmented weighted estimator is reminiscent of the augmented inverse probability weighting (AIPW) estimator (Robins *et al.*, 1994), which has been used to adjust for pre-treatment covariates with the goal of reducing estimator variance while protecting against bias when the true propensity score is known (see, e.g., Cao *et al.* (2009) and Karlsson, Wang *et al.* (2026)). AIPW estimators model both the treatment-dependent conditional expected outcome and treatment probability and are doubly robust in the sense that if either of the estimators for these models is correctly specified, the AIPW estimator remains consistent. In our setting, where the design is randomized and the propensity score is known, a similar robustness guarantee holds. However, unlike the AIPW estimator, our approach relies on a prognostic cluster-level outcome model, $\mathbb{E}[\tilde{Y} \mid X]$, that does not incorporate treatment information, as would typically be required in a doubly robust estimator. The form of the augmented estimator also shares similarities with prediction-powered inference estimators (Angelopoulos *et al.*, 2023), which use an auxiliary prediction model to improve inference, although those estimators do not employ any weighting function.

We can use any flexible, data-adaptive model to learn $\widehat{g}$, and notably, $\widehat{\phi}^{\text{augmented}}(\pi;\omega)$ is robust in the sense that it remains unbiased for the policy value regardless of whether $\widehat{g}(X)$

is correctly specified, provided that the weighting function ω satisfies some conditions. Additionally, to ensure this robustness guarantee, $\widehat{g}$ must be fitted independently of the observations used in the weighting estimator. In practice, this can be achieved using a cross-fitting procedure: the data are split into two folds, $\widehat{g}$ is fitted on one fold and predictions are made on the other fold, and then the roles of the folds are swapped and the procedure is repeated.

Theorem 7.4. *Suppose the weighted estimator $\widehat{\phi}(\pi;\omega)$ is unbiased for the policy value, $\mathbb{E}[\widehat{\phi}(\pi;\omega)] = V(\pi)$, and that the weighting function satisfies $\mathbb{E}[\omega(\boldsymbol{W},\boldsymbol{\pi},X) \mid X] = 1$. Furthermore, assume that $\widehat{g}$ is obtained independently of the observations used to compute its predictions. Then the augmented weighting estimator $\widehat{\phi}^{augmented}(\pi;\omega)$ is unbiased for the policy value, $\mathbb{E}[\widehat{\phi}^{augmented}(\pi;\omega)] = V(\pi)$.*

Proof. We can show

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\phi}^{\text{augmented}}(\pi;\omega)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\omega(\boldsymbol{W}_i,\boldsymbol{\pi}_i,X_i)(\tilde{Y}_i-\widehat{g}(X_i))+\widehat{g}(X_i)\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\omega(\boldsymbol{W}_i,\boldsymbol{\pi}_i,X_i)\tilde{Y}_i\right] - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\Big(1-\omega(\boldsymbol{W}_i,\boldsymbol{\pi}_i,X_i)\Big)\widehat{g}(X_i)\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\omega(\boldsymbol{W}_i,\boldsymbol{\pi}_i,X_i)\tilde{Y}_i\right] - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\Big[1-\omega(\boldsymbol{W}_i,\boldsymbol{\pi}_i,X_i)\mid X_i\Big]\widehat{g}(X_i)\right] \\
&= E\left[\frac{1}{N}\sum_{i=1}^{N}\omega(\boldsymbol{W}_i,\boldsymbol{\pi}_i,X_i)\tilde{Y}_i\right] \\
&= E\Big[\widehat{\phi}(\pi;\omega)\Big] = V(\pi)
\end{aligned}
$$

where the fourth equality follows from the assumption that $\mathbb{E}\big[\omega(\boldsymbol{W}_i,\boldsymbol{\pi}_i,X_i) \mid X_i\big] = 1$ and that $\widehat{g}$ can be taken out of the inner expectation because it is independent of the observations used to compute its predictions. The final equality follows from that the weighted estimator $\widehat{\phi}(\pi;\omega)$ is an unbiased estimator. □

To illustrate how the above theorem applies to the standard IPW and β-IPW estimators, we first note that the standard IPW estimator is unbiased under Assumption 7.1, and that $\mathbb{E}[\omega^{\text{IPW}}(\boldsymbol{W}_i,\boldsymbol{\pi}_i,X_i) \mid X_i] = 1$ as shown in the proof of its unbiasedness in Appendix 7.A.2. Under the additional Assumption 7.3 and the study design considered in this paper, the β-IPW estimator is also unbiased, and we can similarly show that $\mathbb{E}[\omega^{\beta\text{-IPW}}(\boldsymbol{W}_i,\boldsymbol{\pi}_i,X_i) \mid$

$X_i] = 1$ holds as follows:

$$\begin{aligned}
\mathbb{E}[\omega^{\beta\text{-IPW}}(\boldsymbol{W}_i, \boldsymbol{\pi}_i, X_i) \mid X_i] &= \mathbb{E}\left[\sum_{\mathscr{U} \in \mathscr{I}_i^\beta} \prod_{j \in \mathscr{U}} \left(\frac{\mathbf{1}(W_{ij} = \pi_{ij})}{e_{\pi_{ij}}(X_i)} - 1\right) \mid X_i\right] \\
&= \sum_{\mathscr{U} \in \mathscr{I}_i^\beta} \mathbb{E}\left[\prod_{j \in \mathscr{U}} \left(\frac{\mathbf{1}(W_{ij} = \pi_{ij})}{e_{\pi_{ij}}(X_i)} - 1\right) \mid X_i\right] \\
&= 1 + \sum_{\mathscr{U} \in \mathscr{I}_i^\beta \setminus \{\emptyset\}} \mathbb{E}\left[\prod_{j \in \mathscr{U}} \left(\frac{\mathbf{1}(W_{ij} = \pi_{ij})}{e_{\pi_{ij}}(X_i)} - 1\right) \mid X_i\right] \\
&= 1 + \sum_{\mathscr{U} \in \mathscr{I}_i^\beta \setminus \{\emptyset\}} \prod_{j \in \mathscr{U}} \underbrace{\mathbb{E}\left[\left(\frac{\mathbf{1}(W_{ij} = \pi_{ij})}{e_{\pi_{ij}}(X_i)} - 1\right) \mid X_i\right]}_{=0} \\
&= 1
\end{aligned}$$

The third equality follows from explicitly excluding the empty set in the power set $\mathscr{I}i^\beta$ and noting that the product over the empty set is equal to one. The fifth equality follows from the fact that, conditional on X_i, the treatment Wij of unit (i, j) is independent of all other treatments $W_{ij'}$ for $j \neq j'$. This independence allows the product to be moved outside the conditional expectation.

7.B.1. Experiment

Setup We use the same data-generating process as in the main paper, with $N = 20{,}000$ buyers (i.e., clusters) and $K = 11$ items (i.e., units) per buyer. The interference structure is determined using the exponential decay function. To estimate $\mathbb{E}[\tilde{Y} \mid X]$, we use a linear logistic regression model, as the cluster-level outcome $\tilde{Y}$ is binary. We report the bias, variance, and mean-squared error averaged over 150 repetitions.

Results As shown in Table 7.2, the augmented methods may reduce variance compared to their non-augmented counterparts. For the standard IPW estimator, augmentation leads to a decrease in variance (from 0.737 to 0.601) and a lower mean-squared error (from 0.738 to 0.603). For the β-IPW estimator, the variance reduction is negligible, likely because its variance is already low. Overall, augmentation with an outcome model appears to improve estimator efficiency, with the largest gains observed in cases where the initial variance is high, such as with the standard IPW estimator.

7.C. Estimating Qini curves in settings with no interference

We assume the following statement which is equivalent to assuming no interference.

Table 7.2: Comparison of non-augmented and augmented estimators. We used $N = 20.000$ number of buyers (i.e., clusters) and $K = 11$ items (i.e., units) per buyer. We report the average bias, variance and mean-squared error with standard errors (in parentheses) over 150 repetitions.

Estimation strategy	Bias	Variance	MSE
Standard IPW	-0.003 (0.012)	0.737 (0.088)	0.738 (0.032)
Augmented Standard IPW	0.005 (0.010)	0.601 (0.068)	0.603 (0.022)
β-IPW ($\beta = 1$)	-0.016 (0.001)	0.005 (0.001)	0.005 (0.000)
Augmented β-IPW ($\beta = 1$)	-0.016 (0.001)	0.004 (0.001)	0.004 (0.000)

Assumption 7.4. *We assume that* $Y_{ij}(\mathbf{w}) = Y_{ij}(\mathbf{w}')$ *if and only if* $w_{ij} = w'_{ij}$ *for all* $\mathbf{w}, \mathbf{w}' \in \{0,1\}^{M_i}$.

Note that the above assumption is a special case of Assumption 7.2 with the exposure mapping $d_{ij}(\mathbf{W}_i) = W_{ij}$.

We consider the simplest approach in the absence of interference for estimating Qini curves between any units. Consider the estimators based on inverse probability weighting:

$$\widehat{\phi}^{\text{no-interference}}(\pi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{M_i}\frac{\mathbf{1}(W_{ij} = \pi_{ij})}{e_{\pi_{ij}}(X_i)}Y_{ij}$$

$$\widehat{\psi}^{\text{no-interference}}(\pi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{M_i}\frac{\mathbf{1}(W_{ij} = \pi_{ij})}{e_{\pi_{ij}}(X_i)}C_{ij}$$

We can show that under Assumption 7.1 and 7.4, the above estimators are unbiased estimators for the policy value $V(\pi)$ and $C(\pi)$, respectively. Namely, we can show this with the same proof as in Appendix 7.A.4, but replacing $d_{ij}(\mathbf{W_i}) = [W_{ij}, \overline{W}_i]$ with $d_{ij}(\mathbf{W_i}) = W_{ij}$.

7.D. Experimental details

7.D.1. Simulating marketplace dataset

We sample the covariates and treatment as follows: For each buyer $i = 1, \ldots, N$, we sample covariates $X_i \sim \text{U}([0,1]^{12})$. Then, for each item $j = 1, \ldots, M$ we sample covariates $Z_{ij} \sim \text{U}([0,1]^{11})$ and we randomize the treatment assignment by sampling $W_{ij} \sim \text{Bern}(0.5)$. Here, M is the same for all buyers. Since we consider the uniform cost case, we need not sample cost of treatment since they are assumed to be the same each for item.

Next, to introduce heterogeneous treatment effects, we compute an item attractiveness score matrix $\mathbf{A}$ where element A_{ij} relates buyer i's interest in purchasing item j. This matrix depends on both the covariates and treatment as follows

$$A_{ij} = \delta_{ij} \cdot (A_{ij}^{(0)} + W_{ij} \cdot A_{ij}^{(1)}) \,, \tag{7.12}$$

where $A_{ij}^{(w)} = X_i^\top \Omega_w Z_{ij}$ with $\Omega_w \sim \mathrm{U}([0,1]^{12\times 11})$ for $w \in \{0,1\}$. The variable $d_{ij} \sim \mathrm{Bern}(0.5)$ randomly masks some elements in $A_{ij}^{(1)}$ to zero; this emulates that some items will not respond at all to a treatment. Here, A_{ij} typically lies in the range $[0,1]$, but if necessary we clip it to this range so that we later could interpret it as a probability.

Simulating revenue and cost of treating items In our simulations, we also compute the revenue and cost of items in the marketplace if they were to be treated. This allows us to evaluate treatment policies within the simulator that prioritize items based on their predicted expected profit. Specifically, we define the price and margin fraction of item (i,j) as simple linear functions of its covariates Z_{ij}: the price is $P_{ij} = 20+100\cdot Z_{ij,1}$ and the margin fraction is $M_{ij} = 0.01+0.05\cdot Z_{ij,2}$, where $Z_{ij,1}$ and $Z_{ij,2}$ denote the first and second elements of Z_{ij}, respectively. If the treatment $W_{ij} = 1$ corresponds to applying a fixed discount fraction of $d = 0.08$, then the revenue of treating item (i,j) is $R_{ij} = M_{ij}\cdot P_{ij}$ and the cost of treating item (i,j) is $C_{ij} = d\cdot P_{ij}$. Consequently, the potential profit (or loss) from treating the item, conditional on it being converted, is $H_{ij} = R_{ij} - C_{ij} = (M_{ij}-d)\cdot P_{ij}$.

7.D.2. Treatment policy used for evaluation

To construct a simple baseline policy for our simulation studies, we define a treatment prioritization scoring rule S_{baseline} that computes the expected profit from treating item (i,j), ignoring the interference. Since our focus is on evaluating policies rather than developing them, this simplification avoids the complexity of implementing a policy that accounts for interference.

We then compute the scoring rule as

$$S_{\text{baseline}}(X_i, Z_{ij}) = A_{ij}^{(1)} \cdot H_{ij}.$$

Here, $A_{ij}^{(1)}$ represents the incremental change in the probability of conversion after an item has been treated (ignoring interference), while H_{ij} denotes the profit from treating item (i,j). The dependence on the covariates (X_i, Z_{ij}) arises through both $A_{ij}^{(1)}$ and H_{ij}, as described in the previous subsection.

To generate treatment policies with varying performance, we introduce a parameter $\epsilon \in [0,1]$ used to perturb the baseline scoring rule. Specifically, we sample noise from a uniform distribution, $u \sim \mathcal{U}(S_{\min}, S_{\max})$, where $S_{\min}$ and $S_{\max}$ are the minimum and maximum observed values respectively of the scoring rule $S_{\text{baseline}}(X_i, Z_{ij})$. We then define a perturbed policy as

$$S_{\text{baseline}}(X_i, Z_{ij}; \epsilon) = (1-\epsilon)\cdot S_{\text{baseline}}(X_i, Z_{ij}) + \epsilon \cdot u\,.$$

The best-performing policy corresponds to $S_{\text{baseline}}(X_i, Z_{ij}; \epsilon = 0)$, while $S_{\text{baseline}}(X_i, Z_{ij}; \epsilon = 1)$ is equivalent to a completely random policy.

7.D.3. Hardware used for experiments

All experiments have been performed on M1 Macbooks with 16GB of RAM. The runtime is highly dependent on the number of samples being simulated, but anything under 1 million samples takes not more than 10 minutes. The total runtime of our experiments, including all variants required for uncertainty estimation, took about 3 days worth of compute time.

7.D.4. Details on experiment shown in Figure 1

We simulate a dataset with $N = 20.000$ buyers (i.e., clusters) and $K = 3$ items (i.e., units) per buyer with the exponential decay function. We estimate the Qini curve for the treatment prioritization rule $S_{\text{baseline}}(X_i, Z_{ij}; \epsilon = 0)$ described in Appendix 7.D.2 using the naive estimator and the standard IPW estimator. This was repeated 150 times and we plotted the average Qini curve.

7.D.5. Additional experimental results

In this section, we present additional experimental results to support the conclusions in Section 7.6. Figure 7.4 shows the variances of the estimated Qini curves as we vary either the number of buyers N (i.e., clusters) or items M (i.e., units). Meanwhile, Table 7.1 reports the average difference in AUC between the estimated Qini curves and the ground-truth curves over multiple repetitions. These results correspond to the setting with $N = 100{,}000$ and $M = 11$, which is the same as for the Qini curve plots shown in Figure 7.3 in the main paper.

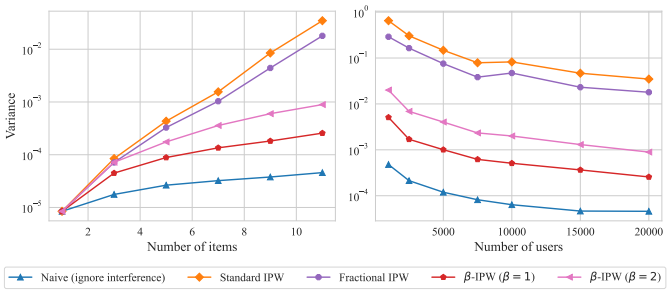


Figure 7.4: Comparing variance of each estimation strategy as we vary the number the buyers N (i.e., clusters) or items M (i.e., units) with $\eta_{\text{exp-decay}}$. While varying one, the other is kept fixed to either $N = 20.000$ or $K = 11$. The variance is reported over 150 repetitions.

Table 7.3: Comparison of estimation strategies via the Area Under the Curve (AUC) relative to the ground-truth Qini curve. Each entry reports the mean AUC and its standard error (in parentheses) across 150 repetitions. We let $N = 100.000$ and $M = 11$.

Interference structure	Naive	Standard IPW	Fractional IPW	β-IPW ($\beta = 1$)	β-IPW ($\beta = 2$)
Max function	0.11 (0.02)	0.00 (0.05)	0.13 (0.04)	-0.44 (0.02)	-0.28 (0.02)
Product function	2.50 (0.03)	-0.04 (0.07)	0.62 (0.06)	-0.05 (0.02)	-0.00 (0.03)
Exponential decay	0.17 (0.02)	0.01 (0.05)	0.11 (0.04)	-0.18 (0.02)	-0.07 (0.02)

8

Discussion

The main aim of this dissertation was to develop methods for safer causal inference. To this end, in Part One we examined approaches to falsify causal identification assumptions in settings with data from multiple sources. Parts Two and Three addressed two different problem settings, proposing new methods that require weaker assumptions than existing approaches. Rather than restating the conclusions to the research questions, which are detailed in each chapter, we now instead discuss broader considerations, reflect on the limitations of the research presented in this dissertation, and suggest possible new research directions.

8.1. Considerations and take-aways

8.1.1. Should we replace one untestable assumption with another?

In Part One, we began by exploring how to test the unconfoundness assumption. We discovered that by introducing another assumption, the independence of causal mechanisms (ICM), we obtained testable implications in the data that allowed us to falsify unconfoundedness under the belief that ICM is valid. While we soon shall discuss the plausibility of the ICM assumption itself, in this subsection we first want to address another, broader point.

When we introduce our falsification strategy, we essentially make the following argument: we can test one assumption which was previously untestable (i.e., unconfoundedness) by introducing another untestable assumption (i.e., ICM). The other falsification strategy we compare to in Part One also makes the same argument, replacing ICM with the transportability condition instead. Since all of these assumptions – unconfoundedness, ICM, and transportability – are untestable conditions, what do we gain from replacing them with each other?

First, because our assumptions should be based on a priori beliefs, we cannot give a

universal reason to prefer one assumption over another. However, as an example, if we are willing to assume both unconfoundedness and ICM, we gain the ability to use to data to test whether at least some parts of our prior beliefs hold. If we find that our beliefs may be false, we must then decide how to update them. Here, it is important that unconfoundedness and ICM are orthogonal assumptions, meaning one could still hold even if the other is violated.

Second, for falsification to be possible, we must be willing to assume at least ICM or transportability, in addition to unconfoundedness. In Part One, we give some arguments for why it may be more natural for ICM to hold even when transportability is violated. In practice, this means that our proposed falsification strategy might be preferred because it does not lead to false positives under transportability violations, as would happen with a transportability-based falsification strategy. However, the converse can also happen. It is logically possible for ICM to be violated even when transportability holds. Thus, neither assumption is strictly weaker than the other. So when we consider one untestable assumption over the other, the decision should ultimately depend on which assumption we find credible. From this perspective, having two distinct falsification strategies, each based on different underlying assumptions, expands the range of scenarios in which a falsification approach can be effectively applied.

Lastly, if one is willing to assume both ICM and transportability, an alternative approach that avoids committing to a single assumption is to evaluate both falsification strategies. Since the falsification strategies rest on different assumptions, applying both can be more informative than relying on either one alone. Specifically, if both strategies do not falsify, this could offer additional evidence for the absence of unmeasured confounding, as we have attempted to refute this assumption in two distinct ways. Conversely, if both tests falsify, one could argue that it becomes more plausible that unmeasured confounding is present, rather than both ICM and transportability being violated. If the tests disagree, meaning that one falsifies while the other does not, this could be explained either by that there is no unmeasured confounding but ICM or transportability is violated, or there is unmeasured confounding but one test failed, for example due to insufficient power.

8.1.2. How plausible is the ICM assumption?

Empirical evidence pointing towards ICM holding in practice is relatively scarce, so consequently many of the arguments in favor of ICM have been conceptual. According to Peters *et al.* (2017, Ch. 2), the ICM assumption covers two aspects: that causal mechanisms are independent of each other and that they are uninformative of each other. One implication of the above aspects is that variables can be locally intervened upon without changing other variables in the same system. This idea of "modularity" has many proponents (Haavelmo, 1944; Pearl, 2009; Schölkopf *et al.*, 2021), and Woodward (2000) even going so far to say that modularity in a system is the single best marker for what makes some relationships causal. Examples of such systems where ICM holds include those where the causal mechanisms correspond to invariant physical laws of nature. However, counterexamples of non-modular systems do exist, see e.g. Cartwright (2004),

which proves the point that ICM in fact can be violated in practice and that caution when assuming ICM is always warranted.

In the search for empirical evidence supporting ICM, one could argue that methods which assume ICM and demonstrate strong performance in real-world applications provide indirect support for its validity. Mooij *et al.* (2016) benchmarked several causal discovery algorithms on a real-world bivariate dataset where the causal directions were known. The two types of algorithms they studied, additive noise models and information-geometric causal inference, both rely on conditions implied by ICM; for detailed discussions of the connection between ICM and additive noise models or information-geometric causal inference, see Peters *et al.* (2017, p. 21) and Daniusis *et al.* (2010), respectively. In the benchmark, additive noise models performed particularly well and were able to distinguish the causal direction from the data, which might suggest that ICM may hold in the real-world settings they examined.

Outside the causal inference literature, as noted in e.g. Chapter 3, the models we propose for falsification resemble Bayesian hierarchical models. These models are commonly applied in multilevel studies and meta-analyses where they often choose the parameter priors to be independent (Gelman, 2007, Ch. 16). This independent prior assumption resembles how the ICM assumption is invoked in our falsification method. Since such hierarchical models are widely used in practice, this fact may provide further support for the idea that the ICM assumption can be sensible.

8.1.3. The bias-variance trade-off when choosing an identification strategy

Traditionally the steps of identification and estimation in causal inference are well-separated. Certain identification conditions must first be justified to ensure the unbiased identification of treatment effects. Once identification is established, the estimation step follows, where additional modeling choices may be introduced to improve the estimation. However, in Parts Two and Three, we observed that this separation is not always clear-cut: some identification conditions directly influence our modeling choices, thereby linking identification and estimation more closely than is typically acknowledged.

For example, in Chapter 7, we examined the β-additive condition which, when satisfied, allowed for unbiased identification of Qini curves in presence of interference. Here, smaller values of β increase the restrictiveness of this condition, while larger values relax it. Interestingly, varying β also altered the estimator which we used, leading us to observe that smaller β values often led to greater variance reduction at the potential cost of bias when the condition was increasingly violated. Despite this bias, using a small β often yielded an estimator with the lowest mean-squared error.

Another example of the link between identification and estimation is provided in Chapter 5 on trial augmentation. Prior work such as X. Li *et al.* (2023) shows that assuming transportability enables a new identification strategy to integrate trial and external data, resulting

in a more efficient estimator than if we had used trial data alone. The drawback is that if the transportability condition is violated, this identification strategy fails and unbiased estimation is not possible. Accepting this trade-off, Yang *et al.* (2023) focused explicitly on minimizing mean-squared error rather than eliminating bias, by selectively using external data based on testable implications of transportability violations.

Taken together, the observations in the above examples raise an interesting question: if relying on an identification condition which introduces a small amount of bias but sufficiently reduces variance to lower mean-squared error, should the resulting estimator truly be considered unsafe? Considering this bias–variance trade-off suggests that, in some cases, allowing limited violations of our identification conditions may actually be beneficial. However, because causal inference has traditionally prioritized unbiased identification, one could argue that preserving this focus remains important for achieving broader adoption by practitioner when developing new methods.

In our own approach to trial augmentation, we address this bias–variance trade-off by explicitly prioritizing unbiasedness. A key insight from us was that the transportability condition could be decoupled from the identification strategy and used solely to improve estimation with external data. This ensures that no bias is introduced, regardless of whether transportability holds. The "safety" of our approach thus lies in this guarantee of no bias, in particular because we further provide conditions under which the variance cannot worsen either. The limitation of this stronger guarantee however is that the variance reduction we achieve is smaller than what could be obtained by directly incorporating transportability into the identification strategy.

Looking ahead, we believe it is worth exploring other problem settings where conditions that improve estimation can be decoupled from the underlying identification strategy, as doing so may help advance the development of new methods for safer causal inference.

8.1.4. In what sense can we ever do causal inference safely?

Most researchers agree that our ability to do causal inference is limited, if not impossible, unless we make assumptions: no causes in, no causes out, as Cartwright (1989, Ch. 2) famously put it. Even randomized experiments, often considered the "gold standard" of causal inference, rely on assumptions which need to be justified. Assumptions of consistency and no interference can fail even in the presence of randomization, and the extent to which experimental results can generalize to broader contexts is often up for debate (Kahan *et al.*, 2015). Assumptions are therefore unavoidable in causal inference, yet incorrect assumptions can lead us to non-trustworthy causal claims; so does this make safer causal inference a hopeless pursuit?

Ogburn and Shpitser (2021) propose a "cautious" approach for causal inference, where assumptions must be explicitly stated and justified about one's data before proceeding with identifying the treatment effect one is interested in estimating. From this perspective, even when taking the stance that our assumptions are untestable, tools that encourage

reflection on these assumptions could contribute towards making causal inference safer. Some assumptions, such as positivity, can be assessed from data (Oberst *et al.*, 2020; Petersen *et al.*, 2012), and in this dissertation we have further broadened the range of methods for falsifying other causal identification assumptions.

In addition, we have seen that for certain tasks we may already be safer than expected, even when some of our assumptions are violated. For example, in Chapter 7 we observed that when ignoring interference, although this led to biased estimation of Qini curves, we were often still able to correctly rank multiple treatment policies from best to worst. Further discovering and formalizing this kind of scenarios could provide great practical value.

Finally, even before stating our assumptions, there is another possibility to come closer to making causal inference safer. By directly improving the study design and data collection, it might be possible to support certain identification assumptions which we would need for identification of treatment effects. For instance, the risk of unmeasured confounding in observational studies could be reduced by measuring a richer set of covariates. Reducing the risk of transportability violations, which depend on the presence of shifted unmeasured effect modifiers, could be addressed in a similar way. In other cases, design choices such as cluster randomization can help mitigate bias from interference (Holtz *et al.*, 2025).

8.2. Limitations

8.2.1. Empirical evaluation of methods

In our work, we faced a challenge common to much of methodological research in causal inference: evaluating methods using real-world data is difficult because the ground truth causal effects are unknown. For this reason, much of the experimental evaluation in each chapter was performed with synthetic data, where the ground truth is known. In addition to this, we used semi-synthetic datasets in Chapters 2 and 3, constructed by combining a real-world dataset with simulated variables to retain some realism while introducing a known ground truth. With these datasets we had realistic distributions of the pre-treatment covariates, including covariate shifts between different data sources. On the other hand, we used hand-made treatment and outcome mechanisms for simulating the treatment and outcome variables, limiting our ability to learn how our methods would be influenced by model misspecification of those mechanisms in real-world settings.

In Chapters 5 and 6, we worked directly with real-world datasets. While real-world data can highlight useful differences in the outputs of methods, it does not give a clear-cut answer for which method has the correct output due to the absence of a known ground truth. In Chapter 6, we tried to address this challenge by constructing proxies for the true treatment effect, thereby circumventing the lack of ground-truth. But while these proxies provided unbiased estimates of the treatment effect, they suffered from high variance, which again limited the conclusions we could draw.

While none of the datasets above on their own allow us to draw definitive conclusions of our methods' performances, they remain valuable as sanity checks: if a method fails or behaves unexpectedly in a controlled setting, there is little reason to trust it on more complex data. For instance, Poinsot *et al.* (2025) emphasize that simulation studies can play an important role in systematically exploring challenging edge cases of a method. In our case, such edge cases often involved violations of key assumptions, such as the absence of unmeasured confounding, transportability, and interference. While it is naturally difficult to capture all possible violations that might occur in practice, an interesting avenue for new sanity checks could also be to use data from tightly controlled physical experiments, where the ground truth is determined by physical laws (Gamella, Bing & Runge, 2025; Gamella, Peters & Bühlmann, 2025), potentially providing more realistic yet controlled edge cases for evaluating methods.

8.2.2. Asymptotic guarantees in trial augmentation

Our safety claims in Part Two are asymptotic: the consistency and optimality of our estimator are guaranteed only in the limit where the proportion of trial size n_1 to external dataset size n_0 converges to a fixed constant, $n_1/n_0 \to q \in (0,1)$. However, somewhat ironically, our motivation is to apply these methods in trials that are typically small, where such asymptotic results likely will not hold. While this point warrants a serious reflection on the usefulness of these results, in some sense, they represent the best we can achieve. Finite-sample guarantees are generally more difficult to establish, and there are fewer theoretical tools available to obtain them. For this reason, we relied on simulation studies of trial augmentation to better understand finite-sample behavior and we encouragingly found that our observations aligned well with the asymptotic theoretical guarantees. Still, the possibility of deriving stronger finite-sample guarantees remains an important direction for future work. One potential avenue is through the Rao–Blackwell theorem, which is frequently used for establishing guarantees on finite-sample efficiency; see examples such as Chen *et al.* (2010) and Karlsson *et al.* (2022).

8.2.3. Proof techniques in falsification results

In Part One, we approached the question of falsification using two very different proof techniques, both aimed at establishing similar claims. In Chapters 2 and 4, we used hierarchical causal models together with graphical arguments commonly found in the causal discovery literature (Glymour *et al.*, 2019). The benefit of this approach was that it is largely nonparametric, relying on few assumptions about the underlying models. On the downside, our conclusions were limited to a particular class of graphs and did not account for more complex structures, such as those that can lead to M-bias or selection bias. In Chapter 3, we took a different route, avoiding graphs entirely and performing an algebraic proof with linear parametric models. This helped to address some of the limitations of the hierarchical approach, but at the same time introduced the parametric assumptions that the graphical method had avoided. Another unexpected benefit of the

second proof technique is that, unlike the the graph-based approach, our results requires no assumptions on the distribution of the observed pre-treatment covariates.

In a sense, arriving at similar conclusions using two different techniques based on different assumptions helped us be more assertive in our claims, but it also leaves open the question of how they might be related or which proof technique is better suited for generalizing to new settings and falsification strategies. While the linear model can be viewed as a special case of the hierarchical causal model, we did not make this connection explicit – and exploring it might offer further insights.

8.3. Future work

8.3.1. Hierarchical models in causal inference

A key tool we found useful in our research is the hierarchical causal model. When combined with the ICM assumption, the hierarchical causal model provided new testable implications for assumptions required to identify treatment effects. Hierarchical causal models have received growing attention in the literature (Guo *et al.*, 2023, 2024; Jensen *et al.*, 2020; Weinstein & Blei, 2024), and it is interesting to think about their wider potential. For instance, while we focused on falsifying the unconfoundedness assumption in Chapters 2 and 3, we also showed in Chapter 4 that hierarchical causal models can be used to test whether a mediator satisfies the front-door criterion or whether a variable meets the requirements of an instrumental variable. However, more sample-efficient procedures like the one developed in Chapter 3 are still needed for these cases.

One avenue worth exploring is whether hierarchical models can be used for partial identification of treatment effects under the presence of unmeasured confounding, for instance by borrowing techniques from flexible frameworks to compute bounds, as in Padh *et al.* (2023), or from data-fusion-based approaches like those in De Bartolomeis, Martinez *et al.* (2024). Another research direction of interest is to investigate whether hierarchical causal models lead to novel ways of integrating data from multiple sources for more efficient treatment effect estimation.

8.3.2. Trial design and different analyses for trial augmentation

In our work on trial augmentation, we have so far focused on a relatively simple setting with a binary treatment administered at a fixed time, with an outcome measured at a fixed follow-up. While this scenario is common, it is also, in some sense, the easiest context in which to develop new methods. In practice, however, more complex settings frequently arise. For example, researchers often conduct survival analyses to study the timing of events after treatment, or also consider time-varying treatment strategies. The field is gradually moving towards exploring trial augmentation in this direction, and some methods for integrating external data in survival analyses exist (C. Huang *et al.*, 2023).

However, these approaches do not yet address the safety concerns we emphasize in Part Two, where the priority is unbiased estimation of treatment effects even when integrating possibly misaligned external data.

A closely related issue concerns how to allocate incoming patients to treatment and control groups when external data are available, a problem that belongs to the broader topic of adaptive trial designs (Chow & Chang, 2008). For instance, if the external data are well-aligned with the trial population but predominantly contain individuals receiving the control treatment, it may be advantageous to adjust the trial allocation to assign more participants to the treatment group inside the trial. This kind of rebalancing can improve efficiency, but it remains an open question whether the same safety guarantees we discussed earlier can be maintained.

8.3.3. Combining information from multiple sources

A common theme in all but one chapter of this dissertation is that we studied problem settings where we have data from multiple distinct sources. Unlike the typical scenario of analyzing data from a single source in isolation, integrating information from several sources presents additional challenges by having to account for the differences between each source. Despite these challenges, we have seen that combining information from multiple sources unlocks new opportunities. In this dissertation, we proposed methods for falsifying identification assumptions and improving the efficiency of treatment effect estimation in randomized trials; both of which would not have been possible without access to multiple data sources. More broadly, combining data from multiple sources has also been used to enable new forms of identification strategies (Athey *et al.*, 2025), extend inferences to new populations beyond the original study sample (Dahabreh & Hernán, 2019; Westreich *et al.*, 2017), and improve causal discovery algorithms (B. Huang *et al.*, 2020; Mooij *et al.*, 2020).

As the availability data keeps increasing in many domains, additional opportunities are emerging for the types of causal questions we can ask and the ways we can answer them. It is therefore becoming increasingly important to develop methods that allow us to use all available data; integrating information not only across study populations but also across different modalities, including tabular data, images, and text. In particular, the growing popularity of foundation models with general-purpose capabilities trained at large scale (Bommasani, 2021) may also offer new potential for improving causal analyses. An open question however is how to leverage these sources of information while preserving the properties we care the most about, such as unbiased identification and statistically valid inferences. Going forward, as we start addressing these challenges, we should ensure that combining information from multiple sources becomes the norm in causal inference, allowing us to make the most of all available data.

8.4. Concluding remarks

Throughout this dissertation, our motivation was to find new ways to avoid a common pitfall when doing causal inference, which is that the quantity we are estimating from data may not correspond to the actual treatment effect we are interested in. The underlying root of this problem is that our causal identification strategy relies on untestable assumptions; hence if those assumptions are violated, it is in general impossible for us to know that our identification strategy has failed. This problem can ultimately lead to untrustworthy causal claims.

The central question thus was whether we could make causal inference safer by detecting or dealing with violations of identification assumptions. To address this, we made contributions along two main lines. First, we investigated and proposed methods to falsify causal identification assumptions when having data from multiple sources under certain conditions. Second, we revisited two specific problem settings in causal inference and demonstrated that it is possible to develop methods that either remove the dependence on strong assumptions. Although these approaches may never be entirely bulletproof against everything that can go wrong, together they make causal inference safer by reducing dependence on strong assumptions and helping to avoid the very pitfall that motivated this dissertation.

References

Almond, D., Chay, K. Y., & Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, *120*(3), 1031–1083.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669–674.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444–455.

Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, *11*(4), 1912.

Asiaee, A., Di Gravio, C., Beck, C., Mei, Y., Pal, S., & Huling, J. D. (2025). Improving precision of rct-based cate estimation using data borrowing with double calibration. *arXiv preprint arXiv:2306.17478*.

Asiaee, A., Di Gravio, C., Mei, Y., & Huling, J. D. (2023). Leveraging observational data for efficient cate estimation in randomized controlled trials. *arXiv preprint arXiv:2306.17478*.

Athey, S., Chetty, R., & Imbens, G. (2020). Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes. *arXiv preprint arXiv:2006.09676*.

Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2025). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *Review of Economic Studies*, rdaf087.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, *31*(2), 3–32.

Atienza, D., Bielza, C., & Larrañaga, P. (2022). Pybnesian: An extensible python package for bayesian networks. *Neurocomputing*, *504*, 204–209.

Baba, K., Shibata, R., & Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, *46*(4), 657–664.

Bajari, P., Burdick, B., Imbens, G. W., Masoero, L., McQueen, J., Richardson, T. S., & Rosen, I. M. (2023). Experimental design in marketplaces. *Statistical Science*, *38*(3), 458–476.

Balke, A., & Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty proceedings 1994* (pp. 46–54). Elsevier.

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

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

Barnatchez, K., Josey, K. P., Hejazi, N. S., Shepherd, B. E., Parmigiani, G., & Nethery, R. (2025). Efficient estimation of causal effects under two-phase sampling with error-prone outcome and treatment measurements. *arXiv preprint arXiv:2506.21777*.

Bhattacharya, R., & Nabi, R. (2022). On testability of the front-door model via verma constraints. *Uncertainty in Artificial Intelligence*, 202–212.

Bliek, L., Guijt, A., Karlsson, R., Verwer, S., & De Weerdt, M. (2023). Benchmarking surrogate-based optimisation algorithms on expensive black-box functions. *Applied Soft Computing*, *147*, 110744.

Bokelmann, B., & Lessmann, S. (2024). Improving uplift model evaluation on randomized controlled trial data. *European Journal of Operational Research*, *313*(2), 691–707.

Bommasani, R. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference: Theory and methods* (Vol. 591). Springer.

Brand, J. E., & Davis, D. (2011). The impact of college education on fertility: Evidence for heterogeneous effects. *Demography*, *48*, 863–887.

Breiman, L. (1996). Stacked regressions. *Mach. Learn.*, *24*(1), 49–64.

Burauel, P. F. (2023). Evaluating instrument validity using the principle of independent mechanisms. *Journal of Machine Learning Research*, *24*(176), 1–56.

Cao, W., Tsiatis, A. A., & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, *96*(3), 723–734.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, *84*(4), 772–793.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Clarendon Press.

Cartwright, N. (2004). Causation: One word, many things. *Philosophy of Science*, *71*(5), 805–819.

Chen, Y., Wiesel, A., Eldar, Y. C., & Hero, A. O. (2010). Shrinkage algorithms for mmse covariance estimation. *IEEE transactions on signal processing*, *58*(10), 5016–5029.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–c68.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*.

Chow, S.-C., & Chang, M. (2008). Adaptive design methods in clinical trials–a review. *Orphanet journal of rare diseases*, *3*(1), 11.

Cinelli, C., Kumor, D., Chen, B., Pearl, J., & Bareinboim, E. (2019). Sensitivity analysis of linear structural causal models. *International conference on machine learning*, 1252–1261.

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

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, *22*(1), 173–203.

Curth, A., & Van der Schaar, M. (2021). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. *International Conference on Artificial Intelligence and Statistics*, 1810–1818.

Dahabreh, I. J., Haneuse, S. J. A., Robins, J. M., Robertson, S. E., Buchanan, A. L., Stuart, E. A., & 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., Hayward, R., & Kent, D. M. (2016). Using group data to treat individuals: Understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International journal of epidemiology*, *45*(6), 2184–2193.

Dahabreh, I. J., & Hernán, M. A. (2019). Extending inferences from a randomized trial to a target population. *European journal of epidemiology*, *34*(8), 719–722.

Dahabreh, I. J., Matthews, A., Steingrimsson, J. A., Scharfstein, D. O., & Stuart, E. A. (2024). Using trial and observational data to assess effectiveness: Trial emulation, transportability, benchmarking, and joint analysis. *Epidemiologic reviews*, *46*(1), 1–16.

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

Dahabreh, I. J., Robins, J. M., Haneuse, S. J., & Hernán, M. A. (2019). Generalizing causal inferences from randomized trials: Counterfactual and graphical identification. *arXiv preprint arXiv:1906.10792*.

Dahabreh, I. J., Robins, J. M., & Hernán, M. A. (2020). Benchmarking observational methods by comparing randomized trials and their emulations. *Epidemiology (Cambridge, Mass.)*, *31*(5), 614–619.

D'Amour, A. (2019). On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. *The 22nd international conference on artificial intelligence and statistics*, 3478–3486.

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

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., & Schölkopf, B. (2010). Inferring deterministic causal relations. *26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, 143–150.

Dasgupta, T., Pillai, N. S., & Rubin, D. B. (2014). Causal inference from 2k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *77*(4), 727–753.

Davidson, M., Emsley, R., Kramer, M., Ford, L., Pan, G., Lim, P., & Eerdekens, M. (2007). Efficacy, safety and early response of paliperidone extended-release tablets (paliperidone er): Results of a 6-week, randomized, placebo-controlled study. *Schizophrenia research*, *93*(1-3), 117–130.

de Finetti, B. (1937). La prévision : Ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, *7*(1), 1–68.

De Bartolomeis, P., Abad, J., Donhauser, K., & Yang, F. (2024). Detecting critical treatment effect bias in small subgroups. *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, 943–965.

De Bartolomeis, P., Abad, J., Wang, G., Donhauser, K., Duch, R. M., Yang, F., & Dahabreh, I. J. (2025). Efficient randomized experiments using foundation models. *arXiv preprint arXiv:2502.04262*.

De Bartolomeis, P., Martinez, J. A., Donhauser, K., & Yang, F. (2024). Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. *International Conference on Artificial Intelligence and Statistics*, 1045–1053.

Deckler, E., Ferland, M., Brazis, S., Mayer, M. R., Carlson, M., & Kantrowitz, J. T. (2022). Challenges and strategies for the recruitment of patients with schizophrenia in a research setting. *International Journal of Neuropsychopharmacology*, *25*(11), 924–932.

Demirel, I., De Brouwer, E., Hussain, Z. M., Oberst, M., Philippakis, A. A., & Sontag, D. (2024). Benchmarking observational studies with experimental data under right-censoring. *International Conference on Artificial Intelligence and Statistics*, 4285–4293.

Devereaux, P., & Yusuf, S. (2003). The evolution of the randomized controlled trial and its role in evidence-based decision making. *Journal of internal medicine*, *254*(2), 105–113.

Di Angelantonio, E., Bhupathiraju, S. N., Wormser, D., Gao, P., Kaptoge, S., De Gonzalez, A. B., Cairns, B. J., Huxley, R., Jackson, C. L., Joshy, G., et al. (2016). Body-mass index and all-cause mortality: Individual-participant-data meta-analysis of 239 prospective studies in four continents. *The Lancet*, *388*(10046), 776–786.

Díaz, I., & van der Laan, M. J. (2013). Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The international journal of biostatistics*, *9*(2), 149–160.

Eckles, D., Karrer, B., & Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, *5*(1), 20150021.

Faller, P. M., Vankadara, L. C., Mastakouri, A. A., Locatello, F., & Janzing, D. (2024). Self-compatibility: Evaluating causal discovery without ground truth. In S. Dasgupta, S. Mandt & Y. Li (Eds.), *Proceedings of the 27th international conference on artificial intelligence and statistics* (pp. 4132–4140, Vol. 238). Pmlr.

Fawkes, J., O'Riordan, M., Vlontzos, A., Corcoll, O., & Gilligan-Lee, C. M. (2025). The hardness of validating observational studies with experimental data. *International Conference on Artificial Intelligence and Statistics*, 1819–1827.

Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I. S., & van der Schaar, M. (2024). Causal machine learning for predicting treatment outcomes. *Nature Medicine*, *30*(4), 958–968.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver; Boyd.

Fisher, R. A. (1958). Cancer and smoking. *Nature*, *182*(4635), 596–596.

Foster, D. J., & Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics*, *51*(3), 879–908.

Gagnon-Bartsch, J. A., Sales, A. C., Wu, E., Botelho, A. F., Erickson, J. A., Miratrix, L. W., & Heffernan, N. T. (2023). Precise unbiased estimation in randomized experiments using auxiliary observational data. *Journal of Causal Inference*, *11*(1), 20220011.

Galilei, G. (1638). *Discorsi e dimostrazioni matematiche intorno a due nuove scienze* [First edition published in Leiden by Elsevier; English translation: *Two New Sciences*, translated by Stillman Drake, University of Wisconsin Press, 1974]. Elsevier.

Gamella, J. L., Bing, S., & Runge, J. (2025). Sanity checking causal representation learning on a simple real-world system. *arXiv preprint arXiv:2502.20099*.

Gamella, J. L., Peters, J., & Bühlmann, P. (2025). Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, *7*(1), 107–118.

Gao, C., Yang, S., Shan, M., Ye, W., Lipkovich, I., & Faries, D. (2023). Integrating randomized placebo-controlled trial data with external controls: A semiparametric approach with selective borrowing. *arXiv preprint arXiv:2306.16642*.

Gelman, A. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd). CRC press.

Ghassami, A., Kiyavash, N., Huang, B., & Zhang, K. (2018). Multi-domain causal structure learning in linear systems. *Advances in neural information processing systems*, *31*.

Ghosal, S., & Van der Vaart, A. (2017). *Fundamentals of nonparametric bayesian inference* (Vol. 44). Cambridge University Press.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

Goldenberg, D., Albert, J., Bernardi, L., & Estevez, P. (2020). Free lunch! retrospective uplift modeling for dynamic promotions recommendation within roi constraints. *Proceedings of the 14th ACM Conference on Recommender Systems*, 486–491.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Multilevel modelling of medical data. *Statistics in medicine*, *21*(21), 3291–3315.

Goplerud, M., Imai, K., & Pashley, N. E. (2025). Estimating heterogeneous causal effects of high-dimensional treatments: Application to conjoint analysis. *arXiv preprint arXiv:2201.01357*.

Graybill, F. A., & Deal, R. (1959). Combining unbiased estimators. *Biometrics*, *15*(4), 543–550.

Greenland, S. (1996). Basic methods for sensitivity analysis of biases. *International journal of epidemiology*, *25*(6), 1107–1116.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical science*, *14*(1), 29–46.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., & Schölkopf, B. (2005). Kernel methods for measuring independence. *Journal of Machine Learning Research*, *6*, 2075–2129.

Grimmett, G., & Stirzaker, D. (2020). *Probability and random processes*. Oxford university press.

Guo, S., Tóth, V., Schölkopf, B., & Huszár, F. (2023). Causal de finetti: On the identification of invariant causal structure in exchangeable data. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 36463–36475.

Guo, S., Zhang, C., Mohan, K., Huszár, F., & Schölkopf, B. (2024). Do finetti: On causal effects for exchangeable data. *Advances in Neural Information Processing Systems*, *37*, 127317–127345.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, iii–115.

Hatt, T., Berrevoets, J., Curth, A., Feuerriegel, S., & van der Schaar, M. (2022). Combining Observational and Randomized Data for Estimating Heterogeneous Treatment Effects. *arXiv preprint arXiv:2202.12891*.

Heard, N. A., & Rubin-Delanchy, P. (2018). Choosing between methods of combining-values. *Biometrika*, *105*(1), 239–246.

Hernan, M. A., & Robins, J. M. (2023). *Causal inference*. CRC Press.

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Willett, W. C., Manson, J. E., & Robins, J. M. (2008). Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*, *19*(6), 766–779.

Hernán, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, *60*(7), 578–586.

Hernán, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, *22*(3), 368–377.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Holtz, D., Lobel, F., Lobel, R., Liskovich, I., & Aral, S. (2025). Reducing interference bias in online marketplace experiments using cluster randomization: Evidence from a pricing meta-experiment on airbnb. *Management Science*, *71*(1), 390–406.

Huang, B., Zhang, K., Gong, M., & Glymour, C. (2020). Causal discovery from multiple data sets with non-identical variable sets. *Proceedings of the AAAI conference on artificial intelligence*, *34*(06), 10153–10161.

Huang, C., Wei, K., Wang, C., Yu, Y., & Qin, G. (2023). Covariate balance-related propensity score weighting in estimating overall hazard ratio with distributed survival data. *BMC medical research methodology*, *23*(1), 233.

Huang, M., Egami, N., Hartman, E., & Miratrix, L. (2023). Leveraging population outcomes to improve the generalization of experimental results: Application to the jtpa study. *The Annals of Applied Statistics*, *17*(3), 2139–2164.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842.

Hussain, Z., Shih, M.-C., Oberst, M., Demirel, I., & Sontag, D. (2023). Falsification of internal and external validity in observational studies via conditional moment

restrictions. *International Conference on Artificial Intelligence and Statistics*, 5869–5898.

Hussain, Z. M., Oberst, M., Shih, M.-C., & Sontag, D. (2022). Falsification before extrapolation in causal effect estimation. *Advances in Neural Information Processing Systems*, *35*, 6161–6174.

Ilse, M., Forré, P., Welling, M., & Mooij, J. M. (2022). Combining Interventional and Observational Data Using Causal Reductions. *arXiv preprint arXiv:2103.04786*.

Imai, K., & Li, M. L. (2023). Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association*, *118*(541), 242–256.

Imbens, G., Kallus, N., Mao, X., & Wang, Y. (2022). Long-term Causal Inference Under Persistent Confounding via Data Combination. *arXiv preprint arXiv:2202.07234*.

Imbens, G., & Rubin, D. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jahanshahi, M., Gregg, K., Davis, G., Ndu, A., Miller, V., Vockley, J., Ollivier, C., Franolic, T., & Sakai, S. (2021). The use of external controls in fda regulatory decision making. *Therapeutic Innovation & Regulatory Science*, *55*(5), 1019–1035.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., & Schölkopf, B. (2012). Information-geometric approach to inferring causal directions. *Artificial Intelligence*, *182*, 1–31.

Janzing, D., & Schölkopf, B. (2010). Causal inference using the algorithmic markov condition. *IEEE Trans. Inf. Theor.*, *56*(10), 5168–5194.

Janzing, D., & Schölkopf, B. (2018). Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, *6*(1), 20170013.

Jensen, D., Burroni, J., & Rattigan, M. (2020). Object conditioning for causal inference. *Uncertainty in Artificial Intelligence*, 1072–1082.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. *Sigir'05*, 154–161.

Kahan, B. C., Rehal, S., & Cro, S. (2015). Risk of selection bias in randomised trials. *Trials*, *16*(1), 405.

Kallus, N., Puli, A. M., & Shalit, U. (2018). Removing Hidden Confounding by Experimental Grounding. *Advances in Neural Information Processing Systems*, *31*.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539.

Karlsson, R., Akker, B. v. d., Moraes, F., Proença, H. M., & Krijthe, J. H. (2025). Qini curve estimation under clustered network interference. *arXiv preprint arXiv:2502.20097*.

Karlsson, R., Bliek, L., Verwer, S., & de Weerdt, M. (2020). Continuous surrogate-based optimization algorithms are well-suited for expensive discrete problems. *BNAIC/BENELEARN (Selected Papers)*, 48–63.

Karlsson, R., Creastă, Ș., & Krijthe, J. H. (2023). Putting causal identification to the test: Falsification using multi-environment data. *Causal Representation Learning Workshop at NeurIPS 2023*.

Karlsson, R., De Bartolomeis, P., Dahabreh, I. J., & Krijthe, J. H. (2026). Robust estimation of heterogeneous treatment effects in randomized trials leveraging external data [Forthcoming]. *International Conference on Artificial Intelligence and Statistics*.

Karlsson, R., & Krijthe, J. (2023). Detecting hidden confounding in observational data using multiple environments. *Advances in Neural Information Processing Systems*, *36*, 44280–44309.

Karlsson, R., & Krijthe, J. (2025). Falsification of unconfoundedness by testing independence of causal mechanisms. *Proceedings of the 42nd International Conference on Machine Learning*, *267*, 29128–29147.

Karlsson, R., Wang, G., De Bartolomeis, P., Krijthe, J. H., & Dahabreh, I. J. (2026). Robust integration of external control data in randomized trials [Forthcoming]. *Biometrics*.

Karlsson, R., Willbo, M., Hussain, Z. M., Krishnan, R. G., Sontag, D., & Johansson, F. (2022). Using time-series privileged information for provably efficient learning of prediction models. *International Conference on Artificial Intelligence and Statistics*, 5459–5484.

Kédagni, D., & Mourifie, I. (2017). Generalized instrumental inequalities: Testing the iv independence assumption. *Available at SSRN 2692274*.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, *17*(2), 3008–3049.

Kravitz, R. L., Duan, N., & Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, *82*(4), 661–687.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The quarterly journal of economics*, *114*(2), 497–532.

Kügelgen, J., Mey, A., Loog, M., & Schölkopf, B. (2020). Semi-supervised learning, causality, and the conditional cluster assumption. *Conference on uncertainty in artificial intelligence*, 1–10.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, *116*(10), 4156–4165.

Kuroki, M., & Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, *101*(2), 423–437.

Lagakos, S. W., et al. (2006). The challenge of subgroup analyses-reporting without distorting. *New England Journal of Medicine*, *354*(16), 1667.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate behavioral research*, *50*(3), 265–284.

Leitgöb, H. (2020). *Analysis of rare events*. SAGE Publications Limited.

Li, F., Buchanan, A. L., & Cole, S. R. (2022). Generalizing trial evidence to target populations in non-nested designs: Applications to aids clinical trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *71*(3), 669–697.

Li, X., Miao, W., Lu, F., & Zhou, X.-H. (2023). Improving efficiency of inference in clinical trials with external control data. *Biometrics*, *79*(1), 394–403.

Liao, L. D., Højbjerre-Frandsen, E., Hubbard, A. E., & Schuler, A. (2023). Prognostic adjustment with efficient estimators to unbiasedly leverage historical data in randomized trials. *arXiv preprint arXiv:2305.19180*.

Lind, J. (1772). *A treatise on the scurvy: In three parts. containing an inquiry into the nature, causes, and cure, of that disease. together with a critical and chronological view of what has been published on the subject*. S. Crowder.

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

Loeb, L. A., Emster, V. L., Warner, K. E., Abbotts, J., & Laszlo, J. (1984). Smoking and lung cancer: An overview. *Cancer research*, *44*(12_part_1), 5940–5958.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, *30*.

Lu, X., & White, H. (2014). Robustness checks and robustness tests in applied economics. *Journal of econometrics*, *178*, 194–206.

Luedtke, A. R., Carone, M., & van der Laan, M. J. (2019). An omnibus non-parametric test of equality in distribution for unknown functions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *81*(1), 75–99.

Mameche, S., Kaltenpoth, D., & Vreeken, J. (2024). Learning causal models under independent changes. *Advances in Neural Information Processing Systems*, *36*.

Mameche, S., Vreeken, J., & Kaltenpoth, D. (2024). Identifying confounding from causal mechanism shifts. *International Conference on Artificial Intelligence and Statistics*, 4897–4905.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, *16*(1), S1–s23.

Marder, S. R., Kramer, M., Ford, L., Eerdekens, E., Lim, P., Eerdekens, M., & Lowy, A. (2007). Efficacy and safety of paliperidone extended-release tablets: Results of a 6-week, randomized, placebo-controlled study. *Biological Psychiatry*, *62*(12), 1363–1370.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V., & Klungel, O. H. (2006). Instrumental variables: Application and limitations. *Epidemiology*, 260–267.

Mccullagh, P., & Nelder, J. (1989). *Generalized linear models*. CRC press.

McDonald, J. H. (2014). *Handbook of biological statistics* (3rd). Sparky House Publishing.

Meek, C. (1995). Strong completeness and faithfulness in bayesian networks. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 411–418.

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

Mooij, J. M., Magliacane, S., & Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, *21*(99), 1–108.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., & Schölkopf, B. (2016). Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, *17*(32), 1–102.

Morzywolek, P., Decruyenaere, J., & Vansteelandt, S. (2023). On weighted orthogonal learners for heterogeneous treatment effects. *arXiv preprint arXiv:2303.12687*.

Mueller, M., D'Addario, M., Egger, M., Cevallos, M., Dekkers, O., Mugglin, C., & Scott, P. (2018). Methods to systematically review and meta-analyse observational

studies: A systematic scoping review of recommendations. *BMC medical research methodology*, *18*(1), 1–18.

National Institute of Mental Health. (2023). Schizophrenia, U.S. Department of Health and Human Services, National Institutes of Health. [Retrieved May 3, 2023, from https://www.nimh.nih.gov/health/topics/schizophrenia].

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Newton, I. (1672). A letter of mr. isaac newton, professor of the mathematicks in the university of cambridge; containing his new theory about light and colors: Sent by the author to the publisher from cambridge, febr. 6. 1671/72; in order to be communicated to the r. society [Originally communicated February 1671/72 (Julian calendar)]. *Philosophical Transactions of the Royal Society of London*, *6*, 3075–3087.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, 1–51.

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

Oberst, M., Johansson, F., Wei, D., Gao, T., Brat, G., Sontag, D., & Varshney, K. (2020). Characterization of overlap in observational studies. *International Conference on Artificial Intelligence and Statistics*, 788–798.

Ogburn, E. L., & Shpitser, I. (2021). Causal modelling: The two cultures. *Observational Studies*, *7*(1), 179–183.

Ogburn, E. L., Shpitser, I., & Lee, Y. (2020). Causal inference, social networks and chain graphs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*(4), 1659–1676.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, *37*(2), 187–204.

Padh, K., Zeitler, J., Watson, D., Kusner, M., Silva, R., & Kilbertus, N. (2023). Stochastic causal programming for bounding treatment effects. *Conference on Causal Learning and Reasoning*, 142–176.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan kaufmann.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 435–443.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd). Cambridge University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perry, R., Von Kügelgen, J., & Schölkopf, B. (2022). Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, *35*, 10904–10917.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(5), 947–1012.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms* (1st). MIT Press.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & Van Der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, *21*(1), 31–54.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of chronic diseases*, *29*(3), 175–188.

Poinsot, A., Panayiotou, P., Leite, A., Chesneau, N., Şimşek, Ö., & Schoenauer, M. (2025). Position: Causal machine learning requires rigorous synthetic experiments for broader adoption. *arXiv preprint arXiv:2508.08883*.

Racine, J. S., Hart, J., & Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, *25*(4), 523–544.

Radcliffe, N. (2007). Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, 14–21.

Reddy, A. G., & Balasubramanian, V. N. (2024). Detecting and measuring confounding using causal mechanism shifts. *Advances in Neural Information Processing Systems*.

Reichenbach, H. (1956). *The direction of time*. Dover Publications.

Riley, R., Lambert, P., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *British Medical Journal (International edition)*, *340*, c221.

Robins, J. M., & Greenland, S. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, *95*(450), 431–435.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Robins, J. M., Rotnitzky, A., & Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials* (pp. 1–94). Springer.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*(427), 846–866.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, *45*(2), 212–218.

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

Rößler, J., & Schoder, D. (2022). Bridging the gap: A systematic benchmarking of uplift modeling and heterogeneous treatment effects methods. *Journal of Interactive Marketing*, *57*(4), 629–650.

Rothenhäusler, D. (2024). Model selection and inference for estimation of causal parameters. *Electronic Journal of Statistics*, *18*(2), 5449–5483.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., & Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *83*(2), 215–246.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, *75*(371), 591–593.

Saito, Y., & Yasui, S. (2020). Counterfactual cross-validation: Stable model selection procedure for causal inference models. *International Conference on Machine Learning*, 8398–8407.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. (2012). On causal and anticausal learning. *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 459–466.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612–634.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Schuler, A., Walsh, D., Hall, D., Walsh, J., Fisher, C., for Alzheimer's Disease, C. P., Initiative, A. D. N., & Study, A. D. C. (2022). Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, *18*(2), 329–356.

Schweisthal, J., Frauen, D., Van Der Schaar, M., & Feuerriegel, S. (2024). Meta-learners for partially-identified treatment effects across multiple environments. *Forty-first International Conference on Machine Learning*.

Shah, R. D., & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, *48*(3), 1514–1538.

Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. *International conference on machine learning*, 3076–3085.

Shi, C., Veitch, V., & Blei, D. M. (2021). Invariant representation learning for treatment effect estimation. *Uncertainty in Artificial Intelligence*, 1546–1555.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, *90*(2), 227–244.

Shyr, C., Ren, B., Patil, P., & Parmigiani, G. (2023). Multi-study r-learner for estimating heterogeneous treatment effects across studies using statistical machine learning. *arXiv e-prints*, arXiv–2306.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, *101*(476), 1398–1407.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Squires, C., Seigal, A., Bhate, S. S., & Uhler, C. (2023). Linear causal disentanglement via interventions. *International conference on machine learning*, 32540–32560.

Stefanski, L. A., & Boos, D. D. (2002). The calculus of m-estimation. *The American Statistician*, *56*(1), 29–38.

Strobl, E. V., Zhang, K., & Visweswaran, S. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, *7*(1), 20180017.

Su, Z., & Henckel, L. (2022). A robustness test for estimating total effects with covariate adjustment. *The 38th Conference on Uncertainty in Artificial Intelligence*, 1886–1895.

Sverdrup, E., Wu, H., Athey, S., & Wager, S. (2025). Qini curves for multi-armed treatment rules. *Journal of Computational and Graphical Statistics*, *34*(3), 948–960.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, *101*(476), 1619–1637.

Tchetgen, E. J. T., & VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical methods in medical research*, *21*(1), 55–75.

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. (2017). Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*, *45*(6), 1887–1894.

Tippett, L. H. C. (1931). *The methods of statistics: An introduction mainly for workers in the biological sciences*. Williams & Norgate.

Tsamardinos, I., & Borboudakis, G. (2010). Permutation testing improves bayesian network learning. In J. L. Balcázar, F. Bonchi, A. Gionis & M. Sebag (Eds.), *Machine learning and knowledge discovery in databases* (pp. 322–337). Springer Berlin Heidelberg.

Ugander, J., Karrer, B., Backstrom, L., & Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 329–337.

Ung, L., Wang, G., Haneuse, S., Hernan, M. A., & Dahabreh, I. J. (2024). Combining an experimental study with external data: Study designs and identification strategies. *arXiv preprint arXiv:2406.03302*.

Valancius, M., Pang, H., Zhu, J., Cole, S. R., Funk, M. J., & Kosorok, M. R. (2023). A causal inference framework for leveraging external controls in hybrid trials. *arXiv preprint arXiv:2305.08969*.

van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B., & Lesaffre, E. (2018). Including historical data in the analysis of clinical trials: Is it worth the effort? *Statistical methods in medical research*, *27*(10), 3167–3182.

Van Der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The international journal of biostatistics*, *2*(1).

van der Laan, M., Qiu, S., & van der Laan, L. (2024). Adaptive-tmle for the average treatment effect based on randomized controlled trial augmented with real-world data. *arXiv preprint arXiv:2405.07186*.

VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the e-value. *Annals of internal medicine*, *167*(4), 268–274.

VanderWeele, T. J., Tchetgen, E. J. T., & Halloran, M. E. (2012). Components of the indirect effect in vaccine trials: Identification of contagion and infectiousness effects. *Epidemiology*, *23*(5), 751–761.

VanderWeele, T. J., Tchetgen, E. J. T., & Halloran, M. E. (2015). Interference and sensitivity analysis. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *29*(4), 687.

Ventz, S., Khozin, S., Louv, B., Sands, J., Wen, P. Y., Rahman, R., Comment, L., Alexander, B. M., & Trippa, L. (2022). The design and evaluation of hybrid controlled trials that leverage external data and randomization. *Nature communications*, *13*(1), 5783.

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

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

Wang, G., Poulin-Costello, M., Pang, H., Zhu, J., Helms, H.-J., Reyes-Rivera, I., Platt, R. W., Pang, M., & Koukounari, A. (2024). Evaluating hybrid controls methodology in early-phase oncology trials: A simulation study based on the morpheus-uc trial. *Pharmaceutical Statistics*, *23*(1), 31–45.

Wang, Y., & Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, *114*(528), 1574–1596.

Weinstein, E. N., & Blei, D. M. (2024). Hierarchical causal models. *arXiv preprint arXiv:2401.05330*.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 9.

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

Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in medicine*, *33*(5), 721–737.

Woodward, J. (2000). Explanation and invariance in the special sciences. *The British journal for the philosophy of science*, *51*(2), 197–254.

Word, E., Johnston, J., Bain, H. P., Fulton, B. D., Zaharias, J. B., Achilles, C. M., Lintz, M. N., Folger, J., & Breda, C. (1990). The state of tennessee's student/teacher achievement ratio (star) project. *Tennessee Board of Education*.

Wright, S. (1934). The method of path coefficients. *The annals of mathematical statistics*, *5*(3), 161–215.

Wu, L., & Yang, S. (2022). Integrative *R*-learner of heterogeneous treatment effects combining experimental and observational studies. *Conference on Causal Learning and Reasoning*, 904–926.

Wu, L., & Yang, S. (2023). Transfer learning of individualized treatment rules from experimental to real-world data. *Journal of Computational and Graphical Statistics*, *32*(3), 1036–1045.

Wynder, E. L. (1997). Tobacco as a cause of lung cancer: Some reflections. *American Journal of Epidemiology*, *146*(9), 687–694.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., & Wager, S. (2024). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 1–14.

Yang, S., Gao, C., Zeng, D., & 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, K., Huang, B., Zhang, J., Glymour, C., & Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 1347–1353.

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 804–813.

Zhang, Y., & Imai, K. (2023). Individualized policy evaluation and learning under clustered network interference. *arXiv preprint arXiv:2311.02467*.

Zhao, Z., Bai, Y., Xiong, R., Cao, Q., Ma, C., Jiang, N., Wu, F., & Kuang, K. (2024). Learning individual treatment effects under heterogeneous interference in networks. *ACM Transactions on Knowledge Discovery from Data*, *18*(8), 1–21.

Zheng, Y., Huang, B., Chen, W., Ramsey, J., Gong, M., Cai, R., Shimizu, S., Spirtes, P., & Zhang, K. (2024). Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, *25*(60), 1–8.

Acknowledgements

Way back as a student, out of all courses, I failed the one in machine learning and yet, somewhat ironically, I chose to pursue a PhD in that very field. A few years later, I hope I could now pass that exam, but I am also no longer the same person who once sat it. Since moving to the Netherlands and spending a majority of my 20s here, I've grown tremendously both as a person and scientist. At the risk of forgetting someone because there are too many people to name, I would like to dedicate this section to some of the people who have helped shape my path in the past years.

First and foremost, I want to direct my most sincere appreciation to my daily supervisor **Jesse**. If in a counterfactual world I would have had another supervisor who was even half as kind and supportive, I would still have counted myself incredibly lucky. The freedom you have provided in letting me pursue my interests has allowed me to carve my own unique path as a scientist, while your attitude and approach to science has taught me to ask the right question for the right reason.

Next, I thank my promotor **Marcel** for his support and advice, making sure I was on the path to completing my degree and especially for helping with how to set up my research visit abroad. And when it comes to research visits abroad, at a moment when I felt stuck in my own research, I'm most grateful to **Issa** for welcoming me to Harvard and introducing me to a new and exciting research line, and ultimately becoming a mentor and close collaborator for the second half of my PhD. I would also like to thank my doctoral committee members **Robin Evans**, **Peter Grünwald**, **Julie Josse**, **Aad van der Vaart**, and **Mathijs de Weerdt** for taking the time to read my thesis and share their input.

From a period when I was uncertain whether to do a PhD in the first place, I want to thank **Laurens** and **Fredrik** respectively for giving me the chance to work with them during my master's, ultimately giving me very positive experiences of practicing science. I would also like to thank my other collaborators from the past years: **Guanbo**, **Piersilvio**, **Alex**, **Rui**, **Ștefan**, **Arthur**, **Sicco**, **Martin**, **Zeshan**, **Rahul**, **David**. Finally, thanks to all my colleagues from my time at Booking.com, in particular **Bram**, **Felipe**, and **Hugo** for showing me what industry work can look like.

I also want to thank my wonderful colleagues in the PRB lab. First, my paranymphs: **Jim**, we started on the same day and now we reached the end together, thanks for introducing me to so many people and helping me feel at home in the Netherlands from day one; **Aurora**, thanks for all the memorable and fun moments we have shared, for organizing so many events – even when you cannot attend them yourself – and for letting me meet your

housemates. For the PRLab, I start with thanking **David** and **Marco** who showed me that research should be about having fun and staying curious. Next, I would also like to thank **Tom**, for making sure that the fun does not only happen in the office but also outside. And thanks to **Chirag** for making sure every new person feels welcome and introduced to our group. I would also like to thank **Merve** and **Jing** for joining the group, providing the group with much energy and ideas from the outside. For the PhD students and postdocs, I want to thank each one for having contributed to making the group more inclusive, supportive, and for all the interesting conversations: **Mahdi**, **Taylan**, **Ojas**, **Ramin**, **Gijs**, **Myrthe**, **Yavuz**, **Yuko**, **Arman**, **Cheng**, **Sayak**, **Stephan**, and many more. I also would like to welcome **Matej** and **Isak** to the group as Jesse's new students; I'm excited to see how you will contribute to the group.

In the sibling labs: **Attila**, thanks for all the adventures together in the first years of my PhD, you're still the only one with whom I've stood together above 4000 meters; **Ombretta**, thanks for being so excited about your cats and sharing that with the rest of us; **Robert-Jan**, thanks for always bringing so much energy to our conversations; **Alejandro**, thanks for showing that hips do not lie; **Gabriel**, thanks for showing how October can be celebrated in September. **Jan**, **Xucong**, **Xiangwei**, **Hesam**, **Chengming**, **Thomas**, **Marian**, **Yancong**, **Nergis**, **Zhi-Yi**, **Yunqiang**, **Osman**, **Peter**, **Sander**, **Thomas**, **Alex**, **Stavros**, **Jana**, **Jasmijn**, **Thomas**, **Joana**, **Yasin**, **Ivan**, **Sara**, **Inez**, **Paul**, **Lorenzo**, **Colm**, **Jasper**, **Gerard**, **Bram**, **Swier**, **Daan**, **Stephanie**, **Hayley**, **Bernd**, **Chenxu**, **Vandana**, **Zonghuan**, **Litian** and many more, thanks for all the interesting conversations, ski trips, and other fun moments that make our labs so great. At last, I would also like to thank **Marunka** and before that **Saskia**, for keeping things running smoothly in the lab and always being within reach, and **Ruud** for saving me when my laptop started shutting down right before a deadline.

I also want to thank my close friends in Rotterdam and Delft, my past and current housemates, everyone from the yearly Lowlands group, everyone at the Wildhearts box, and my friends from Sweden who have visited me in the Netherlands on multiple occasions. You have all kept me sane during this time. I also want to thank all the people I've interacted with through the EA community, for keeping me inspired and focused on things that matter. And while there are too many friends from back home to name individually, it is no overstatement that this thesis would not have happened without **Jack**, **Eric**, **Oliver**, and **Fredrik**, who dispelled the doubts I had about going abroad to do a PhD. Almost five years later, those initial doubts have never returned.

I would like to thank some of the people closest to me. Jag vill tacka min familj i Göteborg, att ni alltid stöttat mina val och visat er kärlek, även när jag flyttat hemifrån. Chcę podziękować dziadkowi, babci, moim kuzynom i dalszej rodzinie w Polsce za całą waszą miłość i wsparcie. Szkoda, że nie mogę was odwiedzać częściej.

Last but not least, **Johanna**, thank you for your love and patience with my lack of patience, with you I feel like the real adventure has only started.

Rickard Karlsson
Rotterdam, February 2026

Curriculum Vitæ

Rickard Karl Axel Karlsson was born in Gothenburg, Sweden, on 6 September 1997. After completing his upper secondary education at LM Engströms Gymnasium (2016), he pursued a Bachelor of Science in Engineering Physics at Chalmers University of Technology (2019). During his undergraduate studies, he completed an internship at NASA Goddard Space Flight Center in Maryland, USA. He then continued at Chalmers to complete a master's degree in Engineering Mathematics & Computational Sciences (2021), also fulfilling the requirements for a Master of Engineering in Engineering Physics (in Swedish: Civilingenjörsexamen i Teknisk Fysik). During his master's studies, he spent a semester at Delft University of Technology in the Netherlands (2020). A year later, he returned to Delft to carry out his PhD research in the Pattern Recognition and Bioinformatics group, supervised by Jesse H. Krijthe and Marcel J.T. Reinders. During his PhD, he was a visiting graduate student at the CAUSALab at the Harvard T.H. Chan School of Public Health, hosted by Issa J. Dahabreh. He also spent several months in Amsterdam as a machine learning scientist intern at Booking.com.

List of publications

Publications included in this dissertation

1. Karlsson, R., & Krijthe, J. (2023). Detecting hidden confounding in observational data using multiple environments. *Advances in Neural Information Processing Systems, 36*, 44280–44309

2. Karlsson, R., Creastă, Ș., & Krijthe, J. H. (2023). Putting causal identification to the test: Falsification using multi-environment data. *Causal Representation Learning Workshop at NeurIPS 2023*

3. Karlsson, R., & Krijthe, J. (2025). Falsification of unconfoundedness by testing independence of causal mechanisms. *Proceedings of the 42nd International Conference on Machine Learning, 267*, 29128–29147

4. Karlsson, R., Akker, B. v. d., Moraes, F., Proença, H. M., & Krijthe, J. H. (2025). Qini curve estimation under clustered network interference. *arXiv preprint arXiv:2502.20097*

5. Karlsson, R., Wang, G., De Bartolomeis, P., Krijthe, J. H., & Dahabreh, I. J. (2026). Robust integration of external control data in randomized trials [Forthcoming]. *Biometrics*

6. Karlsson, R., De Bartolomeis, P., Dahabreh, I. J., & Krijthe, J. H. (2026). Robust estimation of heterogeneous treatment effects in randomized trials leveraging external data [Forthcoming]. *International Conference on Artificial Intelligence and Statistics*

Other publications

7. Karlsson, R., Bliek, L., Verwer, S., & de Weerdt, M. (2020). Continuous surrogate-based optimization algorithms are well-suited for expensive discrete problems. *BNAIC/BENELEARN (Selected Papers)*, 48–63

8. Karlsson, R., Willbo, M., Hussain, Z. M., Krishnan, R. G., Sontag, D., & Johansson, F. (2022). Using time-series privileged information for provably efficient learning of prediction models. *International Conference on Artificial Intelligence and Statistics*, 5459–5484

9. Bliek, L., Guijt, A., Karlsson, R., Verwer, S., & De Weerdt, M. (2023). Benchmarking surrogate-based optimisation algorithms on expensive black-box functions. *Applied Soft Computing*, *147*, 110744

The spherical cow is often used as a comical metaphor for overly simplified, and thus potentially misspecified, scientific models of complex phenomena. In this dissertation, we develop and study methods to test and relax common assumptions in causal inference. By doing so we can avoid the pitfall of estimating causal effects using potentially misspecified models. Because after all, real cows are not spherical.