Title

Three Practical Challenges in Causal Inference and Straightforward Solutions

ISBN

9798270209995

Author

Shinkre, Tanvi

Publication Date

2025-12-12

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Three Practical Challenges in Causal Inference and Straightforward Solutions

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Tanvi Shinkre

2025

ABSTRACT OF THE DISSERTATION

Three Practical Challenges in Causal Inference
and Straightforward Solutions

by

Tanvi Shinkre
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2025
Professor Chad J. Hazlett, Chair

In both observational and experimental studies, common methods for estimating treatment effects can be statistically biased if certain assumptions are not met. Moreover, the assumptions required for some methods to be unbiased can be difficulty to satisfy. This dissertation explores two common pitfalls in assessing treatment effects, and provides straightforward solutions to issues with current methods. The first chapter of this dissertation discusses the issues with using regression adjustment to estimate treatment effects for an observational study, focusing on the "weighting problem" that arises in the estimated effect from this method. The second and third chapters of this dissertation move to an experimental setting, in which researchers would like to condition on a post-treatment variable of interest, but cannot as this would lead to a statistically biased estimate. We propose a framework for accounting for post-treatment variables and obtaining a valid range of estimates for the treatment effect among a subgroup of interest. We demonstrate the use of this approach through both simulated and applied examples.

The dissertation of Tanvi Shinkre is approved.

Thomas R. Belin

Jennie E. Brand

Onyebuchi A. Arah

Chad J. Hazlett, Committee Chair

University of California, Los Angeles

2025

To my mom, whose dedication will always inspire me,

To my dad, who never let me disparage myself,

And to my brother, who always provided a listening ear –

I am forever grateful for your love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Chapter 1 of this thesis is a reprint of co-authored work with Chad Hazlett:

> Shinkre, Tanvi, and Chad Hazlett. "Demystifying and avoiding the OLS "weighting problem": Unmodeled heterogeneity and straightforward solutions." arXiv preprint arXiv:2403.03299 (2024).

As a co-author for this chapter, Chad Hazlett's contributions were: (i) advising me and overseeing my progress on the project, and (ii) contributing to the writing of the entire chapter. Chapter 2 of this thesis is a reprint of a co-authored article under review and appears as submitted to a journal:

> Hazlett, Chad, Nina McMurry, and Tanvi Shinkre. "Post-treatment problems: What can we say about the effect of a treatment among sub-groups who (would) respond in some way?." arXiv preprint arXiv:2505.06754 (2025). *Under Review.*

Chapter 3 of this thesis is based on joint work with Chad Hazlett and Matthew Coates.

It has been a long journey to get to to this point, and I have many people to thank who have been with me along the way. First and foremost, I'd like to acknowledge my advisor Chad Hazlett, who has been a kind and supportive mentor since day one. His advice and mentorship has been invaluable in the completion of this dissertation, and I feel lucky to have been his student. I'd also like to acknowledge my dissertation committee, for their thoughtful feedback throughout this process. Thank you to Jennie Brand, who was one of the reasons I came to UCLA, and who has been a gracious mentor since I was merely a prospective student. My work with her has pushed me to improve on both my methods and domain knowledge, and I'm grateful to have been given this opportunity. Thank you to Onyebuchi Arah, who was always open to answering causal questions (and questions about causal) and who helped shaped my perspective in this field. Thank you to Tom Belin, for his integral role in my causal learning.

In my time at UCLA, I've been lucky to have had the opportunity to work with many amazing people across multiple disciplines. I'd like to acknowledge James Macinko, for the opportunity to work on the type of interdisciplinary work that I wrote about working on when I was applying to graduate school. Through our collaboration, I have learned the intricacies of using causal methods in applied work, and I am grateful for his mentorship. Thank you to Jack Kappelman and Wilson Hammett, for the conversations on research from which I always came away with new knowledge, for the enjoyable collaborations, and for the friendship. Thank you to Falco Bargagli-Stoffi, who provided valuable feedback and advice in the year that we overlapped at UCLA.

I am grateful to have been part of the Practical Causal Inference Lab at UCLA, and grateful to Chad Hazlett and Onyebuchi Arah for their efforts to bring together causal enthusiasts from across campus. Thank you to Elizabeth Frank O'Neill, Duy Pham, Borna Bateni, Keri Lintz, Jack Kappelman, Soonhong Cho, Davis Vo, Nathan Hoffman, and Roxana Rezai, for being great labmates.

I am also grateful to have been a part of the Inequality Data Science Lab at UCLA, and grateful to Jennie Brand and Ian Lundberg for bringing us together. Thank you to Nanum Jeon for being a wonderful collaborator and friend throughout the years.

I'd like to acknowledge Melody Huang, for her mentorship and friendship through my early years in graduate school. Starting graduate school during a pandemic was difficult, and her friendship in that first year helped me to feel more connected to the department. On that note, I'd also like to thank Stella Huang, Elizabeth Frank O'Neill, Surabhi Agarwal, Nancy Zhang, and Christy Lee – that first year could have been incredibly lonely if not for our friendship.

I have been so lucky to have met amazing people during my time in LA, whose friendship has kept me going through the ups and downs of graduate school.

Thank you to Stella Huang, Sophie Phillips, Davis Berlind, Hector Escandin, John Baeirl,

Nicole Lee, and Alex Chen for the community inside the department and beyond.

Thank you to Haein Kim, Brooke Zou, Chris Chien, Kimberly Venegas, and Trinny Tat for being among my first friends at UCLA.

Thank you to Alessandro Bullita, Ami Sheth, Abhishek Devarajan, and Dan Chen for the climbs, conversations, and friendship.

Thank you to Filippo Monti, for the hugs (and the debates that were sometimes entertaining), Ali Mofrad, for the shared laughter, and Kyle Webster for the puns (and for Jet Lag).

Thank you to my long distance friends, Aishy Murali, Angie Chen, and Prachi Khandekar, for supporting me from 400 miles away.

Thank you to Brooke Pearson and Joseph Resch, for your kindness and compassion since our first year in LA, and for all the memories made.

I am so grateful to Rushil Raghavan, for being by my side through all the stresses of the last couple years. Thank you for believing in me when I refused to believe in myself.

I am immensely grateful to my neighbors, my friends who became family, Noor Nakhaei, Emma Landry, and Muratkhan Abdirash, for being the best support system that a person could have. Thank you for being there through all the ups and downs, for letting me come over and cry with no announcement, for celebrating my wins in the biggest ways, and for all the laughter.

Finally, a million thank yous to my parents, who have been there for me to lean on throughout my entire journey, and my brother, who has always been my steadfast supporter. This would not have been possible without them.

VITA

2016–2020 B.A. in Applied Mathematics and Data Science, University of California, Berkeley.

2020–2024 C.Phil. in Statistics, University of California, Los Angeles.

2020– Ph.D. Student in Statistics, University of California, Los Angeles.

PUBLICATIONS AND PRESENTATIONS

Hazlett, Chad, Nina McMurry, and Tanvi Shinkre. "Post-treatment problems: What can we say about the effect of a treatment among sub-groups who (would) respond in some way?." arXiv preprint arXiv:2505.06754 (2025). *Under Review.*

Shinkre, Tanvi, Zachary Wagner, and Louis T. Mariano. 2025."Comparing Hierarchical Linear Model and Ordinary Least Squares Approaches in the Analysis of Cluster Randomized Trials: A Guide for Practitioners." *Under Review.*

Xu, Jiahui, Jennie E. Brand, Tanvi Shinkre, and Nanum Jeon. 2024. "Flexibly Detecting Effect Heterogeneity with an Application to the Effects of College on Reducing Poverty." SocArXiv. September 25. doi:10.31235/osf.io/65hcj. *Under Review.*

Macinko, James, Tanvi Shinkre, Falco J. Bargagli-Stoffi, Jack Kappelman, Diana Silver. "Neighborhood Exposure to Gun Violence and Psychosocial Distress in California: A Cross-Sectional Study." *Under Review.*

Kappelman, Jack, Diana Silver, Jin Yung Bae, Kevin Butler, Tanvi Shinkre, Falco J. Bargagli-Stoffi. "Sensitive Places, Persistent Violence: Effectiveness of "Bar Ban" Laws in Reducing Gun Violence Near Alcohol Vendors." *Under Review.*

Shinkre, Tanvi, and Chad Hazlett. "Demystifying and avoiding the OLS "weighting problem": Unmodeled heterogeneity and straightforward solutions." arXiv preprint arXiv:2403.03299 (2024).

Examining the relationship between community violence and alcohol use in California, 2021-2022. *Oral Presentation.* 2025 National Research Conference for the Prevention of Firearm-Related Harms (November 2025).

Estimating effects of extra time for background check laws on firearm-related deaths, 2000-2020. *Oral Presentation.* 2025 National Research Conference for the Prevention of Firearm-Related Harms (November 2025).

Exploring the Causal Relationship Between Firearm Laws and Firearm-Related Deaths, 2000-2021. *Oral Presentation.* UCLA Interdisciplinary Conference on Gun Violence Reduction (May 2025).

Heterogeneous Treatment Effect Estimation with the grf package. *Workshop.* 2023 Summer Institute for Computation Social Science at UCLA (June 2023).

Visualizing Causal Trees with the htetree package. With Jennie Brand and Jiahui Xu. *Oral Presentation.* Syracuse University, Cornell University, and SUNY-Albany, Center for Aging and Policy Studies, Spring 2021.

CHAPTER 1

Introduction

Applied quantitative researchers are often interested in estimating the effect of a treatment on some outcome of interest. In observational studies, in order to get a valid causal effect estimate researchers must identify and adjust for confounders of the treatment effect. It is regular practice to adjust for these confounders by including them as independent variables in a linear regression of outcome on treatment. The first chapter of this dissertation analyzes the issues with this approach in estimating causal effects for observational studies. Even without unobserved confounding, the coefficient on treatment in the described regression yields a conditional-variance-weighted average of strata-wise effects, not the average treatment effect. Scholars have proposed characterizing the severity of these weights, evaluating resulting biases, or changing investigators' target estimand to the conditional-variance-weighted effect. We aim to demystify these weights, clarifying how they arise, what they represent, and how to avoid them. Specifically, these weights reflect misspecification bias from unmodeled treatment-effect heterogeneity. Rather than diagnosing or tolerating them, we recommend avoiding the issue altogether, by relaxing the standard regression assumption of "single linearity" to one of "separate linearity" (of each potential outcome in the covariates), accommodating heterogeneity. Numerous methods – including regression imputation (g-computation), interacted regression, and mean balancing weights – satisfy this assumption. In many settings, the efficiency cost to avoiding this weighting problem altogether will be modest and worthwhile.

While the first chapter focuses on causal inference in an observational setting, the second

and third chapters of this dissertation aim to tackle an important problem in the causal analysis of randomized experiments. Even when treatment is randomized, researchers can face issues when estimating the causal effect of the treatment. It is possible, for example, that the assigned treatment was not implemented sufficiently for some subjects in the treated group, or that some subjects dropped out of the experiment before an outcome was measured. In these cases, we cannot condition on implementation, or subset to the group that did not drop out, because there may be variables that affect implementation or dropout that also affect the outcome. The second chapter of this dissertation introduces a new quantity of interest for these settings, the Treatment Reactive Average Causal Effect (TRACE), defined as the total effect of the treatment in the group that, if treated, would have a particular value of the relevant post-treatment variable. By reasoning about the effect among the "non-reactive" group, we can identify and estimate the range of plausible values for the TRACE. We demonstrate the use of this approach with three examples: (i) learning the effect of police-perceived race on police violence during traffic stops, a case where point identification may be possible; (ii) estimating effects of a community-policing intervention in Liberia, in communities that meaningfully implemented it, and (iii) studying how in-person canvassing affects support for transgender rights, among participants for whom the intervention would result in more positive feelings towards transgender people.

The proposed framework for estimating the TRACE can also be useful in assessing the benefit of early screening for cancer on mortality among those who would have had detected cancer if screened. The third chapter of this dissertation focuses on the application of the TRACE in this and related screening settings. In these settings, the total effect of screening on mortality is difficult to interpret – it would be the effect of screening on mortality among all the people who were screened. If a person has cancer, screening will help with early diagnosis and could increase their life span, but if a person does not have cancer, that person's life span without screening might not be any different. The more pertinent question may be: "Does screening reduce mortality for those who have detectable cancer at the time

of screening?" We propose using the TRACE framework to partially identify the total effect of screening on mortality for those who, if screened, would have had cancer detected. To do this, we need to reason about the total effect among those who, if screened, would not have had cancer detected. We detail the use of this method to assess effects of screening in the third part of this dissertation.

CHAPTER 2

Demystifying and avoiding the OLS "weighting problem": Unmodeled heterogeneity and straightforward solutions

2.1 Introduction

When estimating the effect of a treatment on an outcome of interest by adjusting for covariates (X), researchers typically hope to interpret their result as a well-defined causal quantity, such as the average effect over strata such as the average treatment effect (ATE). Despite the introduction of many more flexible estimation procedures, regression adjustment remains the "workhorse approach" when adjusting for covariates to make causal claims in many disciplines [Aronow and Samii, 2016].[1] For a binary treatment (D) and outcome (Y), a simple bivariate regression of Y on D gives the difference in means estimate (the mean of Y when $D = 1$ minus the mean of Y when $D = 0$). However, this equivalence breaks down when covariates X are included in the regression (as in $Y \sim D + X$), if the treatment effect may vary across values of X. Instead, the coefficient produced by this regression represents a weighted average of strata-specific difference in means estimates, where the weights are larger

[1] In political science, for example, Keele et al. [2010] found that regression adjustment was used in analyzing 95% of experiments in American Political Science Review, 95% in Journal of Politics, and 74% in American Journal of Political Science. The ubiquity of regression adjustment approaches in practice is echoed by authors in economics and other disciplines as well, such as Angrist and Pischke [2009], Humphreys [2009], Chattopadhyay and Zubizarreta [2023]. This is perhaps due to familiarity, ease of use, the efficiency of OLS and its suitably good performance in many contexts [Hoffmann, 2023, Green and Aronow, 2011, Kang and Schafer, 2007], and well-established uncertainty estimation considerations and tools.

for strata with probability of treatment closer to 50% [Angrist, 1998, Angrist and Pischke, 2009]. Such a weighted average is not typically of direct interest, and under severe enough heterogeneity in treatment effects, this can lead to widely incorrect substantive conclusions, as demonstrated below.

An ongoing literature regards these weights as a nuisance to be coped with or incorporated into analysis. Accordingly, authors have proposed diagnostics that index the potential for bias due to these weights or tools to aid interpretation given these weights [Aronow and Samii, 2016, Słoczyński, 2022, Chattopadhyay et al., 2023, Chattopadhyay and Zubizarreta, 2023, Kline, 2011], and have analyzed the behavior of proposed regression estimators in different settings in light of these weights (e.g. Chernozhukov et al., 2013).

In this article, we seek to demystify these weights, offering a simple derivation and conceptual clarification of why they arise, what they represent, and how to avoid them. We offer two main theoretical clarifications. First, under heterogeneous treatment effects (and varying probability of treatment), the linear regression of Y on D and X will be misspecified (e.g. Rubin, 1979, Imbens and Wooldridge, 2009, Imbens, 2015). Re-expressing the regression coefficient as a weighted average of strata-wise effects simply offers a way to see how this misspecification causes the regression coefficient to diverge from the ATE. Such weights emerge, as we show, simply by relying on the Frisch-Waugh-Lowell (FWL) theorem [Frisch and Waugh, 1933, Lovell, 1963] to first construct the unit-level (not strata-level) weights. We can then group these weights to form a new expression of the strata-wise weights. Our expression does not rely on the usual assumption (as in Angrist [1998] and later sources) that D is linear in X, but reduces to that expression in the special case where D is linear in X. This result corroborates the recent independent contribution of Hahn [2023].

Second, we clarify the assumption under which approaches can avoid these weights, which determines what (existing) estimators avoid this problem. Rather than attempting to diagnose, interpret, or otherwise cope with the undesired affect of these weights on interpretation, they can be avoided by any estimator that works under the *separate linearity*

assumption, meaning each potential outcome $Y(d)$ is assumed to be linear in X, as opposed to the *single linearity* assumption that required Y itself to be linear in X and D. Equivalent assumptions have been stated in analyses by Hahn [2023], Słoczyński [2022], Kline [2011] and Imbens and Wooldridge [2009]. Fortunately, many well-known estimation approaches can be justified under the separate linearity framework including (i) regression imputation/g-computation/T-learner/multi-regression [Peters, 1941, Belson, 1956, Robins, 1986, Künzel et al., 2019, Chattopadhyay and Zubizarreta, 2023]; (ii) regression of Y on D and X and their interaction [Lin, 2013], and (iii) balancing/calibration weights that achieve mean balance on X in the treated and control groups (e.g. Hainmueller, 2012). Approaches (i) and (ii) are shown to be identical in the context of OLS models without regularization. Mean balancing weights (iii), when constructed to target the ATE, can be justified by the same assumption of separate linearity, but uses a different estimation approach and produces different point estimates. All three of these strategies produce unbiased ATE estimates under separate linearity, without the "weighting problem" suffered under the single linearity regression. These benefits do come at an efficiency cost, though it is low in settings with few covariates relative to sample size.

These considerations apply broadly to observational research in which an assumption of no unobserved confounding is relied upon. However, they can also apply to the analysis of even randomized experiments, concurring with Lin [2013]. In particular, we illustrate these implications and recommendations for block-randomized experiments.

2.2 Background: OLS weights

2.2.1 Setting and notation

We consider settings where we are interested in estimating the average effect of binary treatment D on outcome Y, while accounting for confounder X. The average treatment

effect is defined as

$$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{2.1}$$

where $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control, respectively [Neyman, 1923, Rubin, 1974]. In order to estimate this average treatment effect using observed data, we require an absence of unobserved confounders, concisely expressed as the conditional ignorability assumption,

$$D_i \perp\!\!\!\perp \{Y_i(1), Y_i(0)\} | X_i \tag{2.2}$$

For some discrete variables X that satisfy conditional ignorability, and given consistency of the potential outcomes, the ATE can be identified according to:

$$\begin{aligned} \tau_{ATE} &= \sum_{x \in \mathcal{X}} \mathbb{E}[Y(1) - Y(0) | X = x] P(X = x) && (2.3) \\ &= \sum_{x \in \mathcal{X}} \left(\mathbb{E}[Y | D = 1, X = x] - \mathbb{E}[Y | D = 0, X = x] \right) P(X = x) \\ &= \sum_{x} DIM_x P(X = x) && (2.4) \end{aligned}$$

where we suppress the i subscripts for legibility and DIM_x is the estimand for the difference in means in the stratum where $X = x$. In a given sample, this is approximated using the analog estimator,[2]

$$\hat{\tau}_{ATE} = \sum_{x} \widehat{DIM}_x \hat{P}(X = x) \tag{2.5}$$

The term $\hat{P}(X = x)$ gives the empirical proportion of units in each stratum in the sample, and can be thought of as the "natural" strata-wise weights, as they marginalize over the stratum according to the fraction of units falling in that stratum. It would be natural to form

[2]This quantity could also properly be labeled as the (estimate of) the sample average treatment effect (SATE) rather than the ATE. However, we maintain the ATE notation for simplicity as its expectation over-samples is still the ATE, presuming the sample in question is a probability sample from the population of interest.

an analog estimator for this through sub-classification/stratification, simply computing the difference in means in each stratum of X and compiling them per expression 2.5. However, suppose we instead attempt to estimate the treatment effect by fitting a regression according to the model

$$Y = \beta_0 + \beta_1 X + \tau_{reg} D + \epsilon \tag{2.6}$$

While regression-based estimation of treatment effects has been widely used in practice across disciplines for decades, as Angrist [1998] and many scholars since then have emphasized, $\hat{\tau}_{reg}$ do not in general yield $\hat{\tau}_{ATE}$, even when conditional ignorability holds. Rather, it can be understood as a version of Expression 2.5 but with weights on each $\widehat{DIM}_x$ that differ from $\hat{P}(X = x)$. We now turn to a ground-up analysis of these strata-wise weights intended to demystify them and to recognize them as the natural consequence of misspecification of the linear model under heterogeneous effects.

2.2.2 The unit-level weighting representation of OLS

While our plan is to (re-)derive strata-wise weights on the $\widehat{DIM}_x$ components, we start with "unit-level weights", which arise simply because the regression coefficient is a weighted sum of Y_i values. Specifically, let $\hat{d}(X_i) = \hat{p}(D_i = 1|X_i) = \mathbf{X}_i\hat{\beta}$ where $\mathbf{X}_i = \begin{bmatrix} 1 & X_i \end{bmatrix}$, and $\hat{\beta}$ is the estimated coefficients from a linear regression of D on X. The unit-wise weighting formula that corresponds to the OLS estimate is defined according to

$$\hat{\tau}_{reg} = \sum_i w_i Y_i \tag{2.7}$$

for some weights w_i. By the FWL theorem [Frisch and Waugh, 1933, Lovell, 1963],

$$\hat{\tau}_{reg} = \frac{\widehat{\text{Cov}}(Y_i, D_i^{\perp X})}{\widehat{\mathbb{V}}(D_i^{\perp X})} = \frac{\sum_i Y_i(D_i - \hat{d}(X_i))}{\sum_i (D_i - \hat{d}(X_i))^2} \tag{2.8}$$

$$\tag{2.9}$$

which already takes the form of Equation 2.7, with weights given by

$$w_i = \frac{D_i - \hat{d}(X_i)}{\sum_i (D_i - \hat{d}(X_i))^2} \tag{2.10}$$

Connection to propensity score. Though these individual-level weights are an intermediate step in our analysis, we give some consideration to their form here. First, as the denominator is a scalar, the behavior of these weights is revealed through the numerator, $D_i - \hat{d}(X_i)$. This is simply $1 - \hat{d}(X_i)$ for treated units and $\hat{d}(X_i)$ for control units. Compare this to inverse-propensity score weights (IPW), which are proportional to $1/\hat{d}(X_i)$ for treated units and $1/(1 - \hat{d}(X_i))$ for control units. Like the IPW weights, these unit-level weights in Expression 2.10 place greater weight on units when their treatment status is more "surprising" given the covariate values. Unlike IPW, the linear regression weights do not require constructing a ratio that has a denominator that can become close to zero, which can create explosive weights under IPW.[3] This similarity of the unit-level weights to the propensity score approach is explored in detail in Kline [2011].

Negative individual-level weights. The weights w_i are constructed to be both positive and negative because Equation 2.7 is structured as a single weighted sum/average. An isomorphic but useful way of signing the weights is to instead seek the weights that would appear in a "weighted difference in means" estimator,

$$\hat{\tau}_{wdim} = \sum_{i:D=1} \tilde{w}_i Y_i - \sum_{i:D=0} \tilde{w}_i Y_i \tag{2.11}$$

where $\tilde{w}_i = D_i w_i + (1 - D_i)(-w_i)$, which simply changes the sign for control units in order to accommodate the subtraction rather than the summation in the form of Expression 2.7.

[3]Equivalent unit-level weights are explored in depth by Chattopadhyay and Zubizarreta [2023] and in Chattopadhyay and Zubizarreta [2024], where the authors usefully employ these weights to analogize how OLS implicitly compares (weighted) mean outcomes for the treated to (weighted) mean outcomes for the controls, sharpening the analogy to the difference-in-means estimator one might apply in a randomized experiment and discussing implications for qualities such as effective balance and sample size.

The weights $\tilde{w}$ signed as in Expression 2.11 will often be positive, and any positive weight may be naturally interpreted as the relative contribution of each observation. However, they can be negative as well. Chattopadhyay and Zubizarreta [2024] note that negative weights "translate into forming effect estimates that may lie outside the support or range of the observed data", or that they "[leave] room for biases due to extrapolation from an incorrectly specified model." Borusyak and Hull [2024] also discuss negative weights, emphasizing that under randomization, these weights will all be non-negative, and so do not pose a problem for the resulting estimand.

We note that the meaning of negative individual weights is closely tied to estimates one would obtain from a linear probability model regressing the treatment indicators on X, even though that model is not run. Specifically, weights will be negative for a treated unit when $\hat{d}(X_i) > 1$, and for control units when $\hat{d}(X_i) < 0$. If a treated unit lies at an X position "so unique to treated units" that a linearly-fitted model produces a predicted value greater than 1 there, it will have a negative weight. This can be thought of as indicating that a treated unit is "more like treated units than any control units", so that it cannot be well compared to control units. The symmetric argument holds for control units with a negative value.

In this way, negative individual level weights do indicate extreme non-overlap, but in a very peculiar, model dependent sense and not in any non-parametric way that corresponds to broader notions of common support. These weights are also not a sure diagnostic for poor overlap or extreme counterfactuals: one could easily imagine a situation where all units above a certain value of X are treated and all those below another value of X are controls, indicating poor overlap. However, there can be many (or all) units outside the area of common support but with positive weights, because the fitted linear probability model for D given X does not exceed 1 (for treated units) or drop below 0 (for control units). Thus, we emphasize that negative weights can provide a warning that "some units look too much like treated units to have valid comparisons" (and likewise for controls), but a positive weight is not a guarantee of meaningful comparability in the sense of finding units of the opposite

treatment value nearby in the covariate space.

2.2.3 From individual to strata-wise weights

With the individual-level weights in hand, we can now construct the strata-level weights of interest to our analysis. These weights are interesting because they allow us to see how the regression coefficient compares to ATE as a combination of average treatment effects by stratum as in Equation 2.5. To do so we consider the case of discrete X and the strata-wise weights, meaning those that can be expressed as,

$$\hat{\tau}_{reg} = \sum w_x \hat{\tau}_x \tag{2.12}$$

where w_x is the weight for the strata in which $X = x$ and $\hat{\tau}_x = \hat{\mathbb{E}}[Y_i(1) - Y_i(0)|X_i = x]$ is the conditional average treatment effect for subgroup $X_i = x$. The resulting $\hat{\tau}_{reg}$ correspond to the ATE only when $w_x = P(X = x)$. The best known expression for such strata-wise weights of OLS arises from Angrist [1998],

$$\hat{\tau}_{reg} = \frac{\sum_x \hat{\tau}_x \hat{d}(X_i)(1 - \hat{d}(X_i))\hat{P}(X_i = x)}{\sum_x \hat{d}(X_i)(1 - \hat{d}(X_i))\hat{P}(X_i = x)} \tag{2.13}$$

This form shows that regression weights strata-wise $\widehat{DIM}_x$ components not according to $\hat{P}(X_i = x)$ as required to represent the ATE, but proportionally to $\hat{d}(X_i)(1 - \hat{d}(X_i))\hat{P}(X_i = x)$. Such a regression puts the most weight on strata in which the conditional variance of treatment status is largest, i.e. when $\hat{d}(X_i)$ is nearest to 50%. An intuition for this notes that OLS seeks to minimize squared error, and the opportunity to learn the most from strata with middling probabilities of treatment [Angrist and Pischke, 2009]. Absent treatment effect heterogeneity, this is unproblematic. However, if there are high levels of effect heterogeneity in our data, then depending on how they correspond to strata of X and the probability of treatment in those strata, these weights could move the regression coefficient far from the average treatment effect [Aronow and Samii, 2016, Humphreys, 2009,

Angrist and Pischke, 2009, Słoczyński, 2022]. Goldsmith-Pinkham et al. [2022] also discusses similar weights in the context of estimating effects for multiple treatments at once, where this leads to "contamination" of effect estimates.

These weights, however, are obtained under the additional assumption that the probability of treatment is linear in X, and not just modeled that way. Angrist and Pischke [2009] satisfy this by assuming the corresponding outcome regression can be saturated in X. This is often a reasonable assumption with discrete, low-dimensional X. Nevertheless we may wish to have a more general expression for the weights that does not rely on such an assumption, either for completeness or in service of generalizing to the case where X is not discrete or is otherwise infeasible to include in a saturating form (i.e. X with numerous multiple dimensions and levels). The issues that could arise when this type of assumption is not satistifed are highlighted by Blandhol et al. [2022], who similarly analyze the weighted average representation of the two stage least squares estimate for the local average treatment effect presented by Angrist and Pischke [2009], and show that the equivalent linearity assumption that is invoked is often not satisfied in practice.

To obtain a more general expression for the weights, we simply organize the individual-level weights above into strata,

$$\hat{\tau}_{reg} = \sum_i w_i Y_i \tag{2.14}$$

$$= \frac{\sum_i Y_i(D_i - \hat{d}(X_i))}{\sum_i (D_i - \hat{d}(X_i))^2} \tag{2.15}$$

$$= \frac{\sum_x \hat{P}(X_i = x)\widehat{\mathbb{E}}[(D_i - \hat{d}(X_i))Y_i | X_i = x]}{\sum_x \hat{P}(X_i = x)\widehat{\mathbb{E}}[(D_i - \hat{d}(X_i))^2 | X_i = x]} \tag{2.16}$$

Rearranging these terms in the form $\hat{\tau}_{reg} = \sum_x w_x \tau_x$ is not in general possible, as the expected

outcome under treatment and the expected outcome under control are weighted differently:

$$\widehat{\mathbb{E}}[(D_i - \hat{d}(X_i))Y_i|X_i = x] = \widehat{\mathbb{E}}[D_i(1 - \hat{d}(X_i))Y_i - (1 - D_i)\hat{d}(X_i)Y_i|X_i = x] \tag{2.17}$$

$$= (1 - \hat{d}(X_i))\widehat{\mathbb{E}}[D_iY_i|X_i = x] - \hat{d}(X_i)\widehat{\mathbb{E}}[(1 - D_i)Y_i|X_i = x] \tag{2.18}$$

$$= (1 - \hat{d}(X_i))\pi_x\widehat{\mathbb{E}}[Y_i|X_i = x, D_i = 1] \tag{2.19}$$

$$- \hat{d}(X_i)(1 - \pi_x)\widehat{\mathbb{E}}[Y_i|X_i = x, D_i = 0]$$

where $\pi_x = \hat{P}(D_i = 1|X_i = x)$ is the true, or correctly specified, probability of treatment in the sample for a given stratum. Thus the strata-wise weighting generally imposed by OLS for a given sample involves a combination of the true probability of treatment in the sample given X, and the probability of treatment given X estimated using a linear model. This result is equivalent to the weighted representation independently developed by Hahn [2023], who also considers a special case for when the outcome model for the control group is linear. We can also formulate our result in terms of the discrepancy, a_x, between the true and linearly-approximated probability of treatment given X, so that $\hat{d}(X_i) = \pi_x + a_x$. The general strata-wise weights corresponding to OLS are then

$$\hat{\tau}_{reg} = \frac{\sum_x \hat{P}(X_i = x)\pi_x(1 - \pi_x)\tau_x - a_x\hat{\mathbb{E}}[Y_i|X_i = x]}{\sum_x \hat{P}(X_i = x)(\pi_x(1 - \pi_x - a_x) + a_x(\pi_x + a_x))} \tag{2.20}$$

Appendix A.1 gives additional details. Notice that this expression is infeasible to compute when D is not linear in X, as π_x is unknown. However, if $a_x = 0$ (i.e. D is truly linear in X), this representation reduces to the variance-weighted representation of the regression coefficient given by Angrist [1998]. For many purposes then the Angrist [1998] representation provides a clear and intuitive conception of the weights. Nevertheless, we found it necessary to have this more complete formulation in order to obtain the correct answer, as our simulations below show.

2.2.4 From single linearity to separate linearity

An OLS model regressing Y on D and X alone would be correctly specified if the true conditional expectation function, $\mathbb{E}[Y|X, D]$ is linear in D and X as in $\mathbb{E}[Y|X, D] = \beta_0 + \beta_1 X + \beta_2 D$. We call such an assumption the *single linearity* assumption, because it requires a single assumption about Y being linear in some terms. Note that this allows for random variation in treatment effects not correlated with X, but it does not accommodate treatment effect variation that is correlated with X.

Accordingly, if the ATE rather than the regression coefficient is the target of inference, a natural solution is not to characterize or diagnose or bound the difference between the coefficient and the ATE, but rather to avoid the underlying misspecification problem. The strata-wise weights above merely describe the regression coefficient in other terms, and as such, elucidate the impact of misspecification for the difference between the coefficient and the ATE. Resolving this misspecification bias is a natural solution.

Here we consider the minimal change to regression practice that an investigator otherwise comfortable with a linear model could employ to avoid this misspecification concern and the resulting weighting behavior. We first define the *separate linearity* assumption, requiring that each of the potential outcomes is separately linear in X,

$$\mathbb{E}[Y(0)|X] = \alpha_0 + \alpha X \tag{2.21}$$

$$\mathbb{E}[Y(1)|X] = \gamma_0 + \gamma X \tag{2.22}$$

This assumption is slightly weaker than single linearity, which effectively forces $\alpha = \gamma$. While we label this assumption for easy reference and to distinguish it from single linearity, it is not intended to be novel, and further can be understood as the (often implicit) motivation for a number of longstanding approaches described next [Imbens, 2015, Kline, 2011, Imbens and Wooldridge, 2009].

We turn next to a variety of known and straightforward estimation approaches that are

in keeping with this assumption and that wholly avoid the awkwardness of "regression's weighting problem" rather than employing diagnostic or interpretational aids.

2.2.5 Estimation approaches suitable for separate linearity

Interactions, imputation, g-computation, and stratification. Fortunately, a number of existing approaches are suggested by such an assumption. One alternative estimation approach, proposed by Lin [2013] initially to mitigate bias in covariate-adjusted estimates from randomized experiments, is to use a regression model that includes an interaction term for the treatment and confounder. Goldsmith-Pinkham et al. [2022] similarly propose interacted models in the context of multi-valued treatments.

In the binary treatment setting we consider, the treatment effect estimate, $\hat{\tau}_{interact}$, is the estimated coefficient on the treatment variable D_i in the regression of Y_i on D_i, $X_i - \bar{X}$, and $D_i(X_i - \bar{X})$. The centering of X in these terms is useful for interpretation: while the fitted models with and without centering X are isomorphic, centering X allows the coefficient on D to represent the average effect "when X is at its mean". By linearity, this is also exactly the average marginal effect of D on Y taken across observations at their observed values of X.[4]

Another straightforward approach to estimate the ATE under separate linearity is to run separate regressions of Y on X for the treatment group and the control group, and use the results from each regression to predict the unobserved outcomes in the other group. Then, using these predicted outcomes, we can estimate the individual treatment effect for each subject, and take the average of the estimated individual treatment effects to get the ATE. This is also known as g-computation, meaning that an estimate from g-computation will also be unbiased for the ATE in settings with high levels of heterogeneity in treatment effect and

[4]This centering procedure and the resulting interpretation is complicated when one level of X (or the intercept) must be dropped, as when X represents block indicators or more generally group fixed effect. We describe this concern and propose a solution for this in Section 2.3.3.

probability of treatment assignment [Robins, 1986, Snowden et al., 2011]. It has also been referred to even more recently as "multi-regression" [Chattopadhyay and Zubizarreta, 2024], who note earlier uses of this approach as far back as Peters [1941] and Belson [1956]. It is also equivalent to the Oaxaca-Blinder decomposition [Oaxaca, 1973, Blinder, 1973].

Specifically, let $\hat{\mu}_0(X_i)$ be the model fit to the control units and $\hat{\mu}_1(X_i)$ be the model fit to the treated units,

$$\hat{\tau}_{imp} = \frac{1}{n}\left(\sum_{D_i=1}(Y_i - \hat{\mu}_0(X_i)) + \sum_{D_i=0}(\hat{\mu}_1(X_i) - Y_i)\right) \tag{2.23}$$

$$= \frac{1}{n}\left(\sum_{D_i=1}((\hat{\mu}_1(X_i) + \epsilon_{1i}) - \hat{\mu}_0(X_i)) + \sum_{D_i=0}(\hat{\mu}_1(X_i) - (\hat{\mu}_0(X_i) + \epsilon_{0i}))\right) \tag{2.24}$$

$$= \frac{1}{n}\left(\sum_{i}\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right) \tag{2.25}$$

Because this explicitly puts weights of $1/N$ on every unit, the estimate does not suffer from the "weighting problem".[5] Further, the Lin estimate and the regression imputation estimate are easily shown to be identical in this context. Specifically, the Lin estimate models $\mathbb{E}[Y|D, X]$ as $\beta_0 + \beta_1 D + \beta_2 X + \beta_3 DX$, and thus implies

$$\mathbb{E}[Y|X, D = 0] = \beta_0 + \beta_2 X \tag{2.26}$$

$$\mathbb{E}[Y|X, D = 1] = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X \tag{2.27}$$

Meanwhile, the imputation estimate is based on two regression models, one for the untreated group and one for the treated group:

$$\hat{\mu}_0(X_i) = \mathbb{E}[Y|X, D = 0] = \alpha_0 + \alpha_1 X \tag{2.28}$$

[5]We note the equality of Expression 2.23 to Expression 2.25 above implies that one may either (i) compare each unit's observed outcome under the realized treatment status to the modeled outcome for that unit under the opposite, or (ii) for each unit, compare the *modeled* outcome under treatment to the *modeled* outcome under control, without using the observed outcome. This is a result of relying on OLS for each outcome model, since the average fitted value from a given model will be precisely equal to the average observed outcome over the same group. Such a property does not hold with estimators that cannot guarantee $\bar{\hat{\epsilon}} = 0$.

$$\hat{\mu}_1(X_i) = \mathbb{E}[Y|X, D = 1] = \gamma_0 + \gamma_1 X \tag{2.29}$$

Comparing this to the above equations, we can prove $\alpha_0 = \beta_0$, $\alpha_1 = \beta_2$, $\gamma_0 = \beta_0 + \beta_1$, and $\gamma_1 = \beta_2 + \beta_3$ by showing the equivalency of the minimization problems in question (see Appendix A.2 for details). Using these equivalencies, we can show that the ATE estimate from regression imputation is equivalent to the estimate from the interacted regression,

$$\hat{\tau}_{imp} = \frac{1}{n}\sum_{i=1}^{n}(\gamma_0 + \gamma_1 X_i - (\alpha_0 + \alpha_1 X_i)) \tag{2.30}$$

$$= (\gamma_0 - \alpha_0) + (\gamma_1 - \alpha_1)\bar{X} \tag{2.31}$$

$$= (\beta_0 + \beta_1 - \beta_0) + (\beta_2 + \beta_3 - \beta_2)\bar{X} \tag{2.32}$$

$$= \beta_1 = \hat{\tau}_{interact} \tag{2.33}$$

Since these two estimation methods are identical under OLS, they can be used interchangeably. One can directly show the unbiasedness of either under assumptions of consistency and conditional ignorability,

$$\mathbb{E}[\hat{\tau}_{imp}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right)\right] \tag{2.34}$$

$$= \mathbb{E}[\mathbb{E}[Y_i|D_i = 1, X_i]] - \mathbb{E}[\mathbb{E}[Y_i|D_i = 0, X_i]] \tag{2.35}$$

$$= \mathbb{E}[Y_i(1) - Y_i(0)] \tag{2.36}$$

Both are also identical to the stratification estimate when X is discrete,

$$\hat{\tau}_{interact} = \hat{\tau}_{imp} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\right) \tag{2.37}$$

$$= \sum_{x \in X}\hat{P}(X = x)\widehat{\mathbb{E}}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)|X_i = x] \tag{2.38}$$

Mean balancing weights. We can also connect the separate linearity assumption to a justification for calibration/balancing weights. Consider mean balancing weights for both

the treated and control units, so that weighted average of covariates for each is equal to the overall unweighted average of covariates,

$$\sum_{i:D_i=0} w_i X_i = \sum_{i:D_i=1} w_i X_i = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{2.39}$$

While many choices of weights can satisfy these constraints (when feasible), it is desirable to minimize some measure of their variation. In the case of maximum entropy weights [Hainmueller, 2012], this is done by maximizing the entropy, $\sum_i w_i log(w_i/q_i)$. The key idea is that by achieving equal means on X in the treated and control group (both equal to the full samples mean on X), then any linear function of X—which include $Y(1)$ and $Y(0)$ if separate linearity holds— will also have equal means in these groups. A difference in means estimator using these weights is then unbiased for the ATE, without requiring any appeal to the relationship between this weighting procedure and propensity score modeling.[6] Kline [2011] discusses how the regression imputation approach can be formulated as a weighting estimator that balances the means of covariates. An interesting feature of the mean balancing approach is that while it is justified by the same assumptions as regression imputation or the interactive regression above, their estimation strategies are different. Balancing weights do not require estimating the (nuisance) coefficients of any model. This may improve tolerance to misspecification (i.e. non-linear conditional expectations for each potential outcome) at the cost of variance, as borne out in simulations below.

2.3 Simulations

To verify that these estimators behave as our analysis claims, we explore the performance of different estimation approaches under three different data generating processes, first with a discrete covariate X and then with a continuous one.

[6]See Appendix A.3 for proof, with similar results targeting the ATT in Hazlett [2020]. See Zhao and Percival [2017] for a detailed analysis of identification and double-robustness of mean balancing weights.

2.3.1 Discrete covariate

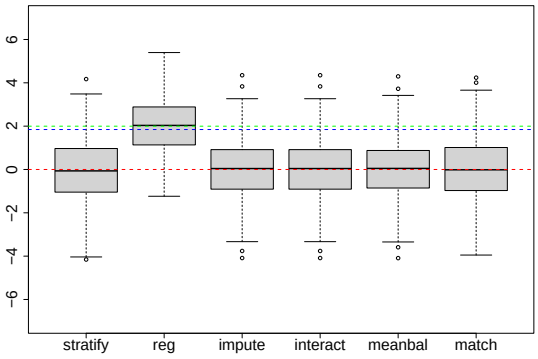
The first three DGPs involve a binary treatment variable D, a discrete covariate X in the range $[-3, 3]$, and an outcome Y which depends on D and X. For each simulation setting, the tables in Figure 2.1 show the possible values of X, the corresponding probability of treatment, probability that $X = x$, and the average treatment effect τ_x for the subgroup where $X = x$. In all simulations, noise is added to the outcome to achieve an R^2 of 0.33 between the systematic (noiseless) portion of Y and the final Y with noise.

As expected, when there is a high level of heterogeneity in treatment effect and in probability of treatment, the regression adjustment estimate will have a substantial amount of bias. Figure 2.1 shows the effect estimates when there is heterogeneity in both treatment effect and probability of treatment between subgroups. In the first two settings, the potential outcomes are linear in X. Here we see that the OLS estimate (`reg`) is heavily biased, and would lead investigators to conclude there is a statistically significant positive effect, though the true ATE is zero. Notably, slightly more extreme simulations would even make it possible for OLS to produce the incorrect sign for the treatment effect estimate. Meanwhile, regression imputation (`impute`), the Lin estimator (`interact`), mean balancing (`meanbal`), and matching (`match`) all successfully address this concern. We recommend the use of regression imputation or equivalently the Lin/interacted adjustment as a simple way to improve on conventional OLS estimation.

Naturally, in the case where the potential outcomes are not linear in the treatment and covariates, the estimates from the interacted model, imputation, and mean balancing are all biased for the ATE. This is expected, as the assumptions of both single and separate linearity are violated. Addressing such non-linearity requires non-linear estimators, such as matching. We also note that the variance-weighted estimate (using Expression 2.13) does not always reproduce the actual OLS estimate, as in the first simulation setting, where $P(D = 1|X)$ is not linear in X. This is easily avoided by fully saturating the model in X, although such a

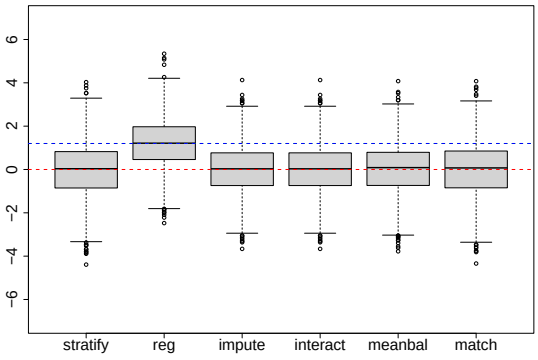
DGP1: $P(D|X)$ increasing in X; $\{Y(1), Y(0)\}$ linear in X

	X	$P(D\|X)$	$P(X=x)$	τ_x
1	-3	0.100	0.143	-9
2	-2	0.100	0.143	-6
3	-1	0.100	0.143	-3
4	0	0.500	0.143	0
5	1	0.500	0.143	3
6	2	0.500	0.143	6
7	3	0.500	0.143	9



DGP2: $P(D|X)$ increasing linearly in X; $\{Y(1), Y(0)\}$ linear in X

	X	$P(D\|X)$	$P(X=x)$	τ_x
1	-3	0.100	0.143	-9
2	-2	0.200	0.143	-6
3	-1	0.300	0.143	-3
4	0	0.400	0.143	0
5	1	0.500	0.143	3
6	2	0.600	0.143	6
7	3	0.700	0.143	9



DGP3: $P(D|X)$ increasing in X; nonlinear $Y(1), Y(0)$

	X	$P(D\|X)$	$P(X=x)$	τ_x
1	-3	0.100	0.143	5
2	-2	0.200	0.143	5
3	-1	0.300	0.143	0
4	0	0.400	0.143	-5
5	1	0.500	0.143	0
6	2	0.600	0.143	5
7	3	0.700	0.143	5

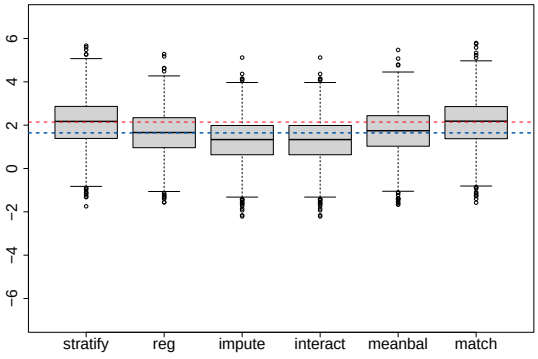


Figure 2.1: Estimates over 500 samples of size $n = 1000$. The dashed red line indicates the true ATE. The dashed blue and green lines show the (expected) regression coefficients reconstructed using the Angrist and Krueger [1999] weights (blue), or our more general weights (green).

manuever would not work for the continuous cases below.

The more general (but infeasible) weighting representation (Expression 2.20) reproduces the OLS estimate exactly regardless of the functional form of $P(D_i|X_i)$.

Standard errors

While these results show the expected variability in estimates under resampling from the given DGP, for an investigator working with one observed dataset, some form of *estimated* standard error is vitally important to inference. Table 2.1 reports the average analytically estimated standard errors for DGP1 above, still with discrete X. Results are similar in other settings.

For stratification, we take a weighted sum of strata-specific Neyman variances. Here and below we write expressions in the plug-in/analog sample estimator form.

$$\widehat{\mathbb{V}}(\hat{\tau}_{strat}) = \sum_{x \in X} \hat{p}(X = x)^2 \left(\frac{\widehat{\mathbb{V}}(Y_i|D = 1, X = x)}{n_1} + \frac{\widehat{\mathbb{V}}(Y_i|D = 0, X = x)}{n_0} \right) \tag{2.40}$$

For the simple and interacted regression, we calculate the HC2 standard error of the coefficient on the treatment variable. For regression imputation, we calculate the standard error again in the Neyman style,

$$\widehat{\mathbb{V}}\left(\frac{1}{n} \sum_i (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) \right) = \widehat{\mathbb{V}}\left(\frac{1}{n} \sum_i (\gamma X_i - \alpha X_i) \right) \tag{2.41}$$

$$= \widehat{\mathbb{V}}(\gamma \bar{X} - \alpha \bar{X}) \tag{2.42}$$

$$= \bar{X}^T \widehat{\Sigma}_1 \bar{X} + \bar{X} \widehat{\Sigma}_0 \bar{X} \tag{2.43}$$

where $\hat{\Sigma}_1$ and $\hat{\Sigma}_0$ are the estimated variance-covariance matrices for the treatment model and the control model, respectively.

For mean balancing we show two types of analytical standard errors. First, we use the (HC2) standard error from the weighted regression of just Y on D. Second, `meanbal-adj`

uses the HC2 standard errors from a regression of Y on D and X, again with the estimated weights. Both methods produce identical point estimates (when perfect mean balance is achieved by the weights), but the analytical standard errors of the `meanbal-adj` approach benefit from partialing out the X. This is akin to how conventional OLS standard errors, under a fixed design, partial X out of Y so that the estimates are built on the conditional/residual variance of Y rather than the total variance. For matching we use the Abadie-Imbens standard error [Abadie and Imbens, 2006].

	bias	rmse	avg analytical SE	empirical SE
stratify	-0.021	1.431	1.478	1.432
reg	2.021	2.376	1.264	1.250
impute	0.005	1.315	1.321	1.316
interact	0.005	1.315	1.321	1.316
meanbal	-0.001	1.328	1.812	1.329
meanbal-adj	-0.002	1.328	1.379	1.329
match	-0.005	1.434	1.514	1.436

Table 2.1: Simulation results under DGP1. The bias and rmse columns repeat information shown graphically in Figure 2.1. The empirical SE gives the actual standard deviation of the estimates across resamples. The average analytical SE is the average of the standard errors that would have been computed analytically in each iteration, using the estimators described in the text.

We find, first, that the empirical standard errors are extremely similar across the methods relying on single or double linearity (`reg`, `impute`, `interact`, `meanbal`, `meanbal-adj`). The approaches not relying on single or separate linearity (`stratify` and `match`) show somewhat larger standard errors, as expected given their greater flexibility. Second, each method's

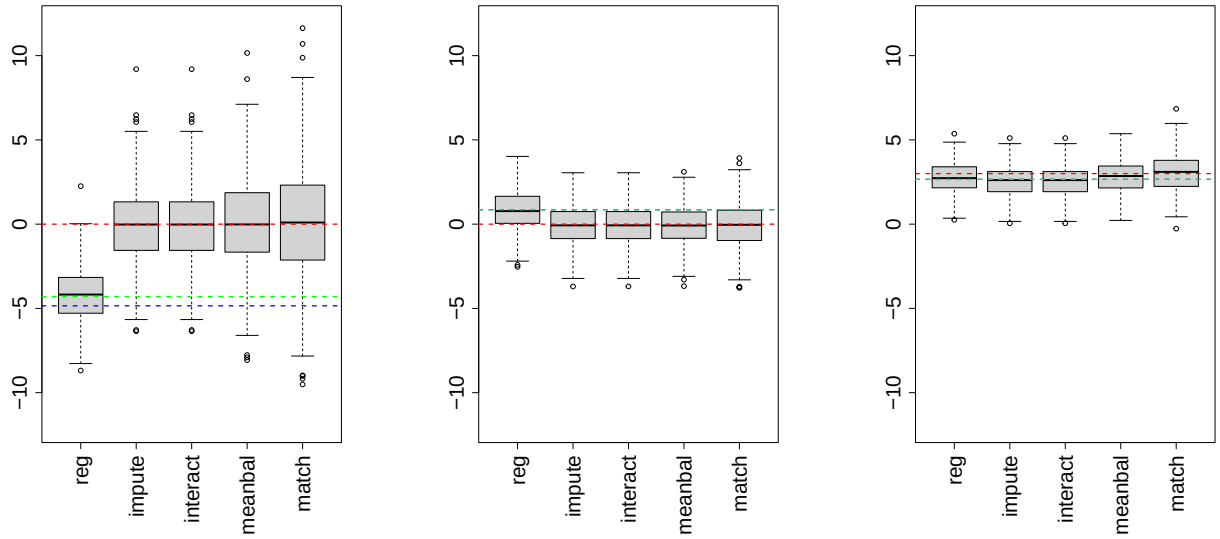
average analytical standard is within 5% of the empirical standard error, with the exception of `meanbal`, with an average analytical standard error almost 40% larger than the empirical value. This is repaired, however, by using `meanbal-adj`.

Finally, some investigators may be concerned with efficiency in the sense of sampling variability in the estimate and its consequences for inference. This can be understood as a bias-variance tradeoff. The comparison between OLS (`reg`) and the Lin/interaction approach (`interact`) offers a simple starting point, since the later will add one additional parameter to the regression for every covariate dimension. Because the number of observations is large relative to the number of covariates, this has very little impact on the estimate's variability across resamples. For example in DGP 1, Table 2.1 shows that the empirical SE is only about 5% larger for `interact` than for `reg`. The behavior of `impute` is of course identical. The empirical SE from `meanbal` and `meanbaladj` are similarly only 6% larger than from `reg`. If an investigator is primarily concerned with root mean square error (RMSE) of the estimates around the true value, RMSE values fall by nearly half for each of these methods relative to `reg`. That is, the small loss of efficiency in these settings (increasing variance) is more than made up for by reductions in bias as they factor into the RMSE.

It is important to recognize, however, the favorable nature of our simulation setting in this regard. If the number of covariates was large enough relative to the sample size, if treatment probability varied little by stratum of X, and/or if efficiency was a greater concern than bias or RMSE, then investigators might have cause to prefer OLS and adopting its weighted-ATE as the target estimand for their inferential purposes.

2.3.2 Continuous covariate

While we have considered discrete X thus far, we do so for the sake of intuition regarding strata, but the lessons apply to settings with continuous X as well. In the simulations shown in Figures 2.2a and 2.2b, X is randomly sampled from uniformly from $[-3, 3]$, and $Y(1) = Y(0) + 3X$. However, $P(D|X)$ takes on a different form in each setting. For the



(a) X is randomly sampled from Unif[-3,3], $P(D|X) = \frac{1}{1+e^{-2-x}}$, and $Y(1) = Y(0) + 3X$

(b) X is randomly sampled from Unif[-3,3], $P(D|X) = 0.1X + 0.4$, and $Y(1) = Y(0) + 3X$

(c) X is randomly sampled from Unif[-3,3], $P(D|X) = 0.1X+0.4$, and $Y(1) = Y(0)+ X^2$.

Figure 2.2: Effect estimates for 200 samples of size n=1000 under data generating processes where X is continuous and different estimation strategies. The dashed red line is set at the true ATE. The dashed blue line shows the average (over population) value for the variance-weighted formulation of the weights per Angrist and Krueger [1999]. The dashed green line shows the (expected) regression coefficients reconstructed using our more general weights.

first continuous specification, $P(D|X)$ takes a logistic form. In the second, $P(D|X)$ increase linearly with X. For the third specification, $P(D|X)$ increases linearly in X, but $Y(1)$ is nonlinear in X.

In each of these settings, we see that the OLS estimate is biased for the true ATE. As before, the interaction, imputation, and mean balancing estimators perform well, except in Figure 2.2c where even separate linearity fails.

2.3.3 Covariate adjustment in experiments, block randomization, and baseline-free average marginal effects

These lessons may at first seem to be directed to researchers working with observational data such that conditioning on X is a requirement. However, they also apply in settings where conditioning on X is used to improve precision/ minimize finite sample differences from the expected result. Our advice essentially echoes that of Lin [2013] and a broad subsequent literature calling for a model that interacts treatment with the (centered) covariates. As noted above, this is identical to analogous imputation/g-computation/T-learner approaches when using OLS for the underlying models.

Consider first the cases where investigators have a randomized experiment, but adjust for covariates in the analysis to improve precision. When D is fully randomized, the probability of treatment across values of X will typically not vary greatly, except by chance. This implies that the bias due to regression's weighting behavior will typically be small. Nevertheless, any discrepancy between the ATE and coefficient on account of these weights can be avoided entirely.

Second, block randomization is a powerful tool often employed by experimentalists to reduce finite sample deviations from expectation. For the estimates to reflect the reduced variance this affords, it is standard to regress the outcome (Y) on the treatment (D) and block indicators (B) as fixed effect covariates. The weights implied by regression of Y on D and block indicators B would be given by

$$\hat{\tau}_{BFE} = \frac{\sum_{b=1}^{P} \hat{\tau}_b P(D_i = 1|B_i = b)(1 - P(D_i = 1|B_i = b))P(B_i = b)}{\sum_{b=1}^{P} P(D_i = 1|B_i = b)(1 - P(D_i = 1|B_i = b))P(B_i = b)} \tag{2.44}$$

where it is innocuous to assume that $P(D = 1|B)$ is linear in the block indicators, as this regression would be fully saturated in B. If all blocks have the same probability of treatment, the weights are equal, so the regression coefficient will represent the difference in means per block, averaged over the size of each block, i.e. the ATE. This will often be the

case. However, if the treatment probability varies by block, either by design or otherwise, then the resulting coefficient estimate would not in general be the ATE. Rather, regression will put higher weight on blocks with probabilities of treatment nearer to 50%.

As above, this is simply a consequence of misspecification generated by heterogeneous treatment effect estimates by block. Including interactions between the treatment and the block fixed effects would address this, under the separate linearity assumption,

$$\mathbb{E}[Y_i(1)|B_i] = \alpha_0 + \alpha_1 \mathbb{1}\{B_i = 1\} + .. + \alpha_b \mathbb{1}\{B_i = P\} \tag{2.45}$$

$$\mathbb{E}[Y_i(0)|B_i] = \gamma_0 + \gamma_1 \mathbb{1}\{B_i = 1\} + ... + \gamma_b \mathbb{1}\{B_i = P\} \tag{2.46}$$

One complication when applying the interacted approach here is that care must be taken regarding the interpretation due to the interaction. In typical usage, one block indicator will be omitted to avoid co-linearity with the intercept. If no centering/de-meaning is done on the block indicators, the coefficient estimate would represent the estimated effect (difference in means) in whichever block had its indicator omitted. The solution of centering covariates [Lin, 2013] as utilized above is now complicated by this omission. It is possible to omit the intercept, rather than the indicator for one block, and utilize the centering. However, a more general solution is to avoid any consideration of centering and omitting one level/the intercept, and works when dealing with one or multiple categorical variables. This is to simply compute the marginal effect estimate for each observation, in whatever block it is in, $\frac{\partial Y}{\partial D}\big|_{B=b}$. Averaging these across observations (giving equal weight to each observations) produces the average marginal effect (AME),

$$\hat{\tau}_{AME} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\frac{\partial Y}{\partial D}}\Big|_{B=b} \tag{2.47}$$

where $\widehat{\frac{\partial Y}{\partial D}}\big|_{B=b}$ is the estimated marginal effect in block b where this individual unit is found.

For example, in the regression

$$Y_i = \beta_0 + \tau D + \beta_1 \mathbb{1}\{B_i = 1\} + \beta_2 \mathbb{1}\{B_i = 2\} + ... + \beta_P \mathbb{1}\{B_i = P\} \tag{2.48}$$

$$+ \alpha_1 \mathbb{1}\{B_i = 1\} D + \alpha_2 \mathbb{1}\{B_i = 2\} D + ... + \alpha_P \mathbb{1}\{B_i = P\} D + \epsilon_i \tag{2.49}$$

the estimated marginal effect in block $B = b$ is $\hat{\tau} + \alpha_p$, and so the average marginal effect of interest over the whole sample is simply

$$\hat{\tau}_{AME} = \hat{\tau} + \frac{1}{n} \sum_{i=1}^{n} \sum_{b=1}^{P} (\hat{\alpha}_p \mathbb{1}\{B_i = b\}) \tag{2.50}$$

Using this approach to interpret the fitted regression with interactions always produces an estimate with the desired interpretation of "the estimated marginal effect, which can differ by block, averaged over blocks according to how many observations fall in each." In the case of block indicators or other categorical variables, this approach avoids errors in relation to choices about what levels to omit in the regression, whether to omit the intercept, or having to center these variables. The variance for this estimator is

$$\mathbb{V}(\hat{\tau}_{AME}) = \mathbb{V}(\hat{\tau}) + \sum_{p=1}^{P} P(B_i = p)^2 \mathbb{V}(\hat{\alpha}_p) + 2 \sum_{k<j} P(B_i = k) P(B_i = j) \text{Cov}(\hat{\alpha}_k, \hat{\alpha}_j) \tag{2.51}$$

$$+ 2 \sum_{p=1}^{P} P(B_i = p) \text{Cov}(\hat{\tau}, \hat{\alpha}_p) \tag{2.52}$$

For comparison, we also consider two other approaches. One practice is to weight the blocked fixed effects regression by the (stabilized) inverse probability of treatment assignment for each block,

$$w_i^{\text{IPW}} = \frac{P(D_i = 1)}{P(D_i = 1|B = b_i)} D_i + \frac{P(D_i = 0)}{P(D_i = 0|B = b_i)} (1 - D_i) \tag{2.53}$$

These weights are constructed so that within each block, the treated and control units make up equal weighted proportions. They therefore neutralize any differences in the $P(D = 1|B)$ across blocks. If we perform a weighted regression of Y on D and the block dummy variables

using these weights, the coefficient estimate for D will be unbiased for the ATE.[7]

Finally we also the stratification estimator (Expression 2.5) but with blocks as strata,

$$\hat{\tau}_{\text{blockDIM}} = \sum_{b=1}^{|B|} \hat{P}(B = b)(\bar{Y}_{D=1,B=b} - \bar{Y}_{D=0,B=b}). \tag{2.54}$$

Figure 2.3 compares estimates from the "plain" block fixed effects (`block FE`) regression to the interacted regression with the average marginal effect interpretation (`block FE interact`), the IPW-weighted regression (`block FE IPW`), and the block DIM average (`mean block DIM`) in a simulation setting where treatment probability varies by block and there is heterogeneity in treatment effect. As expected, the block fixed effects OLS model suffers from the weighting problem. All three alternative approaches produce unbiased and identical estimates.

We recommend the use of one of these alternative estimators to improve upon estimation of the ATE under block randomization.[8]

2.4 Conclusion

Given the wide use of OLS regression of Y on D and X in practice, and the potential gap between the coefficient and target quantities like the ATE, prior work has provided several forms of guidance to practitioners regarding this weighting problem. Humphreys [2009] notes the conditions under which the OLS result will be nearer to ATT or ATC, and observes monotonicity conditions under which the coefficient estimate will fall between these

[7] A weighted difference in means (rather than regression including B) with these weights would produce the same point estimate, but the improvement in efficiency obtained under the block randomization design will not be fully reflected in the estimated standard error. This occurs because the block indicators will be orthogonal to treatment (once weighted) and so do not affect the coefficient estimate on treatment, but the omission of B prevents the model from reducing the residuals that enter the standard error.

[8] These results are consistent with those demonstrated in https://declaredesign.org/blog/posts/biased-fixed-effects.html, which use the DeclareDesign simulation approach [Blair et al., 2019] and conclude that the mean blockwise DiM/ stratification approach is suitable.

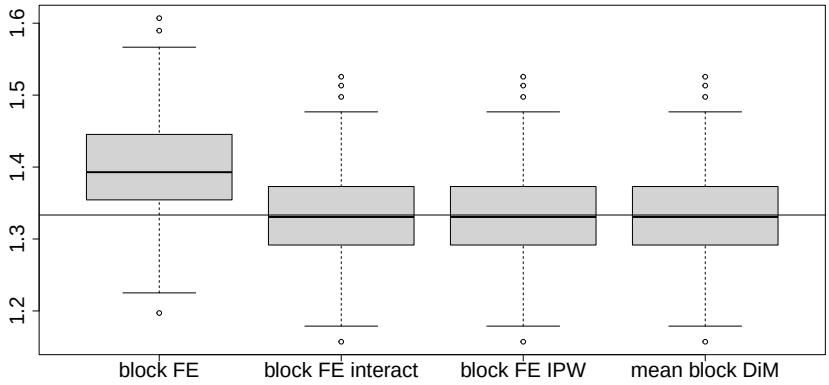


Figure 2.3: Effect estimates for 500 samples of size $n = 1200$ each. D is assigned by (complete) randomization within each block with probability $P(D|B)$, and $Y = 2 + D^*\tau$. For blocks 1-6, $P(D|B) = (0.25, 0.25, 0.5, 0.25, 0.25, 0.5)$ respectively, and $\tau = 4^*P(D|B)$. The black line indicates the true ATE.

two endpoints. Aronow and Samii [2016] suggest characterizing the "effective sample" for which the regression coefficient would be an unbiased estimate of the ATE by applying the variance weights to the covariate distribution. Other papers have provided diagnostics for quantifying the severity of this bias. Chattopadhyay and Zubizarreta [2023] characterize the implied unit-level weights of regression adjustment (as we do as an intermediate result), and propose diagnostics that use these weights to analyze covariate balance, extrapolation outside the support, variance of the estimator, and effective sample size. These ideas relate closely to those of Aronow and Samii [2016] in that they compare the covariate distribution targeted by the estimator to that of the target population. Słoczyński [2022] derives a diagnostic related to variation in treatment probability,

$$\delta = \frac{\rho^2 \mathbb{V}[P(D|X)|D=1] - (1-\rho)^2 \mathbb{V}[P(D|X)|D=0]}{\rho \mathbb{V}[P(D|X)|D=1] + (1-\rho)\mathbb{V}[P(D|X)|D=0]} \tag{2.55}$$

where ρ is the unconditional probability of treatment. The bias due to heterogeneity will be $\delta(\tau_{ATC} - \tau_{ATT})$ where τ_{ATC} and τ_{ATT} are the average treatment effect among treated and control, respectively. Thus δ captures one aspect of the potential for bias—variation in

treatment probability—but reasoning about the severity of bias requires knowledge of the heterogeneity in treatment effects, which is unknown to the investigator.

However, investigators could instead consider different modeling choices that avoid this "weighting problem" altogether. We have sought to clarify the theoretical considerations required to do so, first by understanding the weighting problem as a symptom of misspecification when the treatment effect (and probability of treatment) vary with X. We provide a more general expression for the "weights of regression" by algebraically manipulating a representation of the OLS coefficient. The key fact is that the apparent "weights of regression" simply account for the way misspecification alters the coefficient away from the ATE. Second, viewed this way, a natural proposal for research practice is to rely instead on specifications that will not necessarily be violated by effect heterogeneity. To focus on minimal changes in practice, the conventional *single linearity* assumption (that Y is linear in D and X) can be relaxed at least to the *separate linearity* assumptions (each potential outcome is separately linear in X). This recapitulates equivalent assumptions employed in analyses by Hahn [2023], Słoczyński [2022], Kline [2011] and Imbens and Wooldridge [2009]. It also clarifies which estimators are suitable to avoid these weights, and whose properties we can consider by comparison. Fortunately, a number of existing, straightforward estimation approaches produce unbiased ATE estimates under this assumption. First, regression imputation (including g-estimation and the T-learner) with OLS as well as including the interaction of X and D (as in Lin, 2013) are all suitable, and identical to each other in this setting. This longstanding approach dates back to at least Peters [1941] and Belson [1956], as noted by Chattopadhyay and Zubizarreta [2024] who label it as "multi-regression". In addition, "mean balancing" estimators applied to the ATE—those that choose weights to achieve the same mean of X for the treated and control group as in the full sample—are also unbiased under separate linearity. These are not identical to the above as they avoid fitting any model, and our simulations suggest this property can partially mitigate bias when even the separate linearity assumption fails. These

approaches can be useful not only in observational studies (so far as investigators can claim conditional ignorability), but also when using covariate adjustment after randomization, including the special case of analyzing block randomized experiments.

Avoiding these weights through these alternatives may be desirable in many cases, but is not without cost of limitation. The natural cost to first consider (relative to regression under single linearity) is the potential loss in efficiency. The interact/imputation/g-computation/T-learner approach adds an additional nuisance parameter per covariate. The circumstances investigated here—with just one covariate, a high degree of treatment effect heterogeneity, and large variation in treatment probability—clearly favor these approaches, where they eliminate bias and reduce RMSE by roughly half, while increasing the standard error by only about 5% in our settings. However, there may be settings in which the improved bias does not so assuredly dominate the efficiency cost. For example, where there are many covariates relative to the sample size, little difference in the probability of treatment by stratum, or little theoretical reason to expect treatment effect heterogeneity, investigators may have cause to prefer OLS if efficiency concerns are of paramount interest. In those cases, diagnostics or interpretational aids that gauge the severity of the misspecification bias (weights) may be useful, depending on how investigators prefer to tradeoff bias and efficiency in their inferential or decision-making goals.

Finally, when the linearity in potential outcomes assumption is not satisfied, none of these methods can guarantee unbiasedness. Our work here regards only the relaxation from single linearity to separate linearity. Investigators not satisfied with the linear approximation to treatment effects in this way could reasonably consider more flexible approaches—though doing so may incur additional uncertainty costs as illustrated in Table 2.1. Nevertheless, we agree with assessments such as Keele et al. [2010] and Aronow and Samii [2016] that current practice largely remains reliant on linear approximations to adjust for covariates while hoping to interpret the result as a meaningful causal effect. Our analysis suggests that, especially where there are enough observations relative to covariates to support

imputation/g-computation/T-learner, interactive regression, or mean balancing, this may be preferable to suffering regression's "weighting problem", as it avoids this form of bias under slightly weaker assumptions and requires only a small change to current practice.

CHAPTER 3

Post-treatment problems: What can we say about the effect of a treatment among sub-groups who (would) respond in some way?

3.1 Introduction

It is widely understood that, even in a randomized trial, limiting the analysis to a sub-group defined by a post-treatment variable compromises treatment effect estimates—even within that sub-group.[1] Yet, as cataloged by Montgomery et al. [2018], many investigators have attempted such analyses regardless. The temptation to consider such questions is understandable because, in many scenarios, the effect within such a sub-group is of great interest. For example, we may be interested in the average effect among those who complied with a treatment, who received a well-implemented treatment, who were attentive, or who demonstrated some other mechanistic reaction considered important or even a pre-requisite for a treatment effect. In some cases, these "effect among those who..." questions are of primary interest. In other cases, they arise as post hoc but important questions, particularly if no effect was detected among all study units but we recognize that implementation challenges, low compliance, or inattentive participants may have attenuated the estimated effect.

[1]The term "post-treatment" denotes here any variable that may be influenced by treatment. Variables measured after treatment assignment are not necessarily post-treatment, if they cannot be affected by treatment. Conditioning on such unaffected variables does not introduce the inferential problems discussed here.

We propose a straightforward approach for making bounded inferences about the "Treatment-Reactive Average Causal Effect" (TRACE), the total effect within the sub-group that, if treated, would have realized a specified value of a post-treatment variable. This differs from existing estimands such as the survivor average causal effect (SACE), mediation-related quantities (e.g. the controlled direct effect, CDE), "as-treated" or "per-protocol" estimands, and the local average treatment effect (LATE). The resemblance between the TRACE and the LATE is of particular interest to those familiar with the instrumental variables (IV) framework. As we examine below, the TRACE differs definitionally from the LATE in that it averages over both compliers and always-takers, rather than compliers alone. More importantly, while IV requires the exclusion restriction, meaning here that treatment can have no "direct effect" on the outcome, we are interested in a total effect of treatment on the outcome, which arises in part or whole from that direct effect.

Our partial identification approach for the TRACE relies on positing the total effect of the treatment in the sub-group that would *not* have met that specified criterion even if treated (TRACE(0)). Researchers are not expected to know TRACE(0), but we argue that they are often positioned to place restrictions on it. For example, it may be possible to reasonably argue about the sign of TRACE(0), that it is near zero, or that it is no larger than the TRACE. Such arguments limit the possible range for the TRACE. In situations where researchers estimate a null average effect, our approach can help reason about whether this reflects a true null effect or instead results from poor implementation. In addition, where the mediator is observed for all units and is monotonically related to treatment, an alternative set of bounds can be derived that does not rely on assumptions on TRACE(0). When these bounds partially overlap with those derived by reasoning about TRACE(0), their intersection provides a more informative range than either approach alone.

Finally, in circumstances where the mediating event must occur in order for the outcome to occur, it is sometimes plausible to argue that TRACE(0) = 0, which leads to

point identification of the TRACE. This is an especially useful result because it provides identification in circumstances where the direct effect of D on Y in the "reactive" sub-group is thought to exist and is of interest, violating the exclusion restriction of IV.

We illustrate our approach, first, by demonstrating how a small number of sample statistics could allow point identification of the effect of police-perceived race on police violence during traffic stops, an area where post-treatment selection into the data has been a serious problem [Knox et al., 2020]. We then analyze a randomized trial of a community policing program that produced largely null average effects, but in which investigators observed limited implementation of key components of the prescribed intervention [Morse, 2024]. Finally, we examine the persuasive effects of a 10-minute conversation on support for transgender rights, among participants whose feelings toward transgender individuals became more favorable following the intervention [Broockman and Kalla, 2016].

3.2 Background

3.2.1 The problem

Consider the graph in Figure 3.1. For concreteness, and previewing one of our applications, let D be the treatment assignment for a community policing program we wish to study. Let M be some measure of whether or not the program was implemented in the community. Finally, let Y be an outcome of interest, such as crime rates. If we were interested in the effect of the program on the outcome, it would be natural to compare the outcomes in the treated group ($D = 1$) with the outcomes in the control group ($D = 0$). However, in doing so, we would include communities in the treated group that did not actually implement the program, likely attenuating the treatment effect.

A natural temptation is to look at only the group that implemented the program, effectively, i.e. selecting/conditioning on M. However, this is problematic for two reasons. First, since M is post-treatment, conditioning on it might block part of the effect of D on

Y. Second, it introduces a new bias. We cannot typically rule out unobserved confounders of M and Y (shown as U). Because M is a consequence of both D and U, it is a collider [Pearl, 2009], and conditioning on it can create an association between D and U, leading D and Y to become associated (via $D \leftarrow U \rightarrow Y$) for reasons other than the causal effect of D.

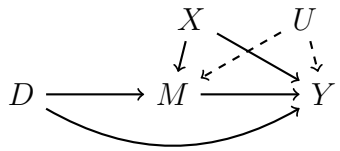


Figure 3.1: *Causal structure of concern.* A treatment (D), which is unconfounded with the outcome of interest (Y) may affect M in some sub-group. While some confounders of M and Y may be measured (X), we cannot rule out the presence of unobserved common causes of M and Y (U).

To avoid confusion when comparing this approach to IV below, note that the estimand of interest here is the effect of the treatment (D) on the outcome (Y) within a sub-group that will be related to M. While the causal structure resembles that assumed in the IV framework (if the direct effect $D \rightarrow Y$ can be ruled out), in that setting the effect of interest is that of M on Y, while D plays the role of instrument. That is, the total effect in our setting corresponds to the intent-to-treat, or reduced form effect in the analogous IV setting.

3.2.2 Examples

This arrangement appears in a wide variety of settings, which we organize into four broad types.

Type 1. Manipulation, attention, and implementation checks. In these cases, M represents a variable that, while not fully necessary for an effect of D on Y, lies on

a path thought to be of central importance for D to be effective. This encompasses our motivating example: a randomized trial in which some communities are assigned to receive a community policing program (D), and where the realization of that program on the ground (M) varies widely. We may wish to examine the effects of the program in communities that implemented it to some satisfactory degree. [2] Other examples include cases where M measures participants' compliance with a protocol, or responses to manipulation checks or attention screens performed after the treatment.

Type 2. Mechanism of interest. In some settings, the way a participant reacts to the treatment is thought to be important to the treatment's effect. For example, in our analysis of Broockman and Kalla [2016] below, we wish to learn if the effect of a perspective-taking intervention on support for transgender non-discrimination policies depends on whether the intervention reduced animus toward transgender individuals. This setting resembles questions in the mediation literature regarding direct vs. indirect effects, but as we describe in Section 3.3.5, the TRACE addresses an entirely different causal question. Our approach also avoids problematic assumptions regarding the absence of mediator-outcome confounders, albeit at the cost of partial rather than point-identification.

Type 3. "Necessary condition" mediators. In this setting, M is a necessary condition for Y to possibly equal 1. An example can be found in studies of the effect of a driver's race (D), as perceived by police, on the potential for police violence (Y) during a traffic stop (M). The chance of a traffic stop occurring ($M = 1$) may be itself a function of police-perceived race. Police violence (Y) can be defined specifically as "police violence during a traffic stop," which ensures that no police violence of this kind can occur if there is no stop, i.e. $Y = 0$ if $M = 0$. While this is merely an extreme case of M being a mediator, we give it a separate

[2]Note that in some cases, M variables capturing implementation may only be measurable in the treated group. For example, if the program entails facilitated community meetings with police, a variable like "attendance" at such meetings may be poorly defined in the control group. Such M variables are difficult to integrate into conventional analyses but can still be utilized for the TRACE.

category because the necessity of $M = 1$ in order for $Pr(Y = 1) > 0$ can lead to point identification as detailed below.

Type 4. Treatment-responsive measurement of M. In another setting, we may wish to know the effect of D on Y among those with some value of M, where measurement of M is affected by D. For example, suppose we wish to know the effect of cancer screening on survival, among those who would have received a positive result for cancer in *that* screening (thus ruling out always-takers). The result here depends on our ability to reason about the effect of screening on survival for patients who, if screened, would have screened negative, on its own or by comparison to the effect in the group who would screen positive. This engages a complicated set of concerns about the effects of screening on behaviors, the false negative rate, potential harms due to false positives, and other challenges that we examine elsewhere.

3.3 Proposal

3.3.1 Setup and Notation

Continuing with the notation defined above, the total effect (TE) of D on Y is defined as

$$TE \equiv \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))] \tag{3.1}$$

where $M_i(d)$ denotes the "potential mediator" [Imai et al., 2010b, 2014]. That is, $M_i(1)$ is the value unit i would have for M if treated, and $M_i(0)$ the value of M for the same unit i had it not been treated. Because $M_i(d)$ is the value of M_i under treatment $D = d$, by consistency $Y_i(d, M(d)) = Y_i(d)$. Thus we can simply write the TE as

$$TE \equiv \mathbb{E}[Y_i(1) - Y_i(0)] \tag{3.2}$$

where $Y_i(\cdot)$ in our usage always refers to $Y_i(d)$. When it is necessary to consider a value of M other than $M(d)$, we explicitly write both arguments in the form $Y_i(D = d, M = m)$. This is

not needed for the quantities we propose; however, it becomes relevant for mediation-related estimands we consider by comparison.

3.3.2 Defining the TRACE

Our target quantity of interest is a total effect (TE) of D on Y, but averaged over only units with $M_i(1) = 1$, i.e. those that if treated would have $M = 1$.[3]

$$\text{TRACE} \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid M_i(1) = 1] \tag{3.3}$$

$$= TE[M_i(1) = 1] \tag{3.4}$$

We also define the total effect among the units who, if treated, would have $M = 0$, which we call TRACE(0):

$$\text{TRACE}(0) \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid M_i(1) = 0] \tag{3.5}$$

$$= TE[M_i(1) = 0] \tag{3.6}$$

3.3.3 Partial identification of the TRACE

Identification of the TRACE is complicated by M's status as a post-treatment variable combined with the possible existence of unobserved common cause confounders of M and Y. If investigators believed all common causes of M and Y were observed (i.e., there was no U on Figure 3.1), it would be possible to identify the effect of D on Y conditioning on M, because all the paths opened by this conditioning could be closed again. However, we consider this unrealistic in many settings, and of no interest for the problems on which we focus. For example, if D is a community-driven development program, and M is an indicator of whether a project chosen by the community was actually built, one expects many

[3]Representing the TRACE on a graphical causal model similar to Figure 3.1 is possible but complicated, as it requires modifications to show $M(d)$ on the graph; see B.2.

unobservable features to be common causes both of "which places manage to implement the project" and the outcomes of interest Y. We therefore focus on the case in which such $M-Y$ confounding cannot be ruled out.

By the law of iterated expectations,

$$TE = \text{TRACE} \cdot Pr(M(1) = 1) + \text{TRACE}(0) \cdot Pr(M(1) = 0) \tag{3.7}$$

$$\text{TRACE} = \frac{TE - \text{TRACE}(0) \cdot Pr(M(1) = 0)}{Pr(M(1) = 1)} \tag{3.8}$$

Given randomization of D, the TE is identifiable, as are $Pr(M(1) = 1)$ and $Pr(M(1) = 0)$:

$$\begin{aligned} Pr(M(1) = m) &= Pr(M(1) = m \mid D = 1) \quad \text{randomization of } D \\ &= Pr(M = m \mid D = 1) \quad \text{consistency of } M(d) \end{aligned}$$

The only unknown on the right hand side of Expression 3.8 is TRACE(0). Thus, for any postulated ranges of values on TRACE(0), we can recover the implied range of values for the TRACE. Note that the range of resulting TRACE values will be narrower when $Pr(M(1) = 0)$ is smaller, i.e. when the fraction of "implementers" is higher. Additionally, the entire procedure can be done conditioning on values of X, i.e. $X = c$.

Anticipated types of assumptions on TRACE(0)

We emphasize that the investigator *is not expected to know, nor postulate a single value of, the effect of D on Y in the group with $M(1) = 0$, i.e. TRACE(0).* On the contrary, maintaining conservative uncertainty over this quantity is central to the approach and its credibility. Nevertheless, inferences can sometimes be made for the price of defensible claims regarding the reasoned bounds on this quantity. Our experience to date suggests several common forms of assumptions, though others are possible.

- *Arguments that $TRACE(0) = 0$.* In some cases we have reason to believe that, absent the activation of M, there can be no effect of D on Y. For example, in the policing

application, if the police do not make a stop ($M = 0$), there cannot be police violence during that stop. Together with an assumption of no defiers (or no effect among defiers), this would support the assumption of TRACE(0) = 0. This assumption is distinct from the exclusion restriction as it allows a direct effect. We elaborate on this distinction in Section 3.3.5.

- *Arguments that* $TRACE(0) \approx 0$. In other circumstances we cannot argue that an effect is precisely impossible when $M = 0$, but are equipped to argue there should be at most a very small effect. For example, if the treated are enrolled in an exercise program but never participate, it would be unlikely to have a substantial effect on their health. However, we cannot rule out some small effect, e.g. through triggering other behavioral changes.

- *Arguments that* $|TRACE(0)| < |TRACE|$. In many other cases, we cannot be certain about the impact of D on Y in units with $M(1) = 0$, but we can be convinced that this effect is less than the effect among those for which $M(1) = 1$. This means that, while we do not know the TRACE, we may be willing to argue that TRACE(0) is smaller (in absolute value) than it. For example, in the community policing application below, we argue that a community randomly assigned to receive the community policing intervention but not showing evidence for its implementation or awareness of the program in the community will not have as large an effect as in those places where we do see evidence that the program occurred and was seen to occur.

- *Arguments that TRACE(0) is opposite in sign to the TRACE.* In some cases investigators may argue that those with $M(1) = 0$ show an effect opposite in direction to that in the group with $M(1) = 1$. For example, if a development program in a village promises to provide something ($D = 1$) but this fails to materialize ($M = 0$), it may engender greater mistrust or anger.

Investigators may argue for one of these assumptions or others to arrive at a defensible

range of TRACE results. Alternatively, this approach allows users to transparently show what values of TRACE(0) are required to reach a particular conclusion about the TRACE (e.g. a significant effect in the intended direction among implementers). The researcher (and reader) can then reason about plausibility and defensibility of these assumed values.

3.3.4 Estimation and inference

For estimation, we use sample analogs for each element in Equation 3.8, i.e. $\widehat{TE} = \overline{Y}_{D=1} - \overline{Y}_{D=0}$, and $\widehat{Pr}(M(1) = m) = \widehat{Pr}(M = m \mid D = 1)$. When conditioning on covariates X is required, $\widehat{TE}$ can instead be estimated via a suitable model. Postulated values of TRACE(0) are then plugged in to produce corresponding values of the TRACE. For inference we employ the percentile bootstrap: resampling the data with replacement, re-estimating the TRACE in each replicate (holding TRACE(0) fixed at its postulated value), and forming 95% confidence intervals using the 2.5th and 97.5th percentiles of the resulting empirical distribution.

3.3.5 Connections to other approaches

Several existing estimands address queries related to post-treatment variables, but differ from the TRACE in the groups they include and the identification opportunities they present. In making these comparisons, it is helpful to follow the structure of the principal stratification framework [Frangakis and Rubin, 2002], on which our estimand and others we consider here can be constructed. Specifically, imagine that all units fall into one of four possible strata: compliers ($M(1) = 1, M(0) = 0$), defiers ($M(1) = 0, M(0) = 1$), always-takers ($M(1) = 1, M(0) = 1$), and never-takers ($M(1) = 0, M(0) = 0$). The TRACE is the average treatment effect for always-takers and compliers, while TRACE(0) is the average effect over never-takers and (if present) defiers.

Table 3.1 compares related estimands. Though ours is unique (excepting equivalences in special cases), the primary emphasis of our contribution lies in the relatively straightforward

Estimand	Definition	Among Whom	Identification Assumptions
LATE (IV)	$\mathbb{E}[Y(1) - Y(0) \mid M(0) = 0, M(1) = 1]$	compliers	relevance, exogeneity, exclusion restriction
SACE	$\mathbb{E}[Y(1) - Y(0) \mid M(0) = M(1) = 1]$	always-takers	monotonicity, $M(1) \perp\!\!\!\perp M(0)$, $\mathbb{E}[Y(1) \mid \text{AT}] = \mathbb{E}[Y(1) \mid \text{C}]$, $\mathbb{E}[Y(0) \mid \text{AT}] = \mathbb{E}[Y(0) \mid \text{D}]$ given X
CDE(m)	$\mathbb{E}[Y(D = 1, M = m) - Y(D = 0, M = m)]$	all units	unconfoundedness of $M - Y$
g-est./ per-protocol	$\mathbb{E}_x\mathbb{E}[Y(1) - Y(0) \mid M(0) = 0, M(1) = 1, X]$	compliers	(conditional) ignorability on $M - Y$ path
TRACE	$\mathbb{E}[Y(1) - Y(0) \mid M(1) = 1]$	compliers, always-takers	postulated TRACE(0))

Table 3.1: Comparison of different estimands for settings in which post-treatment variables are relevant. AT: always-taker; C: complier.

partial identification strategy it allows.

Comparison to instrumental variables. The Local Average Treatment Effect (LATE), also called the Complier Average Causal Effect, (Angrist et al., 1996, Imbens and Angrist, 1994, Baker and Lindeman, 1994) is

$$\mathbb{E}[Y_i(1) - Y_i(0) \mid M_i(0) = 0, M_i(1) = 1] \tag{3.9}$$

These estimands differ in two primary ways.[4] First, the LATE and TRACE differ definitionally in that the LATE considers the average effect among compliers, while the TRACE considers the average effect among compliers and always-takers. These estimands become equivalent when we can rule out always-takers (through one-way non-compliance), or when compliers and always-takers have the same effects.

Second, the LATE is the target estimand for the IV approach, which requires that D only affects the outcome through M, i.e. that there is no direct effect of D on Y. The existence of the $D \to Y$ path in our setup (see Figure 3.1) violates this core IV assumption, and is an important contributor to the total effect we are seeking to learn.[5] Consider, for example, cases where the investigator argues that TRACE$(0) = 0$. This may occur in "Type 3" settings where it is impossible for D to affect Y when $M = 0$, and we can further argue either for monotonicity (no defiers) or for no effect among defiers. The assumption of "no effect of D on Y when $M = 0$" may sound quite similar to the exclusion restriction, but they differ in a key way: the exclusion restriction requires the total absence of a direct $D \to Y$ effect, while TRACE$(0) = 0$ requires no direct effect only for those with $M(1) = 0$ (never-takers and, if present, defiers). This difference is critical in the settings we examine, where we *do expect a direct effect* of D on Y in the "reactive" sub-group. In the police stop violence case below, for example, we wish to study—not rule out—a possible effect of police-perceived race (D) on the use of violence (Y) when a stop occurs ($M = 1$).

Comparison to survivor average causal effect. The "survivor average causal effect" (SACE) arises in settings where the outcome is censored due to a post-treatment factor, such

[4]As noted briefly above, the IV causal structure relevant here would be the one that regards our M as the treatment, and our D as the instrument for it. Relatedly, the $Y_i(\cdot)$ in Expression 3.9 refers to $Y_i(m)$ in the IV setting, and the causal effect among compliers is thus the effect of M on Y, not the effect of D on Y we are investigating. That said, for compliers D and M are required to be the same, so the notational difference is immaterial in this particular case.

[5]For such settings, partial identification strategies have been developed for bounded potential outcomes [Manski, 1990, Manski and Pepper, 1998].

as survival [Tchetgen, 2014, Egleston et al., 2009, Imai, 2008, Hayden et al., 2005, Zhang and Rubin, 2003, Rubin, 2000, Robins, 1986], or sample selection/attrition [Semenova, 2020, Lee, 2005]. The SACE is defined as

$$SACE = \mathbb{E}[Y_i(1) - Y_i(0) \mid M_i(0) = 1, M_i(1) = 1] \tag{3.10}$$

This considers only the effect among always-takers, i.e. "always-survivors" when $M = 1$ is survival to endline. Because we cannot observe which units are always-takers, point identification strategies rely on assumptions that ensure mean values of $Y_i(1)$ or $Y_i(0)$ do not differ between certain principal strata. As an estimand, the TRACE differs from the SACE only by including compliers. However, the survival settings in which the SACE originates differ from those we have considered thus far because missingness in the outcome makes even the total effect unidentified, which poses a problem for our approach. Outside of the survival setting, the SACE is relevant when reinterpreted as the average effect among always-takers, i.e. where the M in question does not affect loss to follow-up (see e.g. Hudgens and Halloran [2006]). Existing work on partial identification for the SACE has proposed trimming bounds [Samii et al., 2025, Semenova, 2020, Lee, 2005, Imai, 2008, Zhang and Rubin, 2003, Horowitz and Manski, 1995] using worst-case values for $\mathbb{E}[Y(d) \mid \text{AT}]$. We discuss similar bounds for the TRACE in Section 3.3.6.

Comparison to mediation analysis. A variety of mediation approaches decompose the total effect into changes in the outcome caused directly by treatment and those caused indirectly through changes in the mediator (e.g. Robins and Greenland, 1992, Pearl, 2001, Imai et al., 2010a, 2014, Pearl, 2014, Acharya et al., 2016). Our causal query is not the same: we seek a *total* effect, for a subgroup defined by the potential value of the mediator. Given this important distinction, efforts we made to directly translate between these two approaches proved more complex than illuminating.

Intention to treat, as-treated, and per-protocol analyses. The intention-to-treat (ITT) approach evaluates causal effects based on the assigned treatment when not all participants adhere to treatment assignment, which is simply the total effect in our setting. Our proposal is essentially a formalization of the understanding that the ITT is "diluted", and a means of exploring the "un-diluted" values implied by varying quantitative assumptions on TRACE(0). Another approach to estimation under nonadherence is "as-treated" analysis, which estimates the effect of the realized treatment rather than the assigned treatment. This estimate will suffer bias from confounding if characteristics that moved a participant to treatment implementation are associated with characteristics that affect a participant's outcome. Similarly, some investigators employ "per-protocol" analyses, including only participants who adhered to treatment assignment. This is simply conditioning on the implementation variable, and thus introduces the biases described above.

G-estimation under a "no unmeasured $M - Y$ confounders" assumption. For both as-treated and per-protocol analyses, if all confounders of the compliance-outcome relationship are known, they can be adjusted for using standard methods like inverse probability weighting or g-estimation. These would target the quantity $\mathbb{E}[Y \mid D = 1, M = 1, X] - \mathbb{E}[Y \mid D = 0, M = 1, X]$, where X satisfies conditional ignorability. However, by conditioning on post-treatment variables, these approaches forego the identification benefits of the initial randomization, and require strong assumptions regarding the absence of mediator-outcome confounders (see e.g. Sheiner and Rubin, 1995, Hernán and Hernández-Díaz, 2012, Hernán et al., 2013).

A related approach is to model $M(1)$ for all units based on available (pre-treatment) covariates. The estimated $\widehat{M(1)}$ can be conditioned on (e.g. Vanacore et al., 2024). This characterizes the total effect across groups defined by their pre-treatment covariates. The predicted $\widehat{M(1)}$ is at best an imperfect proxy for $M(1)$, nevertheless, this approach may prove

valuable when it is less important to precisely learn the TRACE but desirable to have some directional estimate of how it changes as a function of relevant pre-treatment covariates.

Applicability to partial identification using AutoBounds. Duarte et al. [2024] describe a general framework for obtaining bounds on causal estimands by solving a polynomial optimization problem. When formulated as minimization and maximization problems, these programs yield lower and upper bounds on causal queries subject to assumptions and observed data. Although the TRACE is not among the estimands considered explicitly in Duarte et al. [2024], users could treat the TRACE as the target estimand and encode postulated values or constraints on TRACE(0). AutoBounds would then provide a convenient way to obtain bounds on the TRACE, and to combine the assumptions used here with any additional restrictions an investigator is willing to impose, revealing their joint identifying power.

3.3.6 Complementary monotonicity and trimming bounds

Where M is observed for all units, it may seem wasteful that our approach does not seek to learn anything from the naive comparison, $\mathbb{E}[Y \mid D = 1, M = 1] - \mathbb{E}[Y \mid D = 0, M = 1]$. However, we show here that the bias separating this quantity from the TRACE involves other unobservable quantities.[6] Consider the following breakdown of this naive comparison, in which "AT", "C", and "Def" refers to always-takers, compliers, and defiers respectively,

$$\begin{aligned}
&\mathbb{E}[Y \mid D = 1, M = 1] - \mathbb{E}[Y \mid D = 0, M = 1] \\
&\quad = \mathbb{E}[Y(1) \mid M(1) = 1] - \mathbb{E}[Y(0) \mid \text{Def+AT}] \\
&\quad = \mathbb{E}[Y(1) \mid M(1) = 1] - \mathbb{E}[Y(0) \mid M(1) = 1] + \mathbb{E}[Y(0) \mid M(1) = 1] - \mathbb{E}[Y(0) \mid \text{Def+AT}] \qquad (3.11) \\
&\quad = TRACE + \mathbb{E}[Y(0) \mid \text{AT+C}] - \mathbb{E}[Y(0) \mid \text{Def+AT}] \qquad (3.12)
\end{aligned}$$

[6] As discussed in B.1), we could choose to reason about these quantities instead of TRACE(0).

Even when we can rule out defiers, we are left with

$$\begin{aligned}\mathbb{E}[Y \mid D=1, M=1] - \mathbb{E}[Y \mid D=0, M=1] &= TRACE + \mathbb{E}[Y(0) \mid \text{AT+C}] - \mathbb{E}[Y(0) \mid \text{AT}] \\ &= TRACE + \left[\mathbb{E}[Y(0) \mid \text{AT}]\frac{Pr(\text{AT})}{Pr(\text{AT})+Pr(\text{C})} + \mathbb{E}[Y(0) \mid \text{C}]\frac{Pr(\text{C})}{Pr(\text{AT})+Pr(\text{C})}\right] \\ &\quad - \mathbb{E}[Y(0) \mid \text{AT}] \\ &= TRACE + \left[\mathbb{E}[Y(0) \mid \text{C}] - \mathbb{E}[Y(0) \mid \text{AT}]\right]\frac{Pr(\text{C})}{Pr(\text{AT})+Pr(\text{C})} \qquad (3.13)\end{aligned}$$

This is not fully identified simply because we cannot learn $\mathbb{E}[Y(0) \mid \text{C}]$ from the data.[7] However, we can bound this quantity using the observed distribution of outcomes among units with $D=0, M=0$ (compliers and never-takers).[8] To do this, we first need to determine what proportion of this group is made up of compliers. We can estimate $\widehat{Pr}(\text{NT}) = \widehat{Pr}(M = 0 \mid D=1)$, where "NT" refers to never-takers.

Then,

$$\widehat{Pr}(\text{C} \mid D=0, M=0) = \frac{\widehat{Pr}(\text{C})}{\widehat{Pr}(\text{C}) + \widehat{Pr}(\text{NT})} \qquad (3.14)$$

Let $\hat{\pi}$ denote this proportion. We can place simple bounds on the mean $Y(0)$ among compliers by taking the mean of the upper and lower $\hat{\pi}$ percentiles of Y in the group with $D=0$ and $M=0$. By randomization, $\mathbb{E}[Y(0) \mid \text{C}] = \mathbb{E}[Y(0) \mid \text{C}, D=0, M=0]$.[9] Substituting these bounds for $\mathbb{E}[Y(0) \mid \text{C}]$ in Expression 3.13 yields corresponding bounds on the TRACE. We call these the "monotonicity and trimming" (MT) bounds.

A similar logic has been used to bound the survivor average causal effect under monotonicity [Lee, 2005, Imai, 2008, Zhang and Rubin, 2003, Horowitz and Manski, 1995]

[7]Under monotonicity (no defiers), we can identify and estimate $Pr(\text{C})$ using $\widehat{Pr}(\text{C}) = \widehat{Pr}(M=1 \mid D=1) - \widehat{Pr}(M=1 \mid D=0)$ and $Pr(\text{AT})$ using $\widehat{Pr}(\text{AT}) = \widehat{Pr}(M=1 \mid D=0)$.

[8]We thank an anonymous reviewer for suggesting this approach.

[9]Let C denote membership in the complier stratum. By randomization of D, $\mathbb{E}[Y(0) \mid \text{C}] = \mathbb{E}[Y(0) \mid \text{C}, D=0]$. By Law of Iterated Expectations, $\mathbb{E}[Y(0) \mid \text{C}, D=0] = \mathbb{E}[Y(0) \mid \text{C}, D=0, M=0]Pr(M=0 \mid \text{C}, D=0) + \mathbb{E}[Y(0) \mid \text{C}, D=0, M=1]Pr(M=1 \mid \text{C}, D=0)$. By the definition of compliers, $Pr(M=0 \mid \text{C}, D=0, M=0) = 1$ and $Pr(M=1 \mid \text{C}, D=0, M=0) = 0$. Therefore, $\mathbb{E}[Y(0) \mid \text{C}] = E[Y(0) \mid \text{C}, D=0, M=0]$.

or "conditional monotonicity" [Samii et al., 2025, Semenova, 2020]. Where we can plausibly assume (conditional) monotonicity and M is measured for all units, this approach offers a separate route to bounding the TRACE, alongside bounding by reasoning about TRACE(0). This is especially useful when the two bounding regions overlap incompletely, as valid results would then need to fall in the intersection of the two.

3.4 Examples

3.4.1 Hypothetical demonstration: Perceived race, police stops, and violence

Here we provide a brief hypothetical example to demonstrate the use of this approach in the challenging and widely discussed example of understanding the influence of perceived race on police behavior (see e.g. Knox et al. [2020]).[10] Consider a scenario in which police officers observe drivers and have a perception (accurate or not) of each driver's race. Let $D = 1$ designate that the police perceived the driver to be from a minority group. The police may then choose to stop the vehicle (M). During that stop, police violence may or may not occur (Y). Note that if a stop does not occur (so $M = 0$), then there can be no violence, and thus no effect of perceived race on violence. This is a "necessary condition mediator" (Type 3) case.

Recall that TRACE(0) is the total effect among never-takers (with $\{M(1) = 0, M(0) = 0\}$), and defiers (with $\{M(1) = 0, M(0) = 1\}$). Suppose we can rule out defiers, arguing that if a person is not stopped when perceived to be a minority, they would not be stopped when perceived to be non-minority. Consequently, the $M(1) = 0$ group includes only never-takers. Because this group is not stopped (regardless of perceived race) there can be no police violence, and thus no effect of perceived race on violence. This makes $TRACE(0) = 0$, point

[10]This setting is not experimental, so in an actual empirical study of this question, unobserved confounding would pose additional problems beyond the identification concerns related to post-treatment conditioning discussed here and elsewhere.

identifying the TRACE.[11]

Estimation requires only the sample moments required to estimate (i) the total effect, and (ii) the proportion with $M(1) = 1$. The former requires knowing what fraction of encounters with perceived-minority drivers resulted in violence ($Pr(Y = 1 \mid D = 1)$), and the fraction of encounters with perceived-non-minority drivers resulting in violence ($Pr(Y = 1 \mid D = 0)$). This may not be directly available, but could be backed-out from alternative combinations of information, such as the fraction of cases with violence involving perceived minority drivers ($Pr(D = d \mid Y = 1)$), the overall proportion of police-perceived minority drivers ($Pr(D = 1)$) in the relevant population, and the overall proportion of encounters that result in violence ($Pr(Y = 1)$). However, we cannot estimate the second quantity, $Pr(M(1) = 1)$, in cases where data on encounters are not available. This is similar to conclusions reached by other means in Knox et al. [2020], who show that the *total* effect can be point identified in this setting given additional data on the total number of perceived minority and the total number of white encounters ("encounters" describing any possible stop/nonstop). Using the count of total encounters among minority-perceived drivers, we could also estimate $Pr(M(1) = 1)$ using $Pr(M = 1 \mid D = 1)$, identifying the TRACE.

3.4.2 Effects of community policing on mob violence in Liberia

Setting. Our first empirical application is an experimental study by Morse [2024] examining the effects of a community policing intervention in Monrovia, Liberia on mob violence, overall instances of crime, perceptions of security, and crime reporting.[12] The

[11]Even when we cannot rule out defiers, we may still be able to argue that TRACE(0) is nearly zero, which may inform inference. Consider cases in which a police officer would stop an individual only if they were perceived as non-minority, e.g. during a manhunt for a white suspect. Information about the nature of these stops and how often such manhunts occur could put reasonable limits on the ratio of never-takers to defiers. If the proportion of defiers is expected to be much smaller than the proportion of never-takers, $TRACE(0)$ will be very close to 0, even if the effect among defiers is not. Partial identification can then proceed with $TRACE(0)$ values near 0. We thank an anonymous reviewer for raising this consideration.

[12]This study was part of a coordinated multi-site trial across six countries [Blair et al., 2021].

intervention, implemented in collaboration with the Liberian National Police (LNP), involved community town hall meetings, increased foot patrols, and the formation and training of local security groups known as Community Watch Forums. Forty-five out of 93 eligible communities in the capital Monrovia were randomly selected to receive the program over a period of 10 months. Outcomes were measured primarily using community surveys administered three months after the end of the intervention. Morse [2024] finds that the intervention meaningfully reduced instances of mob violence, but did not affect overall crime, perceptions of security, or crime reporting.

The intervention was designed to reflect a coproduction-based approach to policing, aimed at simultaneously increasing reliance on the police and improving communities' abilities to respond to security issues in a lawful manner. This model is tailored to contexts with low existing police capacity, where promoting cooperation with police as the *sole* legitimate means of addressing community security issues can be ineffective, or even backfire, if police lack the capacity to effectively respond. To avoid these potential pitfalls, the intervention incorporated training and capacity building for Community Watch Forums (CWFs), groups of citizens who cooperate directly with the police, including by sharing information, facilitating police investigations, and conducting joint security patrols. [13]Morse [2024] attributes the success of the intervention at reducing mob violence to this component in particular.

We use (community awareness of) the presence of a CWF as our primary mediator of interest M, setting $M = 1$ if more than 20% of community survey respondents at endline reported the existence of a CWF. Our main outcome of interest Y is the average number of instances of mob violence reported by community survey respondents in the past year. The TRACE represents the effect of the community policing intervention on reported mob violence *in communities where assignment to the intervention would have resulted in the*

[13]Community security groups existed in some form in many areas prior to the intervention. As a result, it is possible for respondents in control communities to report the presence of a CWF.

formation of a successful CWF (hereafter "implementing types").

Results. Figure 3.2 illustrates what can be claimed about the effect among implementing types (TRACE) given any assumption about the effect among non-implementing types (TRACE(0)). The vertical axis represents postulated values of TRACE(0), while the horizontal axis represents corresponding estimates of the TRACE.

To illustrate interpretation, consider three possible assumptions. First, suppose we assume that the effect of the community policing intervention is the same among implementing and non-implementing types (i.e. that TRACE = TRACE(0)). Under this assumption, our point estimate of the TRACE is equivalent to that of the ITT, represented by the intersection of the dashed lines at an estimate of -0.23.[14] HC2-based confidence intervals for the ITT are shown using the (blue) vertical line segments around the point estimate. Note that confidence intervals for the TRACE are larger than for the ITT even where the two point estimates are equivalent. This is because the former represents an inference about a sub-group (those with $M(1) = 1$).

Second, if we are willing to posit that the intervention had no effect on mob violence in non-implementing type communities (i.e. that TRACE(0) = 0), we recover an effect size among implementing types of -0.57, more than twice the magnitude of the ITT, in the hypothesized direction (a reduction in mob violence).

Third and more practically, we would likely be unwilling to endorse either of the assumptions above in this case, especially using an arbitrary cutoff for the community awareness indicator of implementation (M). However, we are more willing to entertain the assumption that the intervention produced a larger benefit (a more negative effect on mob

[14]Our point estimate of the ITT (-0.23) differs slightly from the estimate of -0.28 in Morse [2024] for two reasons. First, while Morse [2024] analyzes outcome data at the individual level and clusters standard errors at the community, we measure the outcome at the community level, using the community-level mean, avoiding the clustering requirement. Second, while we use OLS, Morse [2024] uses weighted least squares, weighting by the product of inverse probability of community assignment to treatment and inverse probability of individual selection for the survey.

violence) among implementing types than among non-implementing types. Researchers may argue this directly based on the necessity of M to achieving a meaningful effect.[15] If true, this implies estimates for the TRACE that fall to the left of the vertical dotted line.[16] If we wish to further assume that TRACE and TRACE(0) have the same sign, we bound the TRACE on the left at the point where TRACE(0) = 0. This range is represented by the grey shaded region on the plot, and corresponds to point estimates of TRACE between -0.23 and -0.57 (Figure 3.2, shaded region).

The results also imply that any argument for a TRACE estimate larger (more beneficial) than -0.57 would require us to believe that the intervention produced an average *increase* in mob violence in non-implementing type communities. This could be possible, for example, if police lacked the capacity to respond to increased demand for law enforcement resulting from other aspects of the intervention and – in the absence of a viable lawful community-based alternative – communities (would have) resorted to vigilante justice. Investigators and subject experts can reason about the plausibility of such opposite-directional effects among non-implementing types in this or any given setting.

Assessing null findings. Our approach can also be informative in the interpretation of null results. For example, in contrast to the encouraging findings regarding the effects on mob violence, Morse [2024] does not find evidence that the community policing intervention reduced crime, improved perceptions of security, or increased crime reporting. Given uneven

[15]More rigorously, the assumption can be supported by reference to a combination of direct and indirect effects and their signs. To simplify, suppose we can rule out defiers, as is reasonable here. Then TRACE(0) is simply the effect among never-takers, which is a direct effect of D on Y in that group without the indirect effect through M. We might expect this to be small if D is an intervention that cannot reasonably have a large (direct) effect in the absence of M. Meanwhile, the TRACE is due to both always-takers and compliers. The always-takers again contribute only a direct effect, which may arguably be small. However, compliers can provide both a direct effect and the key indirect effect of interest, which might be much larger when M is clearly important to achieve an effect.

[16]TRACE estimates to the right of the vertical dotted line imply a more beneficial (negative) effect of the community policing intervention on mob violence among non-implementing types than among implementing types.

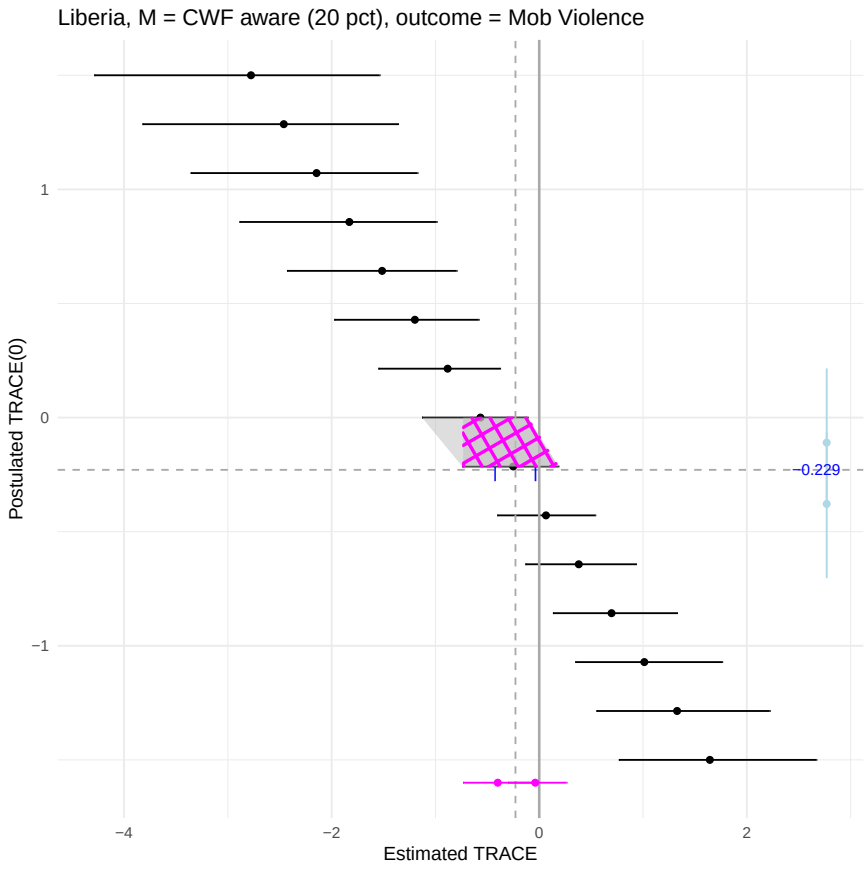


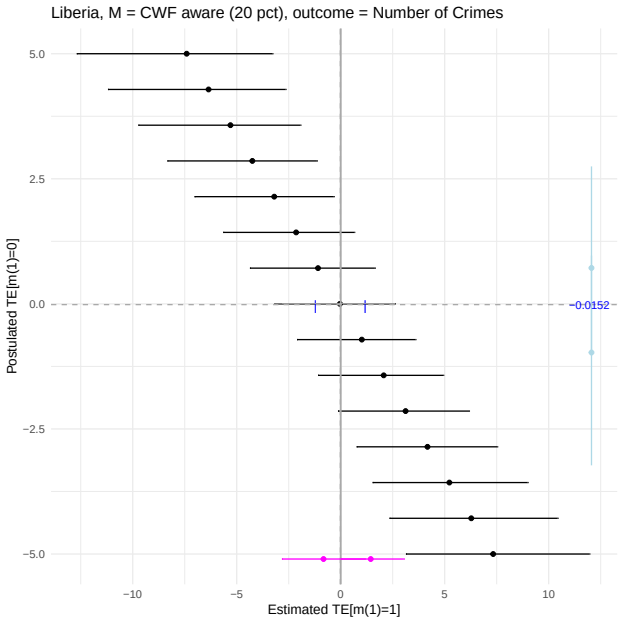
Figure 3.2: Estimated TRACE of community policing intervention on reported mob violence, among communities for which the intervention would have produced a successful Community Watch Forum (implementing types), across postulated values of the treatment effect among non-implementing types (TRACE(0)). Line segments represent 95% confidence intervals. The number on the far right represents the "naive" ITT estimate, assuming constant treatment effects across strata of M. The grey shaded region represents the range of estimates assuming a value for TRACE(0) less than or equal to the value of the TRACE, in the same direction. The pink bounds represent the MT bounds described in Section 3.3.6, and the pink crosshatch shows the intersection of the MT bounds with the assumptions on TRACE(0). All models include police zone (block) fixed effects and baseline levels of reported mob violence, averaged at the community-level, as a covariate.

implementation, we might wonder "what would we have to believe about the effects of the intervention among non-implementing types to recover an effect in hypothesized direction among implementing types?" Figure 3.3 answers this question when using crime incidence as the outcome.[17] In panel 3.3a, we again use CWF-awareness for M ($M = 1$ if 20% of community survey respondents reported awareness of a CWF registered with police at endline). In order to obtain a significant beneficial (negative) effect among implementing types, we would have to assume a harmful effect among non-implementing types larger in magnitude than the corresponding beneficial effect among implementing types. We expect this is unlikely.

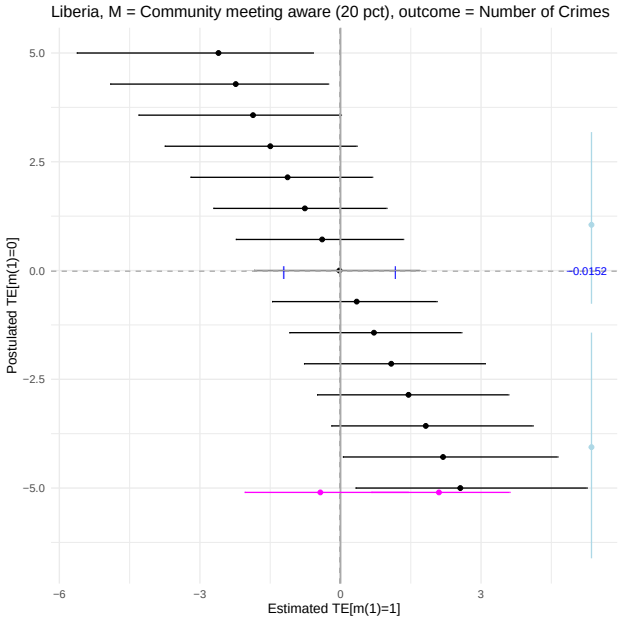
Panel 3.3b repeats this analysis replacing M with an alternative implementation measure capturing another major component of the intervention: community town hall meetings. Here, $M = 1$ if more than 20% of community survey respondents reported being aware of or having attended a community security meeting with police at endline. The results show that we cannot obtain a significant beneficial effect among implementing types (TRACE) even if the non-implementing group had a harmful effect (TRACE(0)) almost twice as large as the beneficial effect among implementers. This increases our confidence that the failure to find an effect of the intervention on crime is not a result of uneven implementation across treated communities, at least insofar as we can consider the implementation "good enough" in those with $M = 1$.

Additional monotonicity and trimming bounds. Results using the alternative MT bounds described in Section 3.3.6 are also shown in Figures 3.2, 3.3a, and 3.3b. This approach relies on an additional "no defiers" assumption, meaning here that no communities would have formed a police-registered CWF only if not assigned to the intervention. For the effect of the intervention on mob violence (Figure 3.2), the upper MT bound on the TRACE

[17]For a list of crimes, survey respondents were asked whether they or anyone in their family were a victim of that type of crime in the past 6 months. The individual-level variable captures the total number of crimes reported across all categories. In our analysis, this variable is averaged at the community level.



(a) Community Watch Forum (CWF) awareness as the mediator



(b) Community meeting attendance or awareness as the mediator

Figure 3.3: Estimated TRACE of community policing intervention on reported incidents of crime among implementing type-communities, across postulated values of the treatment effect among non-implementing-type communities(TRACE(0)), using two different implementation measures (M).

is -0.04, which corresponds to assuming TRACE(0) $= -0.11$. The lower MT bound is -0.40, which corresponds to assuming TRACE(0) $= -0.38$. In this case, combining the MT bounds with the originally proposed bounds results in tighter bounds than either approach alone. Meanwhile, in Figures 3.3a and 3.3b, we see that the MT bounds remain wide, while reasonably defensible assumptions on TRACE(0) lead to more informative conclusions.

3.4.3 Empirical application 2: Effects of in-person canvassing on support for transgender rights

Setting. Our second empirical application is a field experiment by Broockman and Kalla [2016], examining the effects of a door-to-door perspective taking intervention on attitudes toward transgender people and support for a transgender non-discrimination law. The intervention involved a ten-minute conversation between canvassers and voters in Miami, Florida, during which canvassers employed strategies meant to encourage "active processing," asking voters to talk about a time when they were judged for being different, and encouraging them to see connections between their own experiences and those of transgender people. Registered Miami voters who completed a baseline survey (n = 1,825) were randomly assigned to receive either the perspective taking intervention or a placebo intervention involving a conversation about recycling. Among the 501 voters who answered the door in either condition, outcomes were measured in follow-up online surveys administered three days, six weeks, and three months after the intervention.

Broockman and Kalla [2016] find evidence that the intervention both reduced prejudice against transgender people and increased support for policies benefitting them: six weeks and at three months post-intervention, treated individuals were significantly more tolerant of transgender people and more supportive of an ordinance protecting them from discrimination in housing employment and public accommodations.[18]

[18]The authors did not find positive effects on support for the non-discrimination law during the first and second waves of outcome measurement. Beginning in the third wave (six weeks post-intervention), they

Investigators may then wish to know what the effect of the intervention on policy attitudes would be if they could look just at individuals *for whom the intervention (would have, if treated) produced a positive effect on attitudes toward transgender people.* For this application, we term this subset "reactive types," noting that this shorthand is imperfect. We consider this an example of a "mechanism of interest" (type 2) question described above. Our primary outcome Y is support for the transgender non-discrimination law six weeks after treatment (wave 3), measured on a seven-point scale. Our mediator (M) captures the *change* in subjective feelings toward transgender people from baseline. The authors measure attitudes toward transgender people during all waves using a standard 0-100 feeling thermometer, where a higher number represents "warmer" feelings. We code $M = 1$ if an individual's thermometer score increased between baseline and wave 3, and $M = 0$ otherwise. In the treatment group, thermometer scores increased between baseline and wave 3 for approximately 43% of respondents.

Results. Figure 3.4 shows the results from this analysis. We illustrate interpretation with the same three assumptions employed in the previous application. First, if we assume the intervention was equally effective in increasing support for the non-discrimination law among individuals for whom it (would have) produced an improvement in subjective feelings towards transgender people and those for whom it would not have, our estimate of the TRACE is equal to the ITT (in this case, 0.27).[19] Second, if we instead assume no effect of the intervention on policy support among non-reactive types, we recover an effect among reactive types of 0.63. Finally, if we merely wish to assume that the effect on policy support among reactive types is *greater* than that among non-reactive types (but that the latter is still positively-signed), we obtain a range of estimates for the TRACE between 0.27 and 0.63.

---

revised the survey to define the term "transgender." Positive effects on support are detected only after this point. We therefore focus our re-analysis on the third survey wave.

[19]The original paper reports the Causal Average Complier Effect (CACE) rather than the ITT, adjusting for limited non-compliance with treatment assignment. While the CACE estimate reported in the paper (0.36) is larger than the ITT, the ITT estimate is still statistically significant.

Any estimate greater than 0.63 would require an assumption of a negative (e.g. "backlash") effect among non-reactive types.[20]

3.5 Conclusions

Investigators frequently face questions that require reasoning about treatment effects within subgroups defined by post-treatment characteristics, such as implementation quality, compliance, and attentiveness. Yet, directly conditioning on post-treatment variables introduces problematic biases (e.g. Montgomery et al., 2018). The TRACE provides a principled way to partially identify treatment effects for the subgroup that, if treated, would have taken on a particular value of a relevant post-treatment variable. In comparison to existing approaches to utilizing post-treatment variables, ours does not require strong and untestable assumptions such as the exclusion restriction or the absence of mediator-outcome confounders. Additionally, it is possible to use in cases where the relevant post-treatment variable is or can only be measured in the treatment group.

The primary limitation of this approach is that it offers only partial identification, excepting the special case where we can argue for no effect in the "non-reactive" group (TRACE(0) = 0). We emphasize that it is neither necessary nor desirable for users to postulate and defend a single value of TRACE(0), in order to produce a point estimate of the TRACE. Instead, the results show what TRACE values are possible given beliefs we can defend about TRACE(0). For example, in some cases investigators may be able to argue that the TRACE(0) is near zero, exactly zero, has a smaller absolute value but same sign as, or has the opposite sign as that of the TRACE. A defensible assumption of this kind may sometimes lead to an informative boundary; in other cases it may not.

[20]We do not attempt the alternative MT bounds here because we find it unlikely that monotonicity holds. Specifically, some individuals may become irritated with the door-to-door canvassing, developing a more negative attitude towards transgender people in response, while their attitudes may have remained level or improved slightly over-time otherwise.

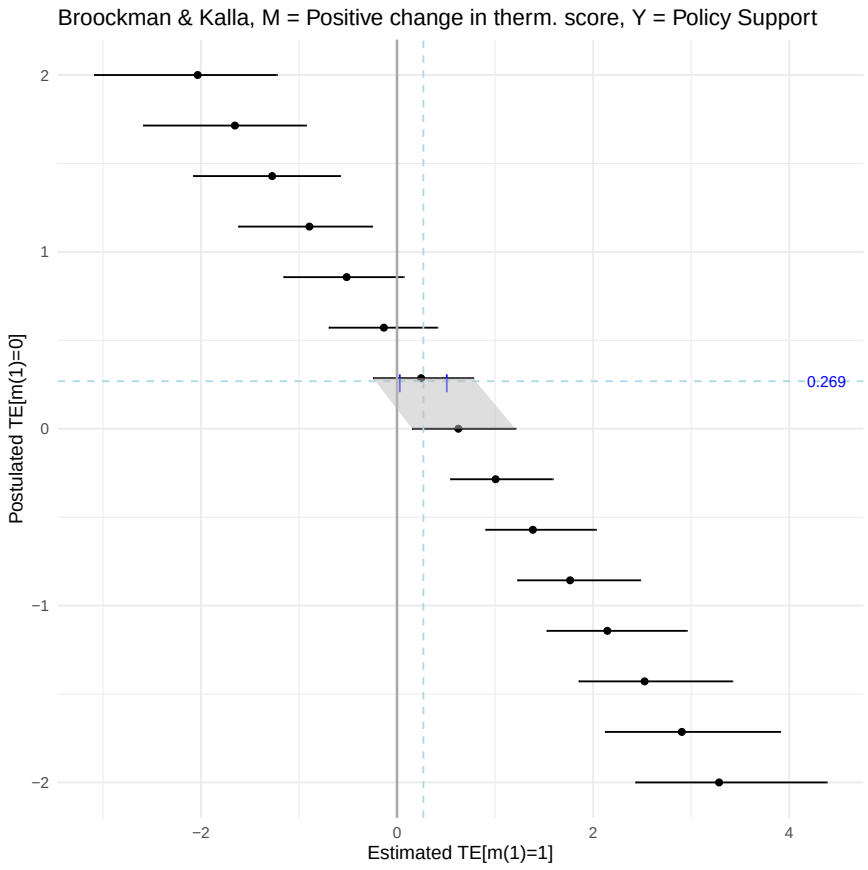


Figure 3.4: Estimated TRACE of in-person canvassing intervention on support for transgender non-discrimination law, among those for which the intervention would have led to an improvement in feelings toward transgender people from baseline (reactive types), across postulated values of the treatment effect among non-reactive types (TRACE(0)). Line segments represent 95% confidence intervals. The number on the far right represents the "naive" ITT estimate, assuming constant treatment effects across strata of M. The shaded region represents the range of estimates assuming a value for TRACE(0) less than or equal to the value of the TRACE, in the same direction. All models include control variables used in Broockman and Kalla [2016], including baseline levels of both transgender attitudes and policy support.

One area of particular interest may be the interpretation of null results from experimental studies in which implementation was uneven. For example, finding no average effects of community policing on citizen-police trust, cooperation with police, or crime across multiple countries, Blair et al. [2021] reasonably suggest that "implementation challenges common to police reforms may have contributed to these disappointing results." In such cases, the approach we propose enables researchers to precisely characterize what we would have to assume in order to attribute null effects to implementation challenges.

CHAPTER 4

Estimating effects of cancer screening among those who would be diagnosed if screened: an application of the Treatment Reactive Average Causal Effect

4.1 Introduction

The effects of early cancer screening on severity of illness and cancer-related mortality are of interest to researchers, physicians, and policymakers alike. In 1999, the Korean government funded the implementation of the Korean National Cancer Screening Program, which provided free biennial Pap smear tests for the screening of cervical cancer. Bui et al. [2021] investigated the effects of early screening on development of invasive cervical cancer and carcinoma in situ (non- or pre- invasive, an earlier stage of cancer) by comparing women who participated in KNCSP to those who did not. They found that women in the ever-screened group were more likely to be diagnosed with carcinoma in situ than with invasive cervical cancer. Similar studies in other countries have found that early screening is associated with reductions in risk of more severe diagnoses of cancer [Amri et al., 2013, Bretthauer et al., 2022].

When assessing a randomized controlled trial on effects of cancer screening on severity and mortality outcomes, it is standard to compare outcomes in the treated group to the control group for all eligible units in the experimental sample to estimate an effect [Bui et al., 2021, Amri et al., 2013] or on incidence of cancer [Bretthauer et al., 2022]. However,

in thinking about the effect of screening on severity of diagnosis, we would ideally like to compare outcomes for only the subset of the population for whom the screening would lead to diagnosis – for those who are never diagnosed with cancer, we expect that screening will have a negligible effect. Focusing on this group would help us to isolate the effect of screening on outcomes of interest, without masking or overstating the effect.

A natural inclination, then, may be to condition on diagnosis in the analysis, i.e. subset the analysis to those who were diagnosed with some stage of cancer. However, this type of conditioning introduces statistical biases that render an invalid causal estimate. Instead, we propose conditioning on "potential" diagnosis, that is, the diagnosis status that an individual would have if they were screened. We describe a partial identification approach that targets the Treatment Reactive Average Causal Effect (TRACE), defined as the effect of treatment among those who, if treated, would have a particular value of the post treatment variable of interest [Hazlett et al., 2025]. Given the proportion of people who screen positive among those screened and the total effect of screening on severity, and by reasoning about the effect of screening for those who, if screened, would not be diagnosed, we can estimate a range of plausible values for the TRACE. There may be reason to believe that the effect of screening for the "screen negative" group is equal to or close to 0. In this case, we can achieve point identification of the TRACE.

In the sections that follow, we detail the use of this approach in estimating the effects of screening on mortality outcomes, both in comparing different types of screening and in comparing screening to no screening. By reasoning about the effect among the group that would not be diagnosed if screened, we can identify and estimate the range of plausible values for the TRACE. We compare our proposed estimand with related estimands in the literature, notably the survivor average causal effect (SACE) and the local average treatment effect (LATE) targeted by instrumental variables. We demonstrate the use of the proposed approach through (i) a simulated example estimating the effects of early screening for cancer on life span and (ii) an applied example estimating the effects of low dose CT screening on

lung cancer mortality, compared to chest radiography screening.

4.2 Background

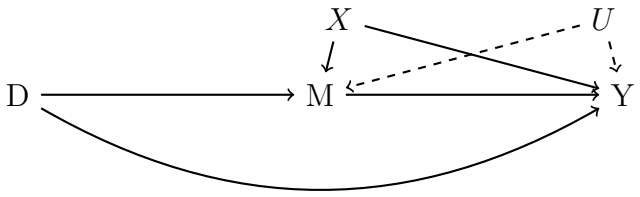


Figure 4.1: Setting of interest

Consider the setting in Figure 4.1. Researchers may be interested in the effect among a group with a particular value of M, but cannot condition on M as it would induce collider bias and/or selection bias. This is especially relevant when estimating the effects of cancer screening or in comparing effects of different types of screening, as we expect that the effect will the strongest among those for whom the screening detects cancer or a precancerous nodule.

There are many related settings in which disease progression analyses are useful, and corresponding estimands have been proposed for analyses of effects in these settings [Gonçalves and Suzuki, 2025]. One proposed estimand is the survivor average causal effect, or the effect among the "always-infected" group. Several identification strategies have been developed for the survivor average causal effect [Tchetgen, 2014, Egleston et al., 2009, Imai, 2008, Hayden et al., 2005, Zhang and Rubin, 2003, Rubin, 2000, Robins, 1986]. However, for our setting, the survivor average causal effect does not fully capture the effect of the exposure. Consider the case where we compare screening to no screening – in this case, it is impossible to screen positive if one is not assigned screening. There is no "always-infected" group, but we still are interested in the effect for those who would screen positive if screened. For this type of setting, Gonçalves and Suzuki [2025] propose the controlled direct effect

(CDE) as a possible quantity of interest. The CDE estimates the effect of screening on outcome for all units when we set diagnosis to 1, which could be useful if correctly identified [Robins and Greenland, 1992, Pearl, 2001, Imai et al., 2010a, Acharya et al., 2016]. However, the identification approach relies on assumption of no unobserved M-Y confounders, which we do not wish to make.

Existing analyses of cancer screening trials mostly focus on intention-to-treat analysis and per-protocol analysis [Bretthauer et al., 2022]. Intention-to-treat analysis is known to be a conservative estimate of effect, and in the case of cancer screening where so few people will be diagnosed with cancer, the ITT estimate is likely to mask any true effects of the exposure [Hernán and Hernández-Díaz, 2012]. Per-protocol analysis focuses on just those people who were screened and did screen positive, this essentially conditions on diagnosis and relies on accurate adjustment for all M-Y confounders. Similar to the CDE, identification of the per-protocol effect is possible only if the corresponding conditional ignorability assumptions hold [Hernán and Robins, 2017, Hernán et al., 2013]. We instead propose a partial identification approach that does not require these assumptions.

Partial identification and sensitivity analysis approaches targeting different estimands have been proposed for the cancer screening setting. Swanson et al. [2015] target the per-protocol effect in estimating the effect of colorectal cancer screening on colorectal cancer incidence and mortality over 10 years of follow-up using data from the Norwegian Colorectal Cancer Prevention trial. Key to this setting is that not all people assigned screening actually did the screening, thus the need for the per-protocol effect, which focuses on the group that was assigned to and did do the screening. In order to estimate bounds without any further assumptions on the setting, they impute the most extreme scenarios for unobserved counterfactual probabilities to get lower and upper bounds. In a similar vein, partial identification strategies have been developed for the survivor average causal effect that use a trimming approach that imputes worst-case values for the expected outcome under treatment for the always-survivor group [Samii et al., 2025, Semenova, 2020, Lee, 2005, Imai, 2008,

Zhang and Rubin, 2003, Horowitz and Manski, 1995]. These trimming approaches rely on an additional assumption of monotonicity (no defiers).

Our approach targets partial identification of a novel estimand, defined as the total effect among the group that would take on a particular value of the post-treatment variable if treated. In the framework of principal stratification [Zhang and Rubin, 2003], this is the effect among always-takers and compliers. This differs from the existing literature in both the quantity of interest that we are targeting, and in the partial identification opportunity that it allows.

4.3 Proposed approach

4.3.1 Setup and notation

Let D be the randomized assignment to screening, M be the detection of cancer from that screening, and Y be an outcome of interest that is defined for all units. We denote the potential outcomes for unit i as $Y_i(d, m)$ for $d = 0, 1$ and $m = 0, 1$ and potential mediator values as $M_i(d)$ for $d = 0, 1$. The total effect is defined as

$$TE = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(0, M_i(0))] \tag{4.1}$$

For our purposes, because we are not interested in manipulating M, we use shorthand $Y(1, M(1)) = Y(1)$ and $Y(0, M_i(0)) = Y_i(0)$. We can then write the total effect as follows:

$$TE = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{4.2}$$

The estimand of interest is the Treatment Reactive Average Causal Effect, defined as the effect of assigned treatment among the group that "if screened, would have been diagnosed",

$$TRACE = \mathbb{E}[Y_i(1) - Y_i(0)|M_i(1) = 1] \tag{4.3}$$

In order to partially identify this quantity, we need to reason about the treatment effect among that group that "if screening, would not be diagnosed,"

$$TRACE(0) = \mathbb{E}[Y_i(1) - Y_i(0)|M_i(1) = 0] \tag{4.4}$$

4.3.2 Partial identification of TRACE

Given randomization of D, the total effect (TE) of D on Y is identifiable, as is $Pr(M(1) = m)$ for $d = 0, 1$. Using these quantities, we can partially identify the TRACE:

$$TE = Pr(M(1) = 1)TRACE + Pr(M(1) = 0)TRACE(0) \tag{4.5}$$

$$TRACE = \frac{TE - Pr(M(1) = 0)TRACE(0)}{Pr(M(1) = 1)} \tag{4.6}$$

We can estimate TE as $\widehat{TE} = \overline{Y_1} - \overline{Y_0}$ and $Pr(M(1) = m)$ as $\widehat{Pr}(M(1) = m) = \widehat{Pr}(M = m|D = 1)$. We are left to reason about the total effect among the $M(1) = 0$ or "screened but not diagnosed" group, $TRACE(0)$. Given a range of plausible values for $TRACE(0)$, we can estimate bounds on $TRACE$. Note that information on M is only required when $D = 1$. This allows us to use this approach in settings where M is not measured or is ill defined for units that do not receive treatment. For example, in comparing screening to no screening, if M is defined as the result from screening, the group that is not screened will not have a measurement for M. Regardless, we can reason about $TRACE(0)$, and estimate all other quantities necessary for partial identification of $TRACE$.

4.3.3 Reasoning about TRACE(0)

In a general setting, there are many different assumptions we can make on $TRACE(0)$ – notably, we can often plausibly assume that the effect of treatment among the "non-implementer" group (or in this case the "not diagnosed" group) will be smaller in magnitude than the effect among the implementer, or diagnosed, group: $|TRACE(0)| < |TRACE|$. For example, and previewing one of our applications, suppose we are interested in the effect

of a new, more sensitive type of screening in comparison to an older screening method. We would expect that the effect of the new screening on mortality among those who would not be diagnosed by the new screening would be smaller in magnitude than the effect for those who would be diagnosed.

We also may be able to assume that $TRACE(0)$ will be opposite in sign from the $TRACE$. Consider the situation where a government plans to provide a program for the community, but fails to properly implement the program. The lack of implementation could engender mistrust or anger among the community where it was not implemented. This would result in a negative effect of the program on trust in the government among communities where it was assigned but not implemented, but a positive effect among those were it was assigned and well implemented.

In some special settings, we can plausibly make the assumption that $TRACE(0) = 0$. We can see an example of this type of setting when estimating the effect of screening compared to no screening. We can plausibly assume that the effect of screening among those who would not be diagnosed if screened is equal to or close to zero. This could be violated if receiving a negative screening result induces people to change their behaviors, e.g. pursue a less healthy lifestyle. In our application in Section 4.4.1, we argue that these kinds of "backlash" effects are minimal if not null. When $TRACE(0) = 0$, our approach allows for point identification,

$$\widehat{TRACE} = \frac{\widehat{TE}}{\widehat{Pr}(M = 1|D = 1)} \tag{4.7}$$

Comparison to exclusion restriction Note that the identification result from postulating that $TRACE(0) = 0$ is similar to the identification result achieved by instrumental variables (IV) approaches that make the exclusion restriction assumption [Imbens and Angrist, 1994, Baker and Lindeman, 1994]. The two results will be equivalent when we can rule out always-takers, that is, $\widehat{Pr}(M = 1|D = 1) = \widehat{Pr}(\text{complier})$. However, stating that $TRACE(0) = 0$ is not the same as the exclusion restriction. The exclusion restriction states that D can only affect Y through M. Meanwhile, $TRACE(0) = 0$ allows

D to affect Y through M when $M(1) = 1$, but not when $M(1) = 0$.

Breaking this down further, in potential outcomes notation, the exclusion restriction is written as:

$$Y(m, d) = Y(m, d') = Y(m) \tag{4.8}$$

for all m and d, d' Imbens and Angrist [1994]. In the setting where we compare screening to no screening, where it is impossible for $M = 1$ when $D = 0$, this is equivalent to stating:

$$Y(\text{complier}|D = 1) = Y(M = 1) \tag{4.9}$$

and

$$Y(\text{never-taker}|D = 1) = Y(\text{never-taker or complier}|D = 0) = Y(M = 0) \tag{4.10}$$

Equation 4.9 holds as there are no always-takers or defiers, so the only group with $M = 1$ is the compliers. Equation 4.10 does not hold. In this setting, never-takers are those who would not be diagnosed with cancer if screened, while compliers are those who would be diagnosed with cancer if screened. The expected mortality outcome will be very different for those who would be diagnosed if screened compared to those who would never be diagnosed. The exclusion restriction is violated in this case, but we are still able to use our identification approach. This allows us to recover an effect when the IV approach cannot be used, and also suggests that the IV assumptions may be overly restrictive.

Practical Considerations Reasoning about $TRACE(0)$ in the screening settings is complicated by the fact that the results from screening are not always accurate. Figure 4.2 shows the different latent types that are represented by the groups $M(1) = 0$ and $M(1) = 1$. Notably, these groups can include false negatives and false positives, which may result in opposite effects than what we would expect for individuals with incorrect screening results.

For example, consider the individuals in the $M(1) = 0$ group who receive a false negative screening. These individuals will continue living with undetected cancer, and it is possible

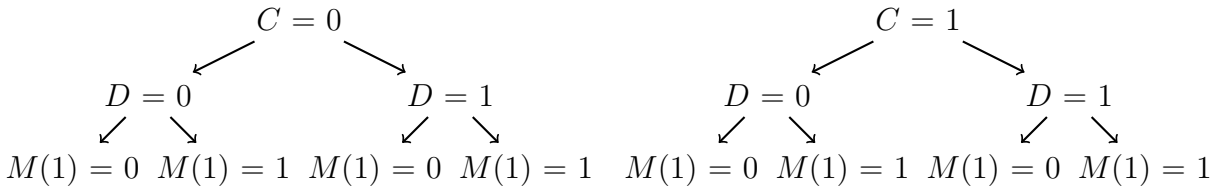


Figure 4.2: Trees representing 8 possible latent groups in the data. In practice, we only observe 4 groups as C (whether a person has cancer) is unknown prior to the screening

that receiving a negative screening result makes them less likely to do another screening in subsequent years. For these individuals, doing the screening and receiving a negative result may be worse than an alternative of not doing the screening. Meanwhile, for individuals who receive false positives, it is possible that they undergo unnecessary treatment and this has negative effects on life span and other relevant outcomes.

If the false positive and false negative rates are relatively small, it is possible to still make reasonable assumptions about $TRACE(0)$. We can break down the $TRACE(0)$ as a weighted combination of the effect among those who actually have cancer and the effect among those who do not:

$$\begin{aligned} TRACE(0) =& \mathbb{E}[Y_i(1) - Y_i(0) | M_i(1) = 0, C_i = 0] P(C_i = 0 | M_i(1) = 0) \\ &+ \mathbb{E}[Y_i(1) - Y_i(0) | M_i(1) = 0, C_i = 1] P(C_i = 1 | M_i(1) = 0) \end{aligned} \tag{4.11}$$

If $P(C_i = 1|M_i(1) = 0)$ is very small in comparison to $P(C_i = 0|M_i(1) = 0)$, then $TRACE(0)$ will be close to the expected effect among the "true negatives". Knowledge about the false negative rate of the screening in question can then help us in reasoning about $TRACE(0)$.

4.3.4 Sharpening bounds under additional assumptions

Under additional assumptions, we may be able to sharpen our bounds using a "trimming" approach [Lee, 2005, Imai, 2008, Zhang and Rubin, 2003, Horowitz and Manski, 1995].

Specifically, consider a monotonicity (no defier) assumption. In this case the first term in the TRACE—the expected treatment potential outcome for always-takers and compliers—is readily identifiable from the data,

$$\mathbb{E}[Y(1) \mid \text{always-takers, compliers}] = \mathbb{E}[Y|\text{always-takers, compliers}, D = 1] \tag{4.12}$$

$$= \mathbb{E}[Y|M = 1, D = 1] \tag{4.13}$$

The difficulty lies in the contrasting term, the expected non-treatment outcome among always-takers and compliers, or

$$\mathbb{E}[Y(0) \mid \text{always-taker}]Pr(\text{always-taker}) + \mathbb{E}[Y(0) \mid \text{complier}]Pr(\text{complier}) \tag{4.14}$$

$$= \mathbb{E}[Y \mid D = 0, M = 1]Pr(M = 1 \mid D = 0) + \mathbb{E}[Y \mid \text{complier}, D = 0]Pr(\text{complier}) \tag{4.15}$$

Under monotonicity we know that those with $D = 0$, $M = 1$ are always-takers and so can identify and estimate $\mathbb{E}[Y(0) \mid \text{always-taker}]$. However, we cannot directly estimate $\mathbb{E}[Y(0) \mid \text{complier}]$, because we cannot know if those with $D = 0, M = 0$ are compliers or never-takers.

We can use the observed distribution of outcomes among the $D = 0, M = 0$ (compliers and never-takers) group to estimate bounds on the expected outcome for the compliers as follows. First, we need to determine what proportion of this group is made up of compliers. This is identifiable as follows: We can estimate $\widehat{Pr}(\text{always-taker}) = \widehat{Pr}(M = 1 \mid D = 0)$, $\widehat{Pr}(\text{never-taker}) = \widehat{Pr}(M = 0 \mid D = 1)$, and $\widehat{Pr}(\text{complier}) = 1 - \widehat{Pr}(\text{always-taker}) - \widehat{Pr}(\text{never-taker})$. Then,

$$\widehat{Pr}(\text{complier} \mid D = 0, M = 0) = \frac{\widehat{Pr}(\text{complier})}{\widehat{Pr}(\text{complier}) + \widehat{Pr}(\text{never-taker})} \tag{4.16}$$

Knowing that $\hat{\pi} \equiv \widehat{Pr}(\text{complier} \mid D = 0, M = 0)$ of this group are compliers, we can place simple bounds on the mean $Y(0)$ of compliers by taking the mean of the upper and lower $\hat{\pi}$ percentiles of Y in the group with $D = 0$ and $M = 0$. We can use these bounds directly in Expression 4.15 to obtain bounded on the mean $Y(0)$ among always

takers and compliers. Differencing these from the estimate of $Y(1)$ among the compliers and always-takers ($\hat{\mathbb{E}}[Y|D=1, M=1]$) (Expression 4.13) produces bounds for the final TRACE estimate.

When these bounds can be calculated (when monotonicity holds and M is measured for all units), they provide a separate approach to partial identification that, in some cases, could lead to narrower bounds. In other cases, these bounds will be wider than those that we can estimate by reasoning about $TRACE(0)$. Nevertheless, the intersection of our bounds with the trimming bounds could be informative, especially when the bounding regions overlap.

4.4 Applications

4.4.1 Hypothetical example: effect of early screening for cancer on severe cancer incidence

Suppose we are interested in the effect of cancer screening (compared to no screening) on life span, among a group of high-risk individuals. Let D represent screening, where $Pr(D = 1) = 0.5$. Let C be a latent variable representing whether a person actually has cancer at screening time, with $Pr(C = 1) = 0.1$. Let M represent the detection of cancer at the screening time point, where $Pr(M = 1|D = 1) = 0.6$. Suppose the screening test has 85% probability of detecting cancer or a pre-cancerous nodule. Then, for the people with cancer, $Pr(M(1) = 1) = 0.85$. Finally, let Y represent life span, and assume average life span with no screening is 70 years and with screening, if the screening detects cancer, is 80 years. We simulate a data set of size $n = 100,000$ for this setting using the following data generating

process:

$$D \sim Bernoulli(0.5) \tag{4.17}$$

$$C \sim Bernoulli(0.1) \tag{4.18}$$

$$M(0) = 0 \tag{4.19}$$

$$M(1)|C = 0 = 0 \tag{4.20}$$

$$M(1)|C = 1 \sim Bernoulli(0.85) \tag{4.21}$$

$$Y(0) \sim Lognormal(4.2, 0.15^2) \tag{4.22}$$

$$Y(1) = Y(0) + 12^* C^* M(1) \tag{4.23}$$

As reflected in Equation 4.23, the ground truth for this data is that $TRACE = 12$ and $TRACE(0) = 0$. Figure 4.3 shows the estimated TRACE (x-axis) for a range of possible values of $TRACE(0)$ (y-axis). The estimated intention-to-treat effect, or the total effect of D, is 0.924 (this is shown by the intersection of the dotted blue lines).

In this setting, without a priori knowledge, we could reason that $TRACE(0)$ is close to 0 as the screening is unlikely to change anything for the people who would not be diagnosed from the screening. If we assume that $TRACE(0) = 0$, we see that our approach estimates that the $TRACE$ is close to 12. This is very different from the total effect of 0.924, and highlights how the effect among those diagnosed can differ greatly from the effect among the whole sample.

If we do not believe that $TRACE(0) = 0$, and instead believe that $|TRACE(0)| < |TRACE|$, then we estimate that the $TRACE$ is between 0.924 and 12. This provides us a range of plausible estimates for $TRACE$, while also taking into account our knowledge of the relationship between $TRACE$ and $TRACE(0)$.

Monotonicity and trimming bounds We can also calculate the monotonicity and trimming (MT) bounds described in Section 4.3.4 for this simulated example. Figure 4.4

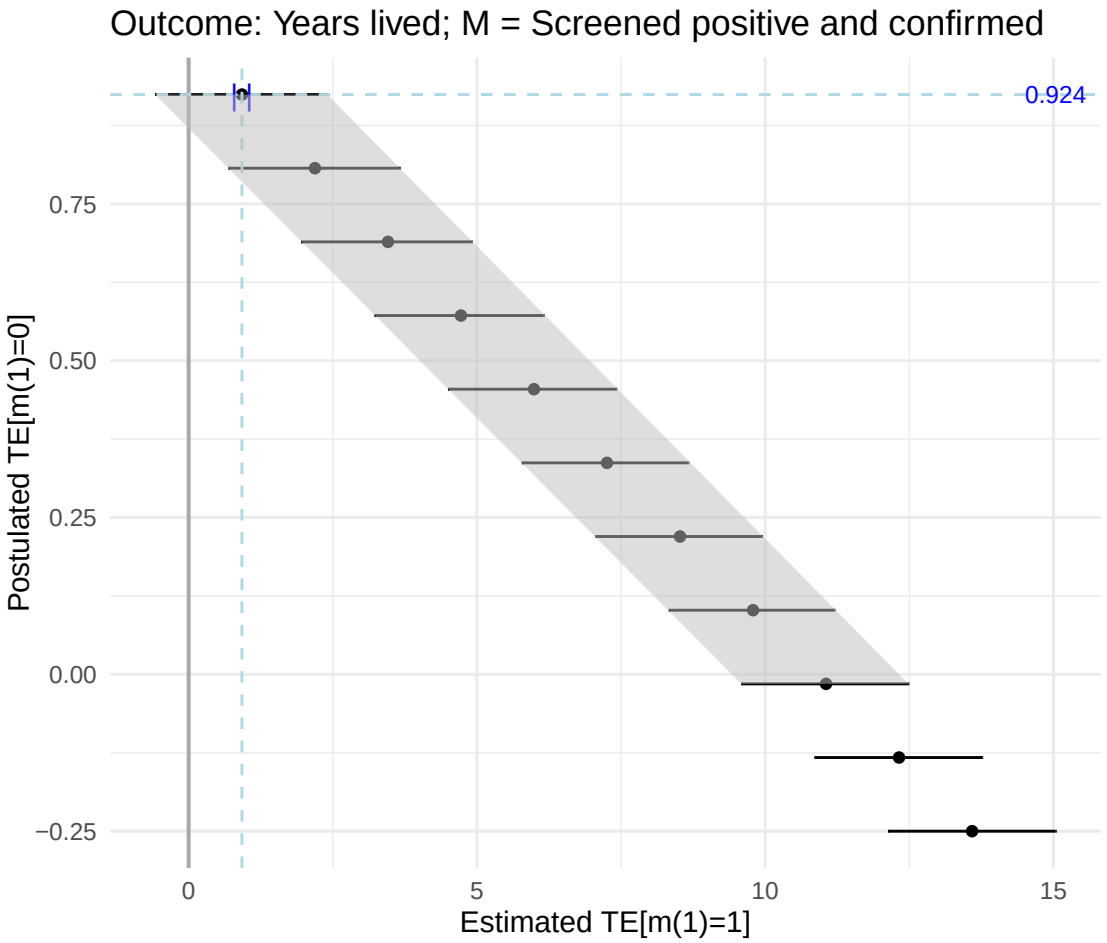


Figure 4.3: Results with simulated data.

shows the results from the MT approach, in addition to our original results. We see that in this particular setting, our bounds from reasoning about $TRACE(0)$ are strictly better than those obtained from the MT approach. This is because the probability of cancer diagnosis is very low, and so the proportion of compliers in the distribution of compliers and never-takers is small. This means that when we look at the lower and upper percentiles corresponding to this proportion, we will likely get values that are on both extremes of our overall outcome distribution, leading to wide bounds. This is not true of most settings – although the MT bounds are uninformative here, they are useful in other settings (see e.g. Hazlett et al., 2025).

Figure 4.4: Results for simulated data, including Monotonicity and Trimming (MT) bounds. The pink data points show the lower and upper bounds (and confidence intervals) on TRACE from the MT approach. The light blue data points show the corresponding TRACE(0) values for those lower and upper bounds on TRACE. The pink crosshatch shows the overlap between our bounds and the MT bounds.

4.4.2 Empirical application: Effects of low-dose CT screening compared to chest radiography on lung cancer mortality

Our empirical application involves estimating the effects on lung-cancer related mortality of low-dose helical computed tomography (CT) for lung-cancer screening. When low-dose CT was introduced for screening, many studies showed that the method detects tumors at earlier stages compared to other screening methods. The National Lung Screening Trial (NLST) aimed to determine whether low-dose CT screening could reduce lung-cancer related mortality among high-risk persons when compared to the previous standard screening method, single-view posteroanterior chest radiography [NLST Research Team, 2011].[1]

Setting. The study was conducted over 33 sites in the United States from August 2002 to April 2004, with 53,452 participants who were each randomly assigned to either low-dose CT or single-view posteroanterior chest radiography. Eligible participants were those with a smoking history of at least 30 pack-years[2], were between 55 and 74 years of age, and if they had quit, had done so in the last 15 years. Participants were invited to three annual screenings, and the rate of adherence to these screenings was more than 90%. The NLST research team found a relative reduction of 20.0% in lung-cancer related mortality for the low-dose CT screening group compared to the radiography group. This corresponds to an absolute reduction of 62 deaths per person-years in the low-dose CT group compared to the radiography group. The analysis included all participants, including those for whom none of the screenings detected cancer and/or any nodules. For those participants who had no detection at any screening, it is plausible to say that the screening had a much smaller effect – in that case, this effect estimate for CT screening would be attenuated by the inclusion of these participants.

---

[1]The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial (NLST), under Project Number NLST-1459.

[2]packs-per-day*number of years smoked

For this reason, we are interested in focusing on the group of participants that have detectable lung cancer, that is, the participants who, if screened with low-dose CT, would have a detected nodule. Focusing on this group will help to isolate the effect of the low-dose CT screening, in the cases where the screening would matter the most.

Let D be the assigned type of screening ($D = 1$ is low-dose CT, $D = 0$ is chest radiography) and let M denote whether a person had cancer or a nodule detected at any of the screenings. Suppose we are interested in the binary outcome Y of whether a person died due to lung cancer. We can use our proposed approach to partially identify and estimate the TRACE, which is the effect of the low-dose CT screening on lung cancer mortality among those who, if screened with low-dose CT, would have had detected cancer or a pre-cancerous nodule.

The relevant quantity of interest from the original NLST analysis is the relative reduction in mortality rate (deaths per person-year) from lung cancer with low-dose CT screening, calculated as

$$\text{Relative reduction} = \frac{\frac{1}{M_{D=1}} \sum_{i:D_i=1} Y_i - \frac{1}{M_{D=0}} \sum_{i:D_i=0} Y_i}{\frac{1}{M_{D=0}} \sum_{i:D_i=0} Y_i} \tag{4.24}$$

where $M_{D=1}$ is the total person-years in the low-dose CT group and $M_{D=0}$ is the total person-years in the radiography group. In our application, we focus instead on the absolute reduction in mortality rate,

$$\text{Absolute reduction} = TE = \frac{1}{M_{D=1}} \sum_{i:D_i=1} Y_i - \frac{1}{M_{D=0}} \sum_{i:D_i=0} Y_i \tag{4.25}$$

In order to have comparable results to the original analysis, we use an adjusted outcome,

$$\tilde{Y}_i = \frac{Y_i N_{D=d}}{M_{D=d}} \tag{4.26}$$

where d is the treatment group to which unit i belongs, $N_{D=d}$ is the total number of units in that treatment group, and $M_{D=d}$ is the total number of person-years in that treatment group. This ensures that the coefficient on D from an unadjusted regression of $\tilde{Y}$ on D will

be equivalent to the estimated effect in Equation 4.25. We estimate a total effect of -63.3, or a reduction in 63.3 lung cancer deaths per 100,000 person years in the low-dose CT screening compared to the radiography group.[3] This is equivalent to a relative reduction of 20.3% in mortality from lung cancer in the low-dose CT group compared to the radiography group. This aligns with the intention-to-treat effect estimated in the original NLST results, with a small discrepancy due to the use of an updated (more complete) database compared to the one used for the primary paper.

TRACE Results. Figure 4.5 shows the estimated TRACE and corresponding confidence intervals at different postulated values of $TRACE(0)$. In this case, $TRACE(0)$ is the effect of the low-dose CT on lung cancer mortality among those who, if screened with the low-dose CT, would not have been diagnosed with cancer. When $TRACE(0) = TRACE$, the estimate for TRACE is equivalent to the estimate for the total effect. The total effect estimate and the corresponding confidence interval is shown in Figure 4.5 by the blue text. Note that the confidence interval for the total effect is smaller than that for TRACE when $TRACE(0) = TRACE$. This is because the TRACE is estimating the effect for a subgroup of the sample, while the total effect is considering the whole sample, and this reduction in sample size results in an increase in the standard error of our estimate.

We are more interested in values of the TRACE when $TRACE \neq TRACE(0)$, that is, when there is a different effect for the group who would be diagnosed compared to the group who would not be diagnosed. It is plausible in this setting that $TRACE(0) \approx 0$, as the new type of screening would have little to no effect for those people who received no new information from it [4]. When $TRACE(0) = 0$, our $TRACE$ estimate is -168.3. This means

[3]All rates are calculated as deaths per 100,000 person-years, adjusted for a calendar date cutoff for the follow-up time variable.

[4]It could be possible that a person had cancer that was not detected in any of the screenings, in which case their outcome would be the same as if they had never been screened. It is also possible that a person did not have cancer and so there was no cancer detected in any screening, in which case their outcome would likely be the same if they hadn't been screened. One could argue for a "backlash" effect, for example if

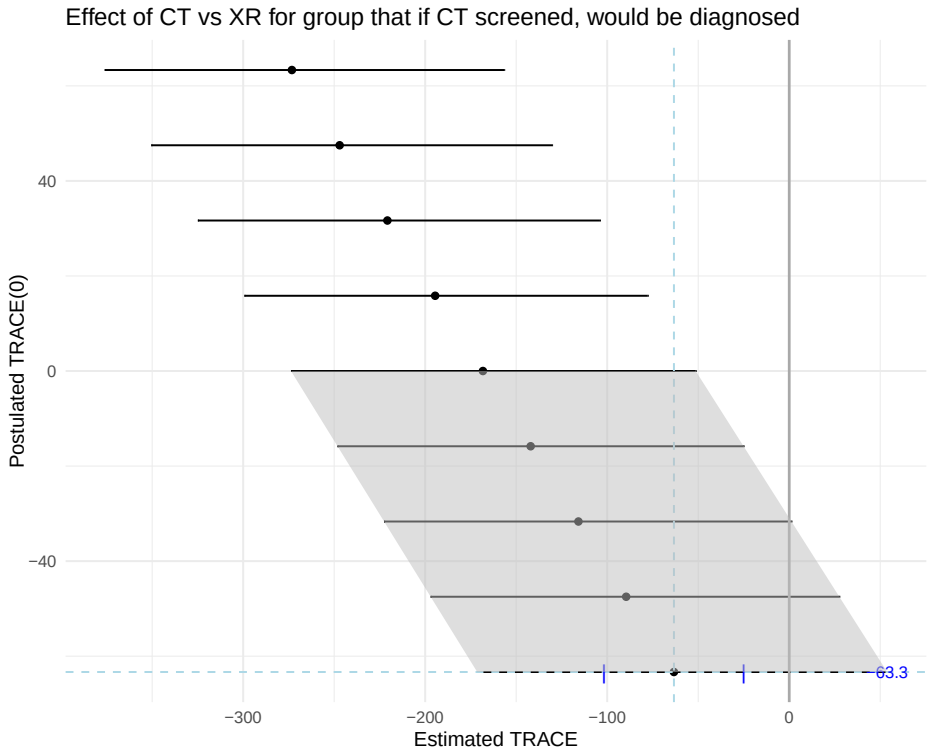


Figure 4.5: Effect of low-dose CT screening on mortality for lung cancer, among the group who would have been diagnosed with cancer or a precancerous nodule if screened.

that, for the group who would have detected cancer if screened with low-dose CT, there was a reduction of 168.3 deaths due to lung cancer (per 100,000 person-years) with low-dose CT screening compared to radiography. This estimated effect is larger than the intent-to-treat (or total effect) estimate, and suggests that the benefits from screening with low-dose CT for those who would be diagnosed are greater than previously established.

Suppose we do not believe that $TRACE(0) \approx 0$ but we believe it is plausible that $|TRACE(0)| \leq TRACE$, that is, the effect among those who would not be diagnosed if screened with CT is less than the effect among those who would be diagnosed if screened with CT. We could then estimate a range of plausible values for $TRACE$, shown by the

screening negative encouraged a person to live an unhealthy lifestyle, but we argue that these effects are negligible in this case.

shaded grey region in Figure 4.5. We see that, if $|TRACE(0)| \leq TRACE$, the $TRACE$ is between -63.3 and -168.3. Substantively, this means that for those who would be diagnosed if screened with CT, we estimate that the reduction in lung cancer mortality from CT screening compared to radiography is between 63.3 and 168.3 deaths per person-years. This indicates that the estimated effect for those who would be diagnosed if screened with low-dose CT is larger than that for the entire sample, and additionally puts an upper bound on how much larger the effect is for the subgroup of interest. For the "diagnosed" group to have a relative reduction in lung cancer mortality that is larger than 168.3, it would require that the "not diagnosed" group had an effect opposite in sign to the "diagnosed" group. This could happen if many people in this group had a "backlash" effect, for example, if they changed their behaviors to live a less healthy lifestyle in response to a negative screening.

Monotonicity and Trimming bounds In this setting, because the outcome of death is a rare event, the lower bound on $\mathbb{E}[Y(0)|\text{complier}]$ produced by the MT approach is equivalent to assuming that the average outcome among compliers is 0 – that is, nobody in the complier group dies from lung cancer. The upper bound on $\mathbb{E}[Y(0)|\text{complier}]$ is equivalent to assuming that nobody in the never-taker group dies from lung cancer. As both of these are infeasible assumptions to make, we do not apply the MT bounds in this setting. However, we may be able to apply partial identification approaches developed for settings with bounded outcomes to supplement our results (see e.g. Manski, 1990).

4.5 Conclusions

Randomized trials assessing the effects of cancer screening on severity and mortality often focus on the total effect of screening, which compares outcomes in the treated group to the control group. The total effect can be useful as a measure of effect, however, in the settings described, the total effect of the randomized intervention is often diluted as we expect the

effect to be strong only among a subgroup of the population defined by a post-treatment variable of interest. Conditioning on the post-treatment variable of interest will result in a biased effect estimate, so we must turn to alternative solutions. We propose the use of the Treatment Reactive Average Causal Effect, which is the total effect among the group who would realize a particular value of the post-treatment variable of interest if treated. In the screening setting, this allows us to focus on the group of people that would be diagnosed with cancer or a pre-cancerous nodule if screened.

Our approach is novel compared to existing approaches in that the choice of estimand allows for an identification strategy that does not rely on conditional ignorability assumptions or the exclusion restriction. The tradeoff is that we can only achieve partial identification, and must reason about the unknown TRACE(0) to use our approach. We emphasize that it is not necessary for researchers to determine one specific value for TRACE(0) – the bounds provided by partial identification are useful as they provide a truthful range of plausible effect estimates given minimal assumptions. Additionally, in some special cases, our approach allows for point identification of the TRACE, even when assumptions required for other point identification strategies do not hold.

The TRACE approach can be used to assess whether estimates of screening effects are being diluted by those who would not be diagnosed with cancer or a pre-cancerous nodule if screened, and allow us to determine how strong the effect could be for those who would be diagnosed. This is especially important in the screening setting, where the group of people who are diagnosed is very small in comparison to the group of those who are not. In the examples discussed, we show that the effect among the "diagnosed" subgroup can be quite different from the overall total effect. The proposed approach for partial identification of the TRACE addresses this issue by providing researchers with a principled way to reason about effects in a subgroup defined by a post-treatment variable.

APPENDIX A

Appendix for Chapter 1

A.1 Derivation of more general weights

The above weights rely on the assumption that $\mathbb{E}[D_i|X_i]$ is linear. We can derive more general weights that do not involve this assumption by using a more general representation of $D^{\perp X}$:

$$\hat{\tau}_{reg} = \frac{\widehat{\text{Cov}}(Y_i, D_i^{\perp X})}{\widehat{\mathbb{V}}(D_i^{\perp X})} \tag{A.1}$$

$$D^{\perp X} = D - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T D = D - \mathbf{X}\theta \tag{A.2}$$

where $\mathbf{X} = \begin{bmatrix}\vec{1} & \vec{X}\end{bmatrix}$. We first derive the unit-wise weights, which are somewhat trivial.

$$\begin{aligned}
\widehat{\text{Cov}}(Y_i, D_i^{\perp X}) &= \hat{\mathbb{E}}[Y_i D_i^{\perp X}] - \hat{\mathbb{E}}[Y_i]\hat{\mathbb{E}}[D_i^{\perp X}] && \text{(A.3)}\\
&= \hat{\mathbb{E}}[Y_i D_i^{\perp X}] && \text{(A.4)}\\
&= \sum_i Y_i(D_i - \mathbf{X}_i\theta) && \text{(A.5)}\\
&= \sum_i Y_i(D_i - \hat{d}(X_i)) && \text{(A.6)}\\
& && \text{(A.7)}
\end{aligned}$$

where $\hat{d}(X_i) = \mathbf{X}_i\beta$ where $\mathbf{X}_i = \begin{bmatrix}1 & X_i\end{bmatrix}$ and $X_i = x$.

$$\widehat{\mathbb{V}}(D_i^{\perp X}) = \hat{\mathbb{E}}[D_i^{\perp X^2}] - \hat{\mathbb{E}}[D_i^{\perp X}]^2 \tag{A.8}$$

$$= \hat{\mathbb{E}}[D_i^{\perp X^2}] \tag{A.9}$$

$$= \sum_i (D_i - \mathbf{X}_i\theta)^2 \tag{A.10}$$

$$= \sum_i (D_i - \hat{d}(X_i))^2 \tag{A.11}$$

$$\tag{A.12}$$

So, the unit-wise weighted representation is as follows:

$$\hat{\tau}_{reg} = \frac{\widehat{\text{Cov}}(Y_i, D_i^{\perp X})}{\widehat{\mathbb{V}}(D_i^{\perp X})} = \frac{\sum_i Y_i(D_i - \hat{d}(X_i))}{\sum_i (D_i - \hat{d}(X_i))^2} \tag{A.13}$$

Let $\pi_x = P(D_i = 1|X_i = x)$ and $\hat{d}(X_i) = \mathbf{X}_i\theta$ where $\mathbf{X}_i = \begin{bmatrix} 1 & X_i \end{bmatrix}$ and $X_i = x$. We can try to manipulate the above representation to get the strata-wise weights:

$$\widehat{\text{Cov}}(Y_i, D_i^{\perp X}) = \sum_x P(X = x)\hat{\mathbb{E}}[Y_i(D_i - \hat{d}(X_i))|X_i] \tag{A.14}$$

$$= \sum_x P(X = x)\hat{\mathbb{E}}[Y_i D_i|X_i] - \hat{d}(X_i)\hat{\mathbb{E}}[Y_i|X_i] \tag{A.15}$$

$$= \sum_x P(X = x)(P(D_i = 1|X_i)\hat{\mathbb{E}}[Y_i|D_i = 1, X_i] - \hat{d}(X_i)\hat{\mathbb{E}}[Y_i|X_i]) \tag{A.16}$$

$$= \sum_x P(X = x)(\pi_x\hat{\mathbb{E}}[Y_i|D_i = 1, X_i] - \hat{d}(X_i)(\pi_x\hat{\mathbb{E}}[Y_i|D_i = 1, X_i] \tag{A.17}$$

$$+ (1 - \pi_x)\hat{\mathbb{E}}[Y_i|D_i = 0, X_i))$$

$$= \sum_x P(X = x)(\pi_x(1 - \hat{d}(X_i))\hat{\mathbb{E}}[Y_i|D_i = 1, X_i] \tag{A.18}$$

$$- \hat{d}(X_i)(1 - \pi_x)\hat{\mathbb{E}}[Y|D_i = 0, X_i])$$

This can be rewritten in a couple different ways, where τ_x is the ATE given $X = x$:

$$\widehat{\text{Cov}}(Y_i, D_i^{\perp X}) = \sum_x P(X = x)(\hat{\mathbb{E}}[Y_i|D_i = 1, X_i](\pi_x(1 - \hat{d}(X_i))) \tag{A.19}$$

$$- \hat{\mathbb{E}}[Y|D_i = 0, X_i](\hat{d}(X_i)(1 - \pi_x))$$

$$= \sum_x P(X = x)(\pi_x \hat{\mathbb{E}}[Y_i|D_i = 1, X_i] - \hat{d}(X_i)\hat{\mathbb{E}}[Y_i|D_i = 0, X_i] - \hat{d}(X_i)\pi_x\tau_x \tag{A.20}$$

$$= \sum_x P(X = x)\pi_x(1 - \hat{d}(X_i))(\hat{\mathbb{E}}[Y_i|D_i = 1, X_i] \tag{A.21}$$

$$- \frac{\hat{d}(X_i)}{1 - \hat{d}(X_i)} \frac{1 - \pi_i}{\pi_i} \hat{\mathbb{E}}[Y_i|D_i = 0, X_i])$$

Next, to get the full expression for $\hat{\tau}_{reg}$, we calculate $\mathbb{V}(D_i^{\perp X})$.

$$\widehat{\mathbb{V}}(D_i^{\perp X}) = \sum_x P(X = x)\hat{\mathbb{E}}[D_i^{\perp X^2}|X] \tag{A.22}$$

$$= \sum_x P(X = x)\hat{\mathbb{E}}[D_i|X_i] - 2\mathbf{X}_i\beta\hat{\mathbb{E}}[D_i|X_i] + (\mathbf{X}_i\beta)^2 \tag{A.23}$$

$$= \sum_x P(X = x)(\pi_x - 2\hat{d}(X_i)\pi_x + \hat{d}(X_i)^2) \tag{A.24}$$

$$= \sum_x P(X = x)(\pi_x(1 - \hat{d}(X_i)) + \hat{d}(X_i)(\hat{d}(X_i) - \pi_x)) \tag{A.25}$$

So we end up with the following expression for $\hat{\tau}_{reg}$.

$$\hat{\tau}_{reg} = \frac{\sum_x P(X = x)(\hat{\mathbb{E}}[Y_i|D_i = 1, X_i](\pi_x(1 - \hat{d}(X_i))) - \hat{\mathbb{E}}[Y|D_i = 0, X_i](\hat{d}(X_i)(1 - \pi_x))}{\sum_x P(X = x)(\pi_x(1 - \hat{d}(X_i)) + \hat{d}(X_i)(\hat{d}(X_i) - \pi_x))} \tag{A.26}$$

Suppose we let $\hat{d}(X_i) = \pi_x + a_x$, where a_x is the difference between the true probability of treatment given $X = x$ and the probability of treatment estimated by a linear model. We

can then write $\widehat{\text{Cov}}(Y_i, D_i^{\perp X})$ in terms of a_x:

$$\widehat{\text{Cov}}(Y_i, D_i^{\perp X}) = \sum_x P(X = x)\pi_x \hat{\mathbb{E}}[Y_i|D_i = 1, X_i] - \hat{d}(X_i)\hat{\mathbb{E}}[Y_i|D_i = 0, X_i] - \hat{d}(X_i)\pi_x\tau_x \tag{A.27}$$

$$= \sum_x P(X = x)\pi_x \hat{\mathbb{E}}[Y_i|D_i = 1, X_i] - (\pi_x + a_x)\hat{\mathbb{E}}[Y_i|D_i = 0, X_i] \tag{A.28}$$

$$- (\pi_x + a_x)\pi_x\tau_x$$

$$= \sum_x P(X = x)\pi_x \hat{\mathbb{E}}[Y_i|D_i = 1, X_i] - \pi_x\hat{\mathbb{E}}[Y_i|D_i = 0, X_i] - a_x\hat{\mathbb{E}}[Y_i|D_i = 0, X_i] \tag{A.29}$$

$$- \pi_x^2\tau_x - a_x\pi_x\tau_x$$

$$= \sum_x P(X = x)\pi_x(1 - \pi_x)\tau_x - a_x(\hat{\mathbb{E}}[Y_i|D_i = 0, X_i] + \pi_x\tau_x) \tag{A.30}$$

$$= \sum_x P(X = x)\pi_x(1 - \pi_x)\tau_x - a_x\hat{\mathbb{E}}[Y_i|X_i = x] \tag{A.31}$$

$$\tag{A.32}$$

Using this, we can write the weighted representation of the regression coefficient in terms of a_x:

$$\hat{\tau}_{reg} = \frac{\sum_x P(X = x)\pi_x(1 - \pi_x)\tau_x - a_x\hat{\mathbb{E}}[Y_i|X_i = x]}{\sum_x P(X = x)(\pi_x(1 - \pi_x - a_x) + (\pi_x + a_x)(\pi_x + a_x - \pi_x))} \tag{A.33}$$

A.2 Equivalency of impute and interact

We can show that the regression imputation estimate is equivalent to the Lin/interacted regression estimate by showing that the minimization problems solved by the two methods are equivalent.

We can start with the Lin estimate. We want to estimate $\hat{\beta}$ in the equation:

$$\mathbb{E}[Y|D, X] = \hat{\beta}_0 + \hat{\beta}_1 D + \hat{\beta}_2 X + \hat{\beta}_3 DX$$

We do this by minimizing the least squares error, in other words, $\hat{\beta}$ is the solution to the minimization problem:

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i X_i))^2$$

Now let's look at the regressions involved in the imputation estimate. The imputation estimate is $\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \hat{Y}_i(1) - \hat{Y}_i(0) = \frac{1}{N} \sum_{i=1}^{N} \gamma_0 + \gamma_1 X - (\alpha_0 + \alpha_1 X)$. Here, γ and α are found by solving the following minimization problems:

$$\min_{\gamma} \sum_{D_i=1} (Y_i - (\gamma_0 + \gamma_1 X_i))^2 \tag{A.34}$$

$$\min_{\alpha} \sum_{D_i=0} (Y_i - (\alpha_0 + \alpha_1 X_i))^2 \tag{A.35}$$

$$\tag{A.36}$$

This is equivalent to solving the single minimization problem:

$$= \min_{\gamma,\alpha} \sum_{D_i=1} (Y_i - (\gamma_0 + \gamma_1 X_i))^2 + \sum_{D_i=0} (Y_i - (\alpha_0 + \alpha_1 X_i))^2 \tag{A.37}$$

$$= \min_{\gamma,\alpha} \sum_{i=1}^{n} D_i(Y_i - (\gamma_0 + \gamma_1 X_i))^2 + (1 - D_i)(Y_i - (\alpha_0 + \alpha_1 X_i))^2 \tag{A.38}$$

$$= \min_{\gamma,\alpha} \sum_{i=1}^{n} D_i(Y_i^2 - 2Y_i(\gamma_0 + \gamma_1 X_i) + (\gamma_0 + \gamma_1 X_i)^2) + (1 - D_i)(Y_i^2 - 2Y_i(\alpha_0 \tag{A.39}$$

$$+ \alpha_1 X_i) + (\alpha_0 + \alpha_1 X_i)^2)$$

$$= \min_{\gamma,\alpha} \sum_{i=1}^{n} Y_i^2 - 2Y_i(\alpha_0 + \alpha_1 X_i + (\gamma_0 - \alpha_0)D_i + (\gamma_1 - \alpha_1)D_i X_i) + (\alpha_0 + \alpha_1 X_i + (\gamma_0 - \alpha_0)D_i \tag{A.40}$$

$$+ (\gamma_1 - \alpha_1)D_i X_i)^2$$

$$= \min_{\gamma,\alpha} \sum_{i=1}^{n} (Y_i - (\alpha_0 + \alpha_1 X_i + (\gamma_0 - \alpha_0)D_i + (\gamma_1 - \alpha_1)D_i X_i))^2 \tag{A.41}$$

$$= \min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i X_i))^2 \tag{A.42}$$

where $\beta_0 = \alpha_0$, $\beta_1 = \gamma_0 - \alpha_0$, $\beta_2 = \alpha_1$, and $\beta_3 = \gamma_1 - \alpha_1$

A.3 Unbiasedness of weighted DiM using mean balancing weights

Assume the potential outcomes are separately linear in X:

$$\mathbb{E}[Y_i(1)|X_i] = \gamma X_i \tag{A.43}$$

$$\mathbb{E}[Y_i(0)|X_i] = \alpha X_i \tag{A.44}$$

Then we can show that the ATE estimate from using mean balancing weights is unbiased:

$$\mathbb{E}\left[\hat{\tau}_{meanbal}\right] = \mathbb{E}\left[\sum_{D_i=1} Y_i w_i - \sum_{D_i=0} Y_i w_i\right] \tag{A.45}$$

$$= \sum_{D_i=1} \mathbb{E}[Y_i w_i | D_i = 1] - \sum_{D_i=0} \mathbb{E}[Y_i w_i | D_i = 0] \tag{A.46}$$

$$= \sum_{D_i=1} \mathbb{E}[Y_i(1) w_i | D_i = 1] - \sum_{D_i=0} \mathbb{E}[Y_i(0) w_i | D_i = 0] \tag{A.47}$$

$$= \sum_{D_i=1} \mathbb{E}[(\gamma X_i + \epsilon_i) w_i | D_i = 1] - \sum_{D_i=0} \mathbb{E}[(\alpha X_i + \epsilon_i) w_i | D_i = 0] \tag{A.48}$$

$$= \gamma \sum_{D_i=1} \mathbb{E}[X_i w_i | D_i = 1] - \alpha \sum_{D_i=0} \mathbb{E}[X_i w_i | D_i = 0] \tag{A.49}$$

$$= \gamma \mathbb{E}\left[\sum_{D_i=1} X_i w_i\right] - \alpha \mathbb{E}\left[\sum_{D_i=0} X_i w_i\right] \tag{A.50}$$

$$= \gamma \mathbb{E}\left[\bar{X}\right] - \alpha \mathbb{E}\left[\bar{X}\right] \tag{A.51}$$

$$= \mathbb{E}[\overline{Y_i(1)}] - \mathbb{E}[\overline{Y_i(0)}] \tag{A.52}$$

$$= \mathbb{E}[Y_i(1) - Y_i(0)] \tag{A.53}$$

where $\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$ is the sample mean of the covariates.

A.4 Unbiasedness of impute and interact

We can show that both the regression imputation method and the interacted regression are unbiased for the ATE. We start with regression imputation. We have one model for Y among the treated, $\hat{\mu}_1$, and one among the controls, $\hat{\mu}_0$.

Then,

$$\hat{\tau}_{imp} = \frac{1}{n}\left(\sum_{D_i=1}(Y_i - \hat{\mu}_0(X_i)) + \sum_{D_i=0}(\hat{\mu}_1(X_i) - Y_i)\right) \tag{A.54}$$

$$= \frac{1}{n}\sum_i(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) \tag{A.55}$$

$$= \sum_{x\in\mathcal{X}} \frac{\mathbb{1}_{\{X=x\}}}{n}\left(\hat{\mu}_1(x) - \hat{\mu}_0(x)\right) \tag{A.56}$$

$$= \sum_{x\in\mathcal{X}} \widehat{P}(X=x)\left(\hat{\mu}_1(x) - \hat{\mu}_0(x)\right) \tag{A.57}$$

$$= \sum_{x\in\mathcal{X}} \widehat{P}(X=x) DIM_x \tag{A.58}$$

We see that the imputation approach weights each stratum by the probability of inclusion in that stratum – these weights result in an estimate that is unbiased for the ATE.

APPENDIX B

Appendix for Chapter 2

B.1 Equivalent alternative quantities to reason about

While we regard the TRACE(0) as entirely unknown, it has two components, one of which ($\mathbb{E}[Y(1) \mid M(1) = 0]$) can be identified (by $E[Y(1) \mid D = 1, M = 0]$) under randomization. The difficulty comes with the second component, $\mathbb{E}[Y(0) \mid M(1) = 0]$. Randomization allows this to be written as $\mathbb{E}[Y(0) \mid D = 1, M = 0]$, but it remains unidentified. Assuming no defiers, $\mathbb{E}[Y(0) \mid D = 1, M = 0]$ is simply the expected non-treatment outcome among never-takers ("NT"), $\mathbb{E}[Y(0) \mid \text{NT}]$. Thus, one has the choice of either reasoning directly about $\mathbb{E}[Y(0) \mid \text{NT}]$, or reasoning about the entirety of TRACE(0).

Second, one could also choose to reason about the average non-treatment outcome among compliers. This is because we observe an estimate of $\mathbb{E}[Y(0) \mid D = 0, M = 0]$, which is composed of averages among compliers ("C") and never-takers ("NT"),

$$\begin{aligned}\mathbb{E}[Y(0) \mid D = 0, M = 0] = \mathbb{E}[Y(0) \mid \ \text{NT}]\frac{Pr(\text{NT})}{Pr(\text{NT}) + Pr(\text{C})} \\ + \mathbb{E}[Y(0) \mid \ \text{C}]\frac{Pr(\text{C})}{Pr(\text{NT}) + Pr(\text{C})}\end{aligned} \tag{B.1}$$

Having observed the left-hand side, assuming a value for either $\mathbb{E}[Y(0) \mid \text{C}]$ or $\mathbb{E}[Y(0) \mid \text{NT}]$ is enough to fix the other. That is, one could make an assumption on $\mathbb{E}[Y(0) \mid \ \text{C}]$, use the value of $\mathbb{E}[Y(0) \mid D = 0, M = 0]$ to back-out $\mathbb{E}[Y(0) \mid \text{NT}]$, and use that in turn to compute the TRACE(0) and identify the TRACE.

This leaves us with a choice among three different quantities that we could reason about

in order to get a range of estimates for TRACE: TRACE(0), $\mathbb{E}[Y(0) \mid \text{NT}]$, or $\mathbb{E}[Y(0) \mid \text{C}]$. Note, however, that using the latter two quantities depends on an assumption of no defiers, which is not required if we directly reason about TRACE(0). All three of these routes to (partial) identification are equivalent, but some may be better suited to reasoning and argumentation than others in a given context. In the cases examined here, we found it profitable to reason directly about TRACE(0) because it refers to a substantive causal effect in a sub-group, and the nature of this sub-group can support arguments about TRACE(0), e.g. that is might have the opposite sign as the TRACE, that it is likely to be zero, negative, smaller than the effect in the other group, or something else useful for bounding.

B.2 Latent DAG representation

One difficulty of using DAGs in this context is that the estimand we propose conditions on membership in a group—those units with $M(1) = 1$—that cannot be conditioned on using observed variables, and thus cannot be directly represented by a conditioning procedure on the DAG. To represent our estimand thus requires modifying what appears on the DAG. Specifically, consider the value of $\{M(1), M(0)\}$ for each unit i as a (latent) variable. The characteristic $M(1) = 1$ represents a latent "type", or membership in either of two principal strata—that with $\{M(1) = 1, M(0) = 1\}$ (or "always-takers") and that with $\{M(1) = 1, M(0) = 0\}$ ("compliers"). The value of $M(1)$ is a function of the unobservables, U, and can therefore be represented as a descendant of U on the graph. The realized $M = M(d)$ is a (deterministic) function of $M(1)$, $M(0)$, and D. These relationships are encoded in Figure B.1. A key feature of this relationship is that there is no $D \to M(1)$ edge, so conditioning on $M(1)$ (and/or $M(0)$) does not jeopardize identification of the $D \to Y$ effect.

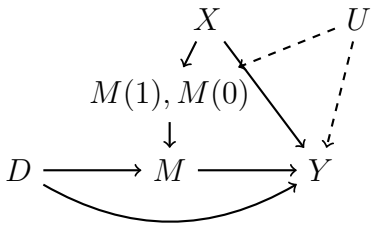


Figure B.1: Latent DAG representation

B.3 Comparing to CDE

In the canvassing application (see Section 3.4.3), our approach and estimand can be compared with that of Blackwell et al. [2024], who use the same empirical application to examine the controlled direct effect (CDE) of the canvassing intervention on support for the non-discrimination law, hypothetically holding constant each participant's feeling thermometer score, leaving only the direct effect of the intervention not through subjective feelings. As noted above, this is very different from our estimand, which is the *total* effect of the canvassing intervention on policy support, among those whose feelings thermometer scores would have increased from baseline if given the intervention. In other words, while the CDE attempts to eliminate the effect through "changes in warmth," we are interested in how the intervention works both directly and through changing warmth, in the sub-group that would show increased warmth due to the intervention. We could expect our TRACE estimate to be larger in magnitude than the CDE, as we are considering both direct and indirect effects and focusing on the group which we believe to have stronger effects. However, this may not be the case if the direct and indirect effects are in opposite directions. Our estimand also differs from the "indirect effect" quantities found in mediation settings, again because we are interested in both the indirect and direct effects, but for a subgroup that cannot be parsed out in mediation analysis.

APPENDIX C

Appendix for Chapter 3

C.1 Calculating variance of TRACE

We want to find

$$\mathbb{V}(\widehat{TRACE}) = \mathbb{V}\left(\frac{\overline{Y_1} - \overline{Y_0} - TRACE(0)\frac{n_{trt} - \sum_{j \in trt} M_j}{n_{trt}}}{\frac{1}{n_{trt}} \sum_{j \in trt} M_j}\right) \tag{C.1}$$

Let

$$B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \overline{Y_1} - \overline{Y_0} - TRACE(0)\frac{n_{trt} - \sum_{j \in trt} M_j}{n_{trt}} \\ \frac{1}{n_{trt}} \sum_{j \in trt} M_j \end{pmatrix} \tag{C.2}$$

and

$$\beta = \begin{pmatrix} TE - TRACE(0)(1 - Pr(M = 1|D = 1)) \\ Pr(M = 1|D = 1) \end{pmatrix} \tag{C.3}$$

It follows from established results about the difference-in-means estimate that $\beta = \mathbb{E}[B]$. We will use the delta method to find the asymptotic variance of $TRACE$. In order to do this, we first need to establish that

$$\sqrt{n}(B - \beta) \xrightarrow{D} \mathcal{N}(0, \Sigma) \tag{C.4}$$

We can do this using a multidimensional version of the Central Limit Theorem. Let probability of treatment be fixed at $p = Pr(D_i = 1)$. We can then approximate $n_{trt} \approx np$.

The elements of B can then be re-written as follows:

$$b_1 = \overline{Y_1} - \overline{Y_0} - TRACE(0)\frac{n_{trt} - \sum_{i \in trt} M_i}{n_{trt}} \tag{C.5}$$

$$= \frac{1}{np}\sum_{i \in trt} Y_i - \frac{1}{n - np}\sum_{i \in ctrl} Y_i - TRACE(0)(1 - \frac{1}{np}\sum_{trt} M_i) \tag{C.6}$$

$$= \frac{1}{n}\left(\frac{1}{p}\sum_{i \in trt} Y_i - \frac{1}{1-p}\sum_{i \in ctrl} Y_i - nTRACE(0) + TRACE(0)\frac{1}{p}\sum_{trt} M_i\right) \tag{C.7}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{p}Y_i D_i - \frac{1}{1-p}Y_i(1 - D_i) - TRACE(0) + \frac{TRACE(0)}{p}M_i D_i\right) \tag{C.8}$$

$$b_2 = \frac{1}{np}\sum_{i \in trt} M_i \tag{C.9}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p}M_i D_i \tag{C.10}$$

Define $T_i := \begin{pmatrix} t_i^{(1)} \\ t_i^{(2)} \end{pmatrix} = \begin{pmatrix} \frac{1}{p}Y_i D_i - \frac{1}{1-p}Y_i(1 - D_i) - TRACE(0) + \frac{TRACE(0)}{p}M_i D_i \\ \frac{1}{p}M_i D_i \end{pmatrix}$.

Then, $B = \overline{\mathbf{T}}_{\mathbf{n}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{T_i}$, where $\mathbf{T_i}$ are random vectors that are independent and identically distributed. By the Central Limit Theorem,

$$\sqrt{N}(B - \beta) \xrightarrow{D} \mathcal{N}(0, \Sigma) \tag{C.11}$$

where $\beta = \mathbb{E}[B]$ is defined above and $\Sigma = \begin{pmatrix} \mathbb{V}(t_1^{(1)}) & \mathrm{Cov}(t_1^{(1)}, t_1^{(2)}) \\ \mathrm{Cov}(t_1^{(1)}, t_1^{(2)}) & \mathbb{V}(t_1^{(2)}) \end{pmatrix}$.

The delta method implies that, if $\sqrt{N}(B - \beta) \xrightarrow{D} \mathcal{N}(0, \Sigma)$, and h is a scalar valued function of B such that $h'(\beta)$ exists and is non-zero valued, then

$$\sqrt{n}(h(B) - h(\beta)) \xrightarrow{D} \mathcal{N}(0, \nabla h(\beta)^T \Sigma \nabla h(\beta)) \tag{C.12}$$

This follows from calculating the variance of $h(B)$ using its first order Taylor approximation. If we know Σ, we can then use delta method to get the variance of TRACE by defining

$h(B) = \frac{b_1}{b_2} = \widehat{TRACE}$. Then,

$$\sqrt{n}(\widehat{TRACE} - TRACE) \xrightarrow{D} \mathcal{N}(0, \nabla h(\beta)^T \Sigma \nabla h(\beta)) \tag{C.13}$$

In order to get this quantity, we first need to estimate Σ, and then apply the delta method. We leave this calculation for future work.

Bibliography

Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, 2006. doi: https://doi.org/10.1111/j.1468-0262.2006.00655.x.

Avidit Acharya, Matthew Blackwell, and Maya Sen. Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects. *American Political Science Review*, 110(3): 512–529, August 2016. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055416000216.

Ramzi Amri, Liliana G. Bordeianou, Patricia Sylla, and David L. Berger. Impact of Screening Colonoscopy on Outcomes in Colon Cancer Surgery. *JAMA Surgery*, 148(8):747–754, August 2013. ISSN 2168-6254. doi: 10.1001/jamasurg.2013.8.

Joshua D. Angrist. Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants. *Econometrica*, 66(2):249–288, 1998. Publisher: Econometric Society.

Joshua D. Angrist and Alan B. Krueger. Empirical strategies in labor economics. In *Handbook of Labor Economics*, volume 3, Part A, pages 1277–1366. Elsevier, 1999.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009. ISBN 9780691120348.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434):444–455, June 1996. ISSN 0162-1459. doi: 10.1080/01621459.1996.10476902. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1996.10476902.

Peter M. Aronow and Cyrus Samii. Does Regression Produce Representative Estimates

of Causal Effects? *American Journal of Political Science*, 60(1):250–267, 2016. ISSN 1540-5907. doi: 10.1111/ajps.12185.

Stuart G. Baker and Karen S. Lindeman. The paired availability design: A proposal for evaluating epidural analgesia during labor. *Statistics in Medicine*, 13 (21):2269–2278, 1994. ISSN 1097-0258. doi: 10.1002/sim.4780132108. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780132108.

William A Belson. A technique for studying the effects of a television broadcast. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 5(3):195–202, 1956.

Matthew Blackwell, Adam Glynn, Hanno Hilbig, and Connor Halloran Phillips. Estimating Controlled Direct Effects with Panel Data: An Application to Reducing Support for Discriminatory Policies. *Working Paper*, January 2024.

Graeme Blair, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. Declaring and diagnosing research designs. *American Political Science Review*, 113(3):838–859, 2019.

Graeme Blair, Jeremy M. Weinstein, Fotini Christia, Eric Arias, Emile Badran, Robert A. Blair, Ali Cheema, Ahsan Farooqui, Thiemo Fetzer, Guy Grossman, Dotan Haim, Zulfiqar Hameed, Rebecca Hanson, Ali Hasanain, Dorothy Kronick, Benjamin S. Morse, Robert Muggah, Fatiq Nadeem, Lily L. Tsai, Matthew Nanes, Tara Slough, Nico Ravanilla, Jacob N. Shapiro, Barbara Silva, Pedro C. L. Souza, and Anna M. Wilke. Community policing does not build citizen trust in police or reduce crime in the Global South. *Science*, 374(6571):eabd3446, November 2021. doi: 10.1126/science.abd3446. Publisher: American Association for the Advancement of Science.

Christine Blandhol, John Bonney, Magne Mogstad, and Alexander Torgovitsky. When is tsls actually late? Working Paper 29709, National Bureau of Economic Research, 2022.

Alan S. Blinder. Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455, 1973. ISSN 0022166X.

Kirill Borusyak and Peter Hull. Negative weights are no concern in design-based specifications. In *AEA Papers and Proceedings*, volume 114, pages 597–600. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2024.

Michael Bretthauer, Magnus Løberg, Paulina Wieszczy, Mette Kalager, Louise Emilsson, Kjetil Garborg, Maciej Rupinski, Evelien Dekker, Manon Spaander, Marek Bugajski, Øyvind Holme, Ann G. Zauber, Nastazja D. Pilonis, Andrzej Mroz, Ernst J. Kuipers, Joy Shi, Miguel A. Hernán, Hans-Olov Adami, Jaroslaw Regula, Geir Hoff, and Michal F. Kaminski. Effect of Colonoscopy Screening on Risks of Colorectal Cancer and Related Death. *New England Journal of Medicine*, 387(17):1547–1556, October 2022. ISSN 0028-4793. doi: 10.1056/NEJMoa2208375. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa2208375.

David Broockman and Joshua Kalla. Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352(6282):220–224, 2016. ISSN 0036-8075. Publisher: American Association for the Advancement of Science.

Cam Nhung Bui, Seri Hong, Mina Suh, Jae Kwan Jun, Kyu Won Jung, Myong Cheol Lim, and Kui Son Choi. Effect of Pap smear screening on cervical cancer stage at diagnosis: results from the Korean National Cancer Screening Program. *Journal of Gynecologic Oncology*, 32(5):e81, June 2021. ISSN 2005-0380. doi: 10.3802/jgo.2021.32.e81.

Ambarish Chattopadhyay and José R Zubizarreta. On the implied weights of linear regression for causal inference. *Biometrika*, 110(3):615–629, 2023.

Ambarish Chattopadhyay and José R. Zubizarreta. Causation, Comparison, and Regression. *Harvard Data Science Review*, 6(1), jan 31 2024. https://hdsr.mitpress.mit.edu/pub/1ybwbmlw.

Ambarish Chattopadhyay, Noah Greifer, and Jose R. Zubizarreta. lmw: Linear Model Weights for Causal Inference, March 2023. arXiv:2303.08790 [stat].

Victor Chernozhukov, Iván Fernández-Val, Jinyong Hahn, and Whitney Newey. Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580, 2013.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547):1778–1793, 2024.

Brian L. Egleston, Daniel O. Scharfstein, and Ellen MacKenzie. On Estimation of the Survivor Average Causal Effect in Observational Studies when Important Confounders are Missing Due to Death. *Biometrics*, 65(2):497–504, June 2009. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2008.01111.x.

Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933. ISSN 00129682, 14680262.

Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár. Contamination bias in linear regressions. Technical report, National Bureau of Economic Research, 2022.

Bronner P Gonçalves and Etsuji Suzuki. Causal approaches to disease progression analyses. *Epidemiology*, pages 10–1097, 2025.

Donald P. Green and Peter M. Aronow. Analyzing Experimental Data Using Regression: When is Bias a Practical Concern?, March 2011.

Jinyong Hahn. Properties of least squares estimator in estimation of average treatment effects. *SERIEs*, 14(3):301–313, 2023.

Jens Hainmueller. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20

(1):25–46, 2012. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpr025. Publisher: Cambridge University Press.

Douglas Hayden, Donna K. Pauler, and David Schoenfeld. An Estimator for Treatment Comparisons among Survivors in Randomized Trials. *Biometrics*, 61(1):305–310, 2005. ISSN 0006-341X. Publisher: [Wiley, International Biometric Society].

Chad Hazlett. Kernel balancing. *Statistica Sinica*, 30(3):1155–1189, 2020.

Chad Hazlett, Nina McMurry, and Tanvi Shinkre. Post-treatment problems: What can we say about the effect of a treatment among sub-groups who (would) respond in some way? *arXiv preprint arXiv:2505.06754*, 2025.

Miguel A Hernán and Sonia Hernández-Díaz. Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials*, 9(1):48–55, February 2012. ISSN 1740-7745. doi: 10.1177/1740774511420743. Publisher: SAGE Publications.

Miguel A. Hernán and James M. Robins. Per-Protocol Analyses of Pragmatic Trials. *New England Journal of Medicine*, 377(14):1391–1398, October 2017. ISSN 0028-4793. doi: 10.1056/NEJMsm1605385. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMsm1605385.

Miguel A. Hernán, Sonia Hernández-Díaz, and James M. Robins. Randomized Trials Analyzed as Observational Studies. *Annals of Internal Medicine*, 159(8):560–562, October 2013. ISSN 0003-4819. doi: 10.7326/0003-4819-159-8-201310150-00709. Publisher: American College of Physicians.

Nathan I Hoffmann. Double robust, flexible adjustment methods for causal inference: An overview and an evaluation. *url: https://osf.io/preprints/socarxiv/dzayg*, Aug 2023. doi: 10.31235/osf.io/dzayg. URL `osf.io/preprints/socarxiv/dzayg`.

Joel L Horowitz and Charles F Manski. Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, pages 281–302, 1995.

Michael G. Hudgens and M. Elizabeth Halloran. Causal Vaccine Effects on Binary Postinfection Outcomes. *Journal of the American Statistical Association*, 101(473):51–64, March 2006. ISSN 0162-1459. doi: 10.1198/016214505000000970.

Macartan Humphreys. Bounds on least squares estimates of causal effects in the presence of heterogeneous assignment probabilities. 2009. URL `http://www.columbia.edu/~mh2245/papers1/monotonicity7.pdf`.

Kosuke Imai. Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statistics & Probability Letters*, 78(2):144–149, February 2008. ISSN 01677152. doi: 10.1016/j.spl.2007.05.015.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334, 2010a. ISSN 1939-1463, 1082-989X. doi: 10.1037/a0020761.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25(1):51–71, February 2010b. ISSN 0883-4237, 2168-8745. doi: 10.1214/10-STS321. Publisher: Institute of Mathematical Statistics.

Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. Comment on Pearl: Practical implications of theoretical results for causal mediation analysis. *Psychological Methods*, 19:482–487, 2014. ISSN 1939-1463. doi: 10.1037/met0000021. Place: US Publisher: American Psychological Association.

Guido W. Imbens. Matching methods in practice: Three examples. *The Journal of Human Resources*, 50(2):373–419, 2015. ISSN 0022166X.

Guido W. Imbens and Joshua D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994. ISSN 0012-9682. doi: 10.2307/2951620. Publisher: [Wiley, Econometric Society].

Guido W. Imbens and Jeffrey M. Wooldridge. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1):5–86, March 2009. ISSN 0022-0515. doi: 10.1257/jel.47.1.5.

Joseph D. Y. Kang and Joseph L. Schafer. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523–539, November 2007. ISSN 0883-4237, 2168-8745. doi: 10.1214/07-STS227. Publisher: Institute of Mathematical Statistics.

Luke Keele, Ismail K White, and Corrine McConnaughy. Adjusting experimental data: Models versus design. In *APSA 2010 Annual Meeting Paper*, 2010.

Patrick Kline. Oaxaca-blinder as a reweighting estimator. *American Economic Review*, 101 (3):532–37, May 2011. doi: 10.1257/aer.101.3.532.

Dean Knox, Will Lowe, and Jonathan Mummolo. Administrative Records Mask Racially Biased Policing. *American Political Science Review*, 114(3):619–637, August 2020. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055420000039.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, March 2019. doi: 10.1073/pnas.1804597116. Publisher: Proceedings of the National Academy of Sciences.

David S Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Working Paper 11721, National Bureau of Economic Research, October 2005. URL `http://www.nber.org/papers/w11721`.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, March 2013. ISSN 1932-6157, 1941-7330. doi: 10.1214/12-AOAS583. Publisher: Institute of Mathematical Statistics.

Michael C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010, 1963. ISSN 01621459.

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Charles F Manski and John V Pepper. Monotone instrumental variables with an application to the returns to schooling, 1998.

Jacob M Montgomery, Brendan Nyhan, and Michelle Torres. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775, 2018.

Benjamin S. Morse. Strengthening the Rule of Law Through Community Policing: Evidence From Liberia. *Comparative Political Studies*, page 00104140241252090, May 2024. ISSN 0010-4140. doi: 10.1177/00104140241252090. Publisher: SAGE Publications Inc.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.

The NLST Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011. doi: 10.1056/NEJMoa1102873.

Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, 1973. ISSN 00206598, 14682354.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Judea Pearl. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4):459–481, 2014. ISSN 1939-1463, 1082-989X. doi: 10.1037/a0036434.

Charles C Peters. A method of matching groups for experiment with no loss of population. *The Journal of Educational Research*, 34(8):606–612, 1941.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, January 1986. ISSN 0270-0255. doi: 10.1016/0270-0255(86)90088-6.

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

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

Donald B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366): 318–328, 1979. ISSN 01621459, 1537274X.

Donald B. Rubin. Causal Inference Without Counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450):435–438, 2000. ISSN 0162-1459. doi: 10.2307/2669382. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Cyrus Samii, Ye Wang, and Junlong Aaron Zhou. Generalizing trimming bounds for endogenously missing outcome data using random forests. *Political Analysis*, page 1–15, 2025. doi: 10.1017/pan.2025.10001.

Vira Semenova. Generalized lee bounds. *arXiv preprint arXiv:2008.12720*, 2020.

Lewis B. Sheiner and Donald B. Rubin. Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology & Therapeutics*, 57(1):6–15, 1995. ISSN 1532-6535. doi: 10.1016/0009-9236(95)90260-0. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/0009-9236%2895%2990260-0.

Jonathan M. Snowden, Sherri Rose, and Kathleen M. Mortimer. Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique. *American Journal of Epidemiology*, 173(7):731–738, April 2011. ISSN 0002-9262. doi: 10.1093/aje/kwq472.

Sonja A Swanson, Øyvind Holme, Magnus Løberg, Mette Kalager, Michael Bretthauer, Geir Hoff, Eline Aas, and Miguel A Hernán. Bounding the per-protocol effect in randomized trials: an application to colorectal cancer screening. *Trials*, 16(1):541, 2015.

Tymon Słoczyński. Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights. *The Review of Economics and Statistics*, pages 1–9, 2022. ISSN 0034-6535. doi: 10.1162/rest_a_00953.

Eric J Tchetgen Tchetgen. Identification and estimation of survivor average causal effects. *Statistics in Medicine*, 33(21):3601–3628, September 2014. ISSN 0277-6715. doi: 10.1002/sim.6181.

Kirk P Vanacore, Ashish Gurung, Adam C Sales, and Neil T Heffernan. Effect of gamification on gamers: Evaluating interventions for students who game the system. *Journal of educational data mining*, 2024.

Junni L Zhang and Donald B Rubin. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368, 2003.

Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1):20160010, 2017.