Causal Mediation in Natural Experiments

Senan Hogan-Hennessy*
Economics Department, Cornell University†

First draft: 12 February 2025
This version: 2 November 2025

Working Paper, newest version available here.

Abstract

Natural experiments are a cornerstone of applied economics, providing settings for estimating causal effects with a compelling argument for treatment randomisation, but give little indication of the mechanisms behind causal effects. Causal Mediation (CM) is a framework for sufficiently identifying a mechanism behind the treatment effect, decomposing it into an indirect effect channel through a mediator mechanism and a remaining direct effect. By contrast, a suggestive analysis of mechanisms gives necessary but not sufficient evidence. Conventional CM methods require that the relevant mediator mechanism is as-good-as-randomly assigned; when people choose the mediator based on costs and benefits (whether to visit a doctor, to attend university, etc.), this assumption fails and conventional CM analyses are at risk of bias. I propose an alternative strategy that delivers unbiased estimates of CM effects despite unobserved selection, using instrumental variation in mediator take-up costs. The method identifies CM effects via the marginal effect of the mediator, with parametric or semi-parametric estimation that is simple to implement in two stages. Applying these methods to the Oregon Health Insurance Experiment reveals a substantial portion of the Medicaid lottery's effect on subjective health and well-being flows through increased healthcare usage — an effect that a conventional CM analysis would mistake. This approach gives applied researchers an alternative method to estimate CM effects when an initial treatment is quasi-randomly assigned, but a mediator mechanism is not, as is common in natural experiments.

Keywords: Direct/indirect effects, quasi-experiment, selection, MTEs.

JEL Codes: C21, C31.

[*] For helpful comments I thank Neil Cholli, Hyewon Kim, Jiwoo Kim, Lukáš Lafférs, Jiwon Lee, Douglas Miller, Zhuan Pei, Brenda Prallon, and Evan Riehl. I thank seminar participants at Cornell University (2025), the European Economic Association Congress at Bordeaux Sciences Économiques (2025), the Econometrics World Congress at Seoul National University (2025), and the Causal Data Science Meeting (2025) for helpful discussion. Some preliminary results previously circulated in an earlier version of the working paper "The Direct and Indirect Effects of Genetics and Education." Any comments or suggestions may be sent to me at `seh325@cornell.edu`, or raised as an issue on the Github project, `https://github.com/shoganhennessy/mediation-natural-experiment`.

[†] Address: Uris Hall #447, Economics Department, Cornell University NY 14853 USA.

Economists use natural experiments to credibly answer social questions, when an experiment was infeasible. For example, does winning access to health insurance causally improve health and well-being (Finkelstein, Taubman, Wright, Bernstein, Gruber, Newhouse, Allen, Baicker & Group 2012)? Natural experiments are settings which answer these questions, but give little indication of how these effects came about. Causal Mediation (CM) aims to estimate the mechanisms behind causal effects, by estimating how much of the treatment effect operates through a proposed mediator mechanism. For example, do causal gains from winning access to health insurance come mostly from physical use of healthcare, or plausible psychological gains from no longer having to worry about being uninsured? This study of mechanisms behind causal effects broadens the economic understanding of social settings studied with natural experiments. This paper shows that the conventional approach to estimating CM effects is inappropriate in a natural experiment setting, provides a theoretical framework for how bias operates, and develops an approach to correctly estimate CM effects under alternative assumptions. These methods contrast the current practice in applied economics of providing suggestive evidence of mechanisms, which gives necessary but not sufficient conditions for the mechanisms behind a treatment effect.

In other disciplines — particularly epidemiology, psychology, and medicine — CM has become a standard empirical framework for decomposing causal effects into direct and indirect components (Imai, Keele & Yamamoto 2010). I refer to this established approach as "conventional CM," which assumes both the treatment and the mediator mechanism are (quasi-)randomly assigned. Applied economics has not typically adopted conventional CM, because mediators in economic applications — such as schooling, labour supply, or healthcare use — are choice variables decided by individuals' costs and benefits, likely violating this assumption. Nevertheless, economists often pursue the same goal informally: they present descriptive and suggestive evidence on plausible mediating mechanisms, or occasionally test whether a treatment effect remains after controlling for a plausible mediator mechanism. The first approach gives necessary not sufficient evidence on the mediating mechanisms, and the second is a conventional CM analysis — despite rarely being named as such by economists.

This paper starts by considering how conventional CM methods (whether applied explicitly or implicitly) perform in a natural experiment setting. Conventional CM methods rely on assuming the initial treatment, and the subsequent mediator mechanism, are both quasi-randomly assigned (Imai et al. 2010). Assuming the mediator is as-good-as-randomly assigned requires either (1) selection is fully captured by observed control variables, or (2) decisions are effectively random. While these assumptions can be plausible when a mediator is directly manipulated, or the data include extremely rich control variables, they are not credible in most economic applications; when take-up decisions reflect unobserved costs and benefits, mediator assignment is likely not random. For example, in the Oregon Health Insurance Experiment, those who got off the wait-list were randomly chosen by a lottery, but made the choice to visit healthcare in the following year of their own free will; this choice considered their own individual costs and benefits, and thus was not a random choice. I formally derive the selection bias that arises from this non-random assignment, and show through simulations that these biases can be large in settings consistent with standard economic models of selection. This result explains why conventional CM methods can yield misleading conclusions in applied economics, and motivates the need for a framework consistent with quasi-experimental causal reasoning.

Conventional CM's identifying assumptions are at odds with selection based on costs and benefits, so I import methods grounded in labour economic theory to solve the identification problem. This approach identifies CM effects via the Marginal Treatment Effect (MTE) of the mediator, and requires three main assumptions. (1) Mediator take-up must respond only positively to the initial treatment (monotonicity), which implies mediator selection follows a selection model. (2) Mediator take-up is motivated by mediator benefits. (3) A valid instrument for mediator take-up must exist, to avoid relying on parametric assumptions on unobserved selection. While these assumptions are strong, they are plausible in many applied settings. Mediator monotonicity aligns with conventional theories for selection-into-treatment, and is accepted widely in many applications using an instrumental variables research design. Selection based on costs and benefits is central to economic theory, and is the dominant concern for judging observational designs that identify causal effects. Access to valid instrumental variation is a strong condition, though is important to avoid further modelling assumptions; the most compelling example is using variation in mediator take-up costs as an instrument.

Applying the new methods to the Oregon Health Insurance Experiment shows that unobserved selection matters in an analysis of a real-world natural experiment. A substantial portion of the wait-list lottery's impact on subjective health and well-being is mediated indirectly through extra healthcare usage, after instrumenting for healthcare usage with respondents' usual provider. A conventional CM analysis would put this indirect mediated share at practically zero, so that my methods expose that negative selection into healthcare usage would be hiding evidence for this mechanism. These estimates give sufficient evidence that extra healthcare use mediates a sizeable share of the Medicaid-lottery's benefits, though wider confidence intervals underscore the inherent uncertainty on the proportion of the effect operating through the single mediator mechanism of healthcare usage.

The methods I propose for CM analyses are not perfect for every setting: the structural assumptions are strong, and are tailored to selection-into-mediator based on the economic principle of selection based on costs and benefits. Indeed, this approach provides no safe harbour for estimating CM effects if these structural assumptions do not hold true. This approach imports insights from the instrumental variables literature, connecting the conventional Imai et al. (2010) approach to CM with the economics literature on selection-into-treatment and MTEs (Vytlacil 2002, Heckman & Navarro-Lozano 2004, Heckman & Vytlacil 2005, Florens, Heckman, Meghir & Vytlacil 2008, Brinch, Mogstad & Wiswall 2017, Kline & Walters 2019).

Applied economists mostly investigate mediating mechanisms for causal effects with suggestive analyses of mechanisms. These descriptive analyses are informative, but do not generally identify a mediating pathway without additional assumptions — see also Blackwell, Ma & Opacic (2024), Green, Ha & Bullock (2010). A new strand of the econometric literature has arisen in implicit acknowledgement that suggestive evidence of mechanisms, or a conventional approach to CM, can lead to biased inference and needs alternative methods for credible inference. These include identifying CM effects with overlapping quasi-experimental research designs (Deuchert, Huber & Schelker 2019, Frölich & Huber 2017), functional form restrictions (Heckman & Pinto 2015, Heckman, Pinto & Savelyev 2013), partial identification (Flores & Flores-Lagunes 2009), or an hypothesis test of full mediation through observed channels (Kwon & Roth 2024) — see Huber (2020) for an overview.[1]

I develop a framework showing exactly how selection bias contaminates conventional CM estimates when mediator choices are driven by unobserved gains — settings where none of the existing econometric approaches to CM directly apply. The recent econometric literature has made important progress in adapting CM to observational settings; my work builds on this literature by addressing the common case in which the mediator is not quasi-randomly assigned, when selection into a mediator mechanism aligns with economic theory for selection. Frölich & Huber (2017) is the most similar paper to mine, though I extend the framework for CM from an economic perspective in multiple ways. First, by linking CM to the MTE literature for a binary mediator, I identify average CM effects rather than complier-specific effects (and provide a different estimation approach). Second, I show formally how unobserved selection biases conventional CM estimates and propose an alternative identification strategy that retains the natural experiment structure while relaxing mediator ignorability. Last, this paper provides a rigorous warning to applied economists against uncritically importing conventional CM methods to avoid selection biases term derived, and clarifies the conditions and practices to avoid them.

This paper proceeds as follows. Section 1 describes the dominant approach in economics for

[1] An alternative method to estimate CM effects is ensuring treatment and mediator quasi-random assignment holds by a running two randomised controlled trials for both treatment and mediator, at the same time. This set-up has been considered in the literature previously, in theory (Imai, Tingley & Yamamoto 2013) and in practice (Ludwig, Kling & Mullainathan 2011).

studying mechanisms behind treatment effects, illustrating with data from the Oregon Health Insurance Experiment. Section 2 introduces the formal framework for CM, and develops expressions for bias in conventional CM estimates in natural experiment and observational settings. Section 3 describes this bias in applied settings with (1) a regression framework, (2) a setting with selection based on costs and benefits. Section 4 purges bias from CM estimates by identifying CM effects via an MTE approach. Section 5 demonstrates how to estimate CM effects with this approach, with either parametric or semi-parametric methods, and gives simulation evidence. Section 6 returns to the Oregon Health Insurance Experiment, providing credible estimates of effects on subjective health and well-being mediated through increased healthcare usage. Section 7 concludes.

1 Mechanisms in the Oregon Health Insurance Experiment

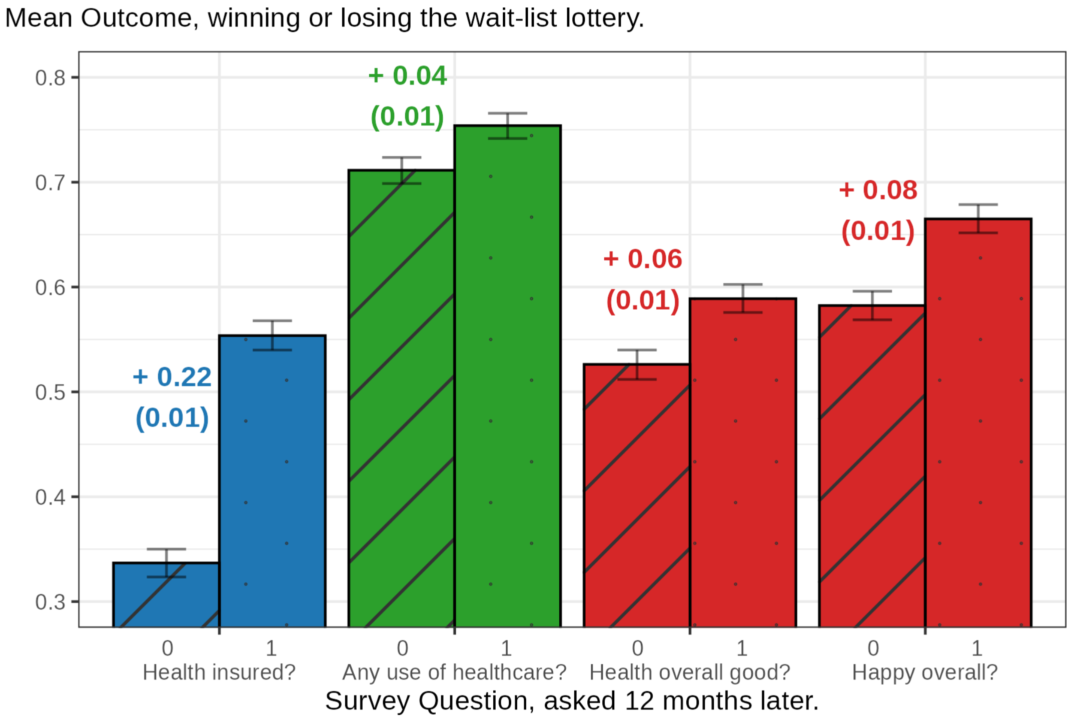
In the United States, healthcare is generally not provided directly by the government. Instead, consumers purchase health insurance to fund healthcare expenses, with the government providing insurance only for elderly individuals (Medicare) and for those with low-incomes (Medicaid). In 2004, the state of Oregon ceased accepting new applications for Medicaid due to budgetary constraints, and did not reopen applications until 2008. When the state resumed enrolment, 90,000 individuals applied, vastly exceeding the programme's capacity. Oregon therefore allocated the opportunity to apply for Medicaid via a lottery system among those on the wait-list. Winning this wait-list lottery significantly increased healthcare usage, plus subjective health and well-being.

Winning the wait-list lottery increased the average health insurance coverage rate by 22 percentage points (pp), and subjective visitation of any healthcare provider in the following 12 months by 4 pp. In addition, wait-list lottery winners agreed 6 pp more with the question "In general, would you say your health is excellent, very good, or good" (hereafter, subjective health), and 8 pp for "How would you say things are these days-would you say that you are very or pretty happy" (hereafter, subjective well-being). These numbers are calculated among the 9,957 people eligible for the Oregon wait-list lottery who responded to a survey sent by Finkelstein et al. (2012).[2] Figure 1 summarises these results.

These results show that winning the wait-list lottery led to large gains in subjective health and well-being. The economics, medicine, and health policy literatures have primarily

[2] This number restricts to those who gave non-missing answers to all relevant questions, including questions on pre-lottery location of usual healthcare, using anonymised data from the Oregon Health Insurance Experiment replication package (Finkelstein & Baicker 2014).

Figure 1: Effects of the Oregon Health Insurance Experiment Wait-list Lottery.



Note: This figure summarises the relevant results of the Oregon Health Insurance Experiment (Finkelstein et al. 2012). $\mathbb{E}\left[Y_i \mid Z_i = z'\right]$ is the mean outcome, where $z' = 0$ refers to the case of losing the wait-list lottery (not given access to Medicaid) and $z' = 1$ winning. The numbers beside the bars are estimates of the mean difference; winning the Medicaid wait-list lottery increased average health insurance rate by 22 percentage points (pp), with standard errors of the difference reported in brackets.

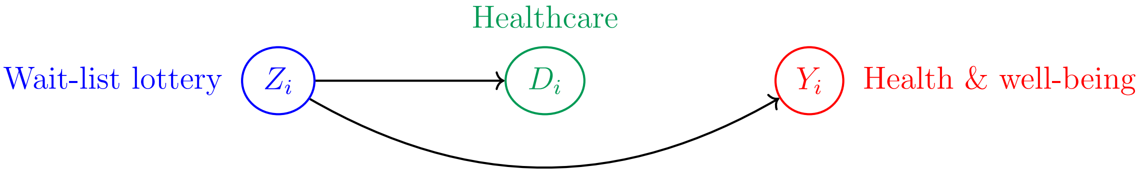
focused on the health benefits — often interpreted as healthcare benefits from new access to government provided health insurance.[3] However, the original authors also noted other benefits, including complete elimination of catastrophic out-of-pocket medical debt among those with new access to Medicaid. These are plausibly income effects that benefit recipients directly, not only through increased use of healthcare, but also by reducing stress and improving financial security. These plausible direct effects have not been explored in the applied literature.

Accepted practice in applied economics is to investigate mechanisms behind causal effects with suggestive evidence. This involves estimating the average causal effect of the wait-lottery on a proposed mediator (healthcare usage) and separately estimating its effect on the final outcomes (subjective health and well-being). When both estimates are positive, and the mediator precedes the outcome, it is taken as de facto evidence that the mediator transmits

[3] Finkelstein et al. (2012) use the wait-list lottery as an IV because health insurance is not randomly assigned; this paper focuses on the average effects of winning the wait-list lottery (which is randomly assigned).

the treatment effect. In the case of the Oregon Health Insurance Experiment, this amounts to concluding that increased healthcare usage mediates the positive effects of winning the lottery on health and well-being. Figure 2 illustrates this approach, which is also prevalent in other social science fields — see Blackwell et al. (2024), Green et al. (2010).

Figure 2: Structural Causal Model for Suggestive Evidence of a Mechanism.



Note: This figure shows the structural causal model behind a suggestive analysis for effects of the Oregon Health Insurance Experiment, where arrows represent causal effects — e.g., $Z_i \to D_i$ means Z_i affects D_i with no reverse causality.

This approach gives necessary, but not sufficient, identification of healthcare as a mediating mechanism. It is not sufficient because it provides no evidence for the effect of healthcare on health and well-being, so does not identify the causal mechanism. Studying this mechanism with suggestive evidence require an additional, hidden assumption that healthcare positively affects health outcomes. While this assumption is not unreasonable in general, it may not apply within the data under study; subjective well-being is measured only 12 months after the Medicaid lottery, which may not be enough time for health gains to accrue. Second, this approach does not quantify the mechanism effects. Healthcare could only have a very large effect on subjective health and well-being, or possibly a very large small effect — it is a priori unclear. In addition, the mediator mechanism effect refers not to the average effect of healthcare usage, but to the effect for Oregon residents who were induced to use more healthcare after winning the wait-list lottery (mediator compliers). This local effect could differ substantially from a population average, and potentially mislead conclusions about the magnitude or generality of the mechanism. Together, these concerns mean that a suggestive analysis of healthcare as a mediating mechanism are not dispositive; without additional assumptions, they neither identify nor quantify the mediator mechanism channel.

CM offers a compelling alternative framework, explicitly defining the average direct and indirect effects and clear assumptions under which they are identified. Moreover, it delivers quantitative answers to the key question: how much of a treatment effect operates through a specific mediator mechanism? CM is widely used in fields such as epidemiology, psychology, and medicine where researchers regularly decompose treatment effects into component pathways. However, CM methods have not yet been examined from an economic perspective to assess their applicability in observational causal research, such as natural experiments.

2 Causal Mediation (CM)

CM decomposes causal effects into two channels, through a mediating mechanism (indirect effect) and through all other paths (direct effect). To develop notation, write $Z_i = 0, 1$ for a binary treatment, $D_i = 0, 1$ a binary mediator mechanism, and Y_i a continuous outcome for individuals $i = 1, \dots, n$.[4] D_i and Y_i are a sum of their potential outcomes,

$$\begin{aligned} D_i &= (1 - Z_i) D_i(0) + Z_i D_i(1), \\ Y_i &= (1 - Z_i) Y_i(0, D_i(0)) + Z_i Y_i(1, D_i(1)). \end{aligned}$$

Assume treatment Z_i is quasi-randomly assigned,[5]

$$Z_i \perp\!\!\!\perp D_i(z'), Y_i(z, d'), \text{ for } z', z, d' = 0, 1.$$

There are only two average effects which are identified without additional assumptions.

1. The average first-stage refers to the effect of the treatment on mediator, Z_i on D_i:

$$\mathbb{E}\left[D_i \,|\, Z_i = 1\right] - \mathbb{E}\left[D_i \,|\, Z_i = 0\right] = \mathbb{E}\left[D_i(1) - D_i(0)\right].$$

   It is common in the economics literature to assume that Z_i influences D_i in at most one direction, $\Pr\left(D_i(0) \leq D_i(1)\right) = 1$ — monotonicity (Imbens & Angrist 1994). I assume mediator monotonicity (and its conditional variant) holds throughout to simplify notation.

2. The Average Treatment Effect (ATE) refers to the effect of the treatment on outcome, Z_i on Y_i, and is also known as the average total effect or intent-to-treat effect in social science settings, or reduced-form effect in the instrumental variables literature:

$$\mathbb{E}\left[Y_i \,|\, Z_i = 1\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0\right] = \mathbb{E}\left[Y_i(1, D_i(1)) - Y_i(0, D_i(0))\right].$$

Z_i affects outcome Y_i directly, and indirectly via the $D_i(Z_i)$ channel, with no reverse causality. Figure 3 visualises the design, where the direction arrows denote the causal

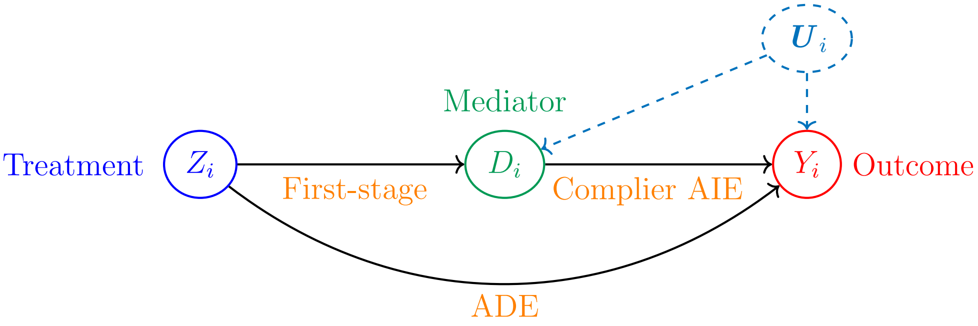
[4] This paper exclusively focuses on the binary case. See Huber, Hsu, Lee & Lettry (2020) or Frölich & Huber (2017) for a discussion of CM with continuous treatment and/or mediator, and the assumptions required.

[5] This assumption can hold conditional on a covariate vector, $\boldsymbol{X}_i$. To simplify notation in this section, leave the conditional part unsaid, as it changes no part of the identification framework.

direction. CM aims to decompose the ATE of Z_i on Y_i into these two separate pathways:

$$\begin{aligned} \text{Average Direct Effect (ADE):} \quad & \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right], \\ \text{Average Indirect Effect (AIE):} \quad & \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right]. \end{aligned}$$

Figure 3: Structural Causal Model for CM.



Note: This figure shows the structural causal model behind CM. The Complier AIE refers to the AIE local to $D_i(Z_i)$ compliers, so that AIE = Average First-stage × Complier AIE. $\boldsymbol{U}_i$ shows an unobserved confounding variable for the causal effect $D_i \to Y_i$, representing this paper's focus on the case that D_i is not quasi-randomly assigned. Subsection 3.1 defines $\boldsymbol{U}_i$ in an applied setting.

Estimating the AIE answers the following question: how much of the causal effect Z_i on Y_i goes through the D_i channel? When studying the health gains of winning the Medicaid wait-list lottery (Finkelstein et al. 2012), the AIE represents how much of the effect comes from using the hospital more often. Estimating the ADE answers the following equation: how much is left over after accounting for the D_i channel?[6] For the example, how much of the wait-list lottery effect is a direct effect, other than increased healthcare usage — e.g., income effects of lower medical debt, or less worry over health shocks thanks to government support. An Instrumental Variables (IV) approach assumes this direct effect is zero for everyone (the exclusion restriction). CM is a similar, yet distinct, framework attempting to explicitly model the direct effect, and not assuming it is zero.

The ADE and AIE are not separately identified without further assumptions.

2.1 Identification of CM Effects

The conventional approach to estimating direct and indirect effects assumes both Z_i and D_i are quasi-randomly assigned, conditional on a vector of control variables $\boldsymbol{X}_i$.

[6] In a non-parametric setting it is not necessary that ADE + AIE = ATE. See Imai et al. (2010) for this point in full.

Definition 1. *Sequential quasi-random assignment (Imai et al. 2010)*

$$Z_i \perp\!\!\!\perp D_i(z'), Y_i(z, d') \;\mid\; \boldsymbol{X}_i, \qquad \text{for } z', z, d' = 0, 1 \tag{1}$$

$$D_i \perp\!\!\!\perp Y_i(z', d') \;\mid\; \boldsymbol{X}_i, Z_i = z', \qquad \text{for } z', d' = 0, 1. \tag{2}$$

Sequential quasi-random assignment assumes that the initial treatment Z_i is quasi-randomly assigned conditional on $\boldsymbol{X}_i$ (as has already been assumed above). It then also assumes that, after Z_i is assigned, that D_i is quasi-randomly assigned conditional on $\boldsymbol{X}, Z_i$ (hereafter, mediator quasi-random assignment). If 1(1) and 1(2) hold, then the ADE and AIE are identified by two-stage mean differences conditioning on $\boldsymbol{X}_i$.[7]

$$\mathbb{E}_{D_i, \boldsymbol{X}_i}\left[\underbrace{\mathbb{E}\left[Y_i \mid Z_i = 1, D_i, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \mid Z_i = 0, D_i, \boldsymbol{X}_i\right]}_{\text{Second-stage regression, } Y_i \text{ on } Z_i \text{ holding } D_i, \boldsymbol{X}_i \text{ constant}}\right] = \underbrace{\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right]}_{\text{Average Direct Effect (ADE)}}$$

$$\mathbb{E}_{Z_i, \boldsymbol{X}_i}\left[\underbrace{\Big(\mathbb{E}\left[D_i \mid Z_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[D_i \mid Z_i = 0, \boldsymbol{X}_i\right]\Big)}_{\text{First-stage regression, } D_i \text{ on } Z_i} \times \underbrace{\Big(\mathbb{E}\left[Y_i \mid Z_i, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \mid Z_i, D_i = 0, \boldsymbol{X}_i\right]\Big)}_{\text{Second-stage regression, } Y_i \text{ on } D_i \text{ holding } Z_i, \boldsymbol{X}_i \text{ constant}}\right]$$
$$= \underbrace{\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right]}_{\text{Average Indirect Effect (AIE)}}$$

I refer to the estimands on the left-hand side as CM estimands, which are typically estimated by a composition of two-stage Ordinary Least Squares (OLS) estimates (Imai et al. 2010). While this is the most common approach in the applied literature, I do not assume the linear model for my identification analysis. Linearity assumptions are not necessary for identification, and it suffices to note that heterogeneous treatment effects and non-linear confounding can bias OLS estimates of CM estimands in the same manner that is well documented elsewhere (see e.g., Angrist 1998, Słoczyński 2022). This section focuses on problems that plague conventional CM, regardless of estimation method.

2.2 Non-identification of CM Effects

Applied research often uses a natural experiment to study settings where treatment Z_i is quasi-randomly assigned, justifying assumption 1(1). Rarely do they also have access to an additional, overlapping natural experiment to isolate random variation in D_i — to justify mediator quasi-random assignment 1(2). One might consider conventional CM methods in such a setting to learn about the mechanisms behind the causal effect Z_i on Y_i. This approach leads to estimates at risk of bias, contaminating inference on direct and indirect effects.

[7] In addition, a common support condition for both Z_i, D_i (across $\boldsymbol{X}_i$) is necessary. Imai et al. (2010) show a general identification statement; I show identification in terms of two-stage regression. See Appendix A.1.

Theorem 1. *Absent an identification strategy for the mediator, conventional CM estimates are at risk of selection bias. If 1(1) holds, and 1(2) does not, then conventional CM estimands are contaminated by selection bias and group differences. Proof: see Appendix A.2.*

Below I present the relevant selection bias and group difference terms, omitting the conditional on $\boldsymbol{X}_i$ notation for brevity.
For the direct effect: conventional CM estimand = ADE + selection bias + group differences.[8]

$$
\begin{aligned}
&\mathbb{E}_{D_i}\Big[\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i\right]\Big] \\
&\quad = \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] \\
&\qquad + \mathbb{E}_{D_i = d'}\Big[\mathbb{E}\left[Y_i(0, D_i(Z_i)) \,|\, D_i(1) = d'\right] - \mathbb{E}\left[Y_i(0, D_i(Z_i)) \,|\, D_i(0) = d'\right]\Big] \\
&\qquad + \mathbb{E}_{D_i = d'}\left[\Big(1 - \Pr\left(D_i(1) = d'\right)\Big)\begin{pmatrix}\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, D_i(1) = 1 - d'\right] \\ - \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, D_i(1) = d'\right]\end{pmatrix}\right]
\end{aligned}
$$

For the indirect effect: conventional CM estimand = AIE + selection bias + group differences.

$$
\begin{aligned}
&\mathbb{E}_{Z_i}\left[\Big(\mathbb{E}\left[D_i \,|\, Z_i = 1\right] - \mathbb{E}\left[D_i \,|\, Z_i = 0\right]\Big) \times \Big(\mathbb{E}\left[Y_i \,|\, Z_i, D_i = 1\right] - \mathbb{E}\left[Y_i \,|\, Z_i, D_i = 0\right]\Big)\right] \\
&= \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] \\
&\quad + \Pr\left(D_i(1) = 1, D_i(0) = 0\right)\Big(\mathbb{E}\left[Y_i(Z_i, 0) \,|\, D_i = 1\right] - \mathbb{E}\left[Y_i(Z_i, 0) \,|\, D_i = 0\right]\Big) \\
&\quad + \Pr\left(D_i(1) = 1, D_i(0) = 0\right) \times \\
&\quad\left[\begin{array}{l}\Big(1 - \Pr\left(D_i = 1\right)\Big)\begin{pmatrix}\mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i = 1\right] \\ - \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i = 0\right]\end{pmatrix} \\ - \left(\dfrac{1 - \Pr\left(D_i(1) = 1, D_i(0) = 0\right)}{\Pr\left(D_i(1) = 1, D_i(0) = 0\right)}\right)\begin{pmatrix}\mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i(1) = 0 \text{ or } D_i(0) = 1\right] \\ - \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0)\right]\end{pmatrix}\end{array}\right]
\end{aligned}
$$

The selection bias terms come from systematic differences between the groups taking or refusing the mediator ($D_i = 1$ versus $D_i = 0$), differences not fully unexplained by $\boldsymbol{X}_i$. These selection bias terms would equal zero if the mediator had been quasi-randomly assigned 1(2), but do not necessarily average to zero if not. In the Oregon Health Insurance Experiment, the wait-list gave random variation in the treatment (the Medicaid wait-list lottery) but there was not a similar natural experiment for healthcare usage; correspondingly,

[8] The bias terms here mirror those in Heckman, Ichimura, Smith & Todd (1998), Angrist & Pischke (2009) for a single D_i on Y_i treatment effect, when D_i is not quasi-randomly assigned:
$\mathbb{E}\left[Y_i \,|\, D_i = 1\right] - \mathbb{E}\left[Y_i \,|\, D_i = 0\right] = \text{ATE} + \underbrace{\Big(\mathbb{E}\left[Y_i(.,0) \,|\, D_i = 1\right] - \mathbb{E}\left[Y_i(.,0) \,|\, D_i = 0\right]\Big)}_{\text{Selection Bias}} + \underbrace{\Pr\left(D_i = 0\right)\left(\text{ATT} - \text{ATU}\right)}_{\text{Group-differences Bias}}.$

the selection-on-observables approach to CM has selection bias.

The group differences represent the fact that a matching approach gives an average effect on the treated group, which is systematically different from the average effect if selection-on-observables does not hold. These terms are a non-parametric framing of the bias from controlling for intermediate outcomes, previously studied only in a linear setting (i.e., bad controls in Cinelli, Forney & Pearl 2024, or M-bias in Ding & Miratrix 2015).

The AIE group differences term is longer, because the indirect effect is comprised of the effect of D_i local to $D_i(Z_i)$ compliers.

$$\begin{aligned}\text{AIE} &= \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] \\ &= \mathbb{E}\left[D_i(1) - D_i(0)\right] \underbrace{\mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i(0) = 0, D_i(1) = 1\right]}_{\text{Average } D_i \text{ on } Y_i \text{ effect among } D_i(Z_i) \text{ compliers}}\end{aligned}$$

It is important to acknowledge the mediator compliers here, because the AIE is the treatment effect going through the $D_i(Z_i)$ channel, thus only refers to individuals pushed into mediator D_i by initial treatment Z_i. If we had been using a population average effect for D_i on Y_i, then this is losing focus on the definition of the AIE; it is not about the causal effect D_i on Y_i, it is about the causal effect $D_i(Z_i)$ on Y_i.

The group difference bias term arises because the convention approach to CM assumes the complier average effect is equal to the population average effect, which does not hold true if the mediator is not quasi-randomly assigned.

3 CM in Applied Settings

Unobserved confounding is particularly problematic when studying the mechanisms behind treatment effects. For example, in studying health gains from the Oregon wait-list lottery, we might expect that health gains came about because those who won access to Medicaid started visiting their healthcare provider more often, when in past they avoided it over financial concerns. Applying conventional CM methods to investigate this expectation would be dismissing unobserved confounders for how often individuals visit healthcare providers, leading to biased results.

The wider population does not have one uniform bill of health; many people are born predisposed to ailments, due to genetic variation or other unrelated factors. These conditions can exist for years before being diagnosed. People with severe underlying conditions may visit healthcare providers more often than the rest of the population, to investigate or begin treating the ill-effects. It stands to reason that people with more serve underlying conditions may gain more from more often attending healthcare providers once given health insurance.

These underlying causes cannot be controlled for by researchers, as we cannot hope to observe and control for health conditions that are yet to even be diagnosed. This means underlying health conditions are an unobserved confounder, and will bias estimates of the ADE and AIE in this setting.

In this section, I further develop the issue of selection on unobserved factors in a general CM setting. First, I show the non-parametric bias terms from Section 2 can be written as omitted variables bias in a random coefficients regression framework. Second, I show how selection bias operates in a basic model for selection-into-mediator based on costs and benefits.

3.1 Regression Framework

Inference for CM effects can be written in a regression framework with random coefficients, showing how correlation between unobserved error terms and the mediator disrupts identification.

Start by writing potential outcomes $Y_i(.,.)$ as a sum of observed and unobserved factors, following the notation of Heckman & Vytlacil (2005). For each $z', d' = 0, 1$, put $\mu_{d'}(z'; \boldsymbol{X}_i) = \mathbb{E}\left[Y_i(z', d') \,|\, \boldsymbol{X}_i\right]$ and the corresponding error terms, $U_{d',i} = Y_i(z', d') - \mu_{d'}(z'; \boldsymbol{X}_i)$, so we have the following expressions:

$$Y_i(Z_i, 0) = \mu_0(Z_i; \boldsymbol{X}_i) + U_{0,i}, \quad Y_i(Z_i, 1) = \mu_1(Z_i; \boldsymbol{X}_i) + U_{1,i}.$$

With this notation, observed data $Z_i, D_i, Y_i, \boldsymbol{X}_i$ have the following random coefficient outcome formulae — which characterise direct effects, indirect effects, and selection bias.

$$D_i = \theta + \overline{\pi} Z_i + \zeta(\boldsymbol{X}_i) + \eta_i, \tag{3}$$

$$Y_i = \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) + \underbrace{(1 - D_i)\, U_{0,i} + D_i U_{1,i}}_{\text{Correlated error term.}} \tag{4}$$

This is not consequence of linearity assumptions; the outcome formulae allow for unconstrained heterogeneous treatment effects, because the coefficients are random. If either Z_i, D_i were continuously distributed, then this representative would not necessarily hold true. First-stage (3) is identified, with $\theta + \zeta(\boldsymbol{X}_i)$ the intercept, and $\overline{\pi}$ the first-stage average compliance rate (conditional on $\boldsymbol{X}_i$). Second-stage (4) has the following definitions, and is not identified thanks to omitted variables bias. See Appendix A.3 for the derivation.

(a) $\alpha = \mathbb{E}\left[\mu_0(0; \boldsymbol{X}_i)\right]$ and $\varphi(\boldsymbol{X}_i) = \mu_0(0; \boldsymbol{X}_i) - \alpha$ are the intercept terms.

(b) $\beta = \mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)$ is the AIE conditional on $Z_i = 0, \boldsymbol{X}_i$.

(c) $\gamma = \mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)$ is the ADE conditional on $D_i = 0, \boldsymbol{X}_i$.

(d) $\delta = \mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - \big(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\big)$ is the average interaction effect conditional on $\boldsymbol{X}_i$.

(e) $(1 - D_i)\, U_{0,i} + D_i U_{1,i}$ is the disruptive error term.

The ADE and AIE are averages of the random coefficients:

$$\begin{aligned} \text{ADE} &= \mathbb{E}\left[\gamma + \delta D_i\right], \\ \text{AIE} &= \mathbb{E}\left[\overline{\pi}\big(\beta + \delta Z_i + \widetilde{U}_i\big)\right], \qquad \text{with } \widetilde{U}_i = \underbrace{\mathbb{E}\left[U_{1,i} - U_{0,i} \,\middle|\, \boldsymbol{X}_i, D_i(0) = 0, D_i(1) = 1\right]}_{\text{Unobserved complier gains.}}. \end{aligned}$$

The ADE is a simple sum of the coefficients, while the AIE includes a group differences term because it only refers to $D_i(Z_i)$ compliers.

By construction, $\boldsymbol{U}_i \coloneqq (U_{0,i}, U_{1,i})$ is an unobserved confounder. The regression estimates of β, γ, δ in second-stage (4) give unbiased estimates only if D_i is also conditionally quasi-randomly assigned: $D_i \perp\!\!\!\perp \boldsymbol{U}_i$. If not, then estimates of CM effects suffer from omitted variables bias from failing to adjust for the unobserved confounder, $\boldsymbol{U}_i$.

3.2 Selection on Costs and Benefits

CM is at risk of bias because $D_i \perp\!\!\!\perp \boldsymbol{U}_i$ is unlikely to hold in applied settings. A separate identification strategy could disrupt the selection-into-D_i based on unobserved factors, and lend credibility to the mediator quasi-random assignment assumption. Without it, bias will persist, given how we conventionally think of selection-into-treatment.

Consider a model where individual i selects into a mediator based on costs and benefits (in terms of outcome Y_i), after $Z_i, \boldsymbol{X}_i$ have been assigned. In a natural experiment setting, an external factor has disrupted individuals selecting Z_i by choice (thus Z_i is quasi-randomly assigned), but it has not disrupted the choice to take mediator (thus D_i is not quasi-randomly assigned). In the Oregon Health Insurance Experiment, the treatment variation comes from the wait-list lottery, while healthcare usage was not subject to a similar lottery. Write C_i for individual i's costs of taking mediator D_i, and $\mathbb{1}\{.\}$ for the indicator function. The Roy model has i taking the mediator if the benefits exceed the costs,

$$D_i\left(z'\right) = \mathbb{1}\left\{ \underbrace{C_i}_{\text{Costs}} \;\le\; \underbrace{Y_i\left(z', 1\right) - Y_i\left(z', 0\right)}_{\text{Benefits}} \right\}, \quad \text{for } z' = 0, 1. \tag{5}$$

The Roy model provides an intuitive framework for analysing selection mechanisms because it captures the fundamental economic principle of decision-making based on costs and benefits in terms of the outcome under study (Roy 1951, Heckman & Honore 1990). In the Oregon Health Insurance Experiment, this models choice to visit the doctor in terms of health and well-being benefits relative to costs.[9] This makes the Roy model useful as a base case for CM, where selection-into-mediator may be driven by private information (unobserved by the researcher).

By using the Roy model as a benchmark, I explore the practical limits of the mediator quasi-random assignment assumption. If selection follows a Roy model, and the mediator is quasi-randomly assigned, then unobserved benefits can play no part in selection. The only driver of selection are individuals' differences in costs (and not benefits). If there are any selection-into-D_i benefits unobserved to the researcher, then mediator quasi-random assignment cannot hold.

Proposition 1. *Suppose mediator selection follows a Roy model (5), and selection is not fully explained by costs and observed gains. Then mediator quasi-random assignment does not hold.*

This is an equivalence statement: selection based on costs and benefits is only consistent with mediator quasi-random assignment if the researcher observed every single source of mediator benefits. See Appendix A.4 for the proof. This means than the vector of control variables $\boldsymbol{X}_i$ must be incredibly rich. Together, $\boldsymbol{X}_i$ and unobserved cost differences $U_{C,i}$ must explain selection-into-D_i one hundred percent. In the Roy model framework, however, individuals make decisions about mediator take-up based on gains — whether the researcher observes them or not. The unobserved gains are unlikely to be fully captured by an observed control set $\boldsymbol{X}_i$, except in very special cases.

In practice, the best setting to believe in the mediator quasi-random assignment assumption is to study a setting where the researcher has two causal research designs, one for treatment Z_i and another for mediator D_i, at the same time. Absent this, mediator quasi-random assignment become hard to believe, and the corresponding conventional CM estimates are at risk of selection bias.

4 Solving Identification via the Mediator MTE

If your goal is to estimate CM effects, and you could control for unobserved selection terms $U_{0,i}, U_{1,i}$, then you would. This ideal (but infeasible) scenario would yield unbiased estimates

[9] If the choice is considered over a sum of outcomes, then a simple extension to a utility maximisation model maintains this same framework with expected costs and benefits. See Heckman & Honore (1990), Eisenhauer, Heckman & Vytlacil (2015).

for the ADE and AIE. Identification via the mediator Marginal Treatment Effect (MTE) takes this insight seriously, providing conditions to model the implied confounding by $U_{0,i}, U_{1,i}$, and then controlling for it.

The main problem is that second-stage regression equation (4) is not identified, because $U_{0,i}, U_{1,i}$ are unobserved, and lead to omitted variables bias.

$$\begin{aligned}\mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i\right] \;&=\; \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\ &\quad + \underbrace{(1-D_i)\,\mathbb{E}\left[U_{0,i} \,|\, D_i = 0, \boldsymbol{X}_i\right] + D_i \mathbb{E}\left[U_{1,i} \,|\, D_i = 1, \boldsymbol{X}_i\right]}_{\text{Unobserved confounding.}} \end{aligned} \tag{6}$$

My approach to identifying CM effects models the contaminating terms in (6) via the mediator MTE, avoiding the bias terms derived in Section 2. Thinking on MTEs began with corrections for sample selection problems (Heckman 1974), and were extended to a general selection problem of the same form as Equation (6) in parametric settings (Heckman 1979, Björklund & Moffitt 1987) and a general case (Heckman & Vytlacil 2005). The approach works in the following manner: (1) assume that the variable of interest follows a selection model, where unexplained first-stage selection informs unobserved second-stage confounding; (2) extract information about unobserved confounding from the first-stage; and (3) incorporate this information as control terms in the second-stage equation to adjust for selection-into-mediator. Identification in MTE methods typically relies on identifying the corresponding Control Functions (CFs) with an external instrument or distributional assumptions; this paper focuses exclusively on the case that an instrument is available. By explicitly accounting for the information contained in the first-stage selection model, the MTE approach enables consistent estimation of causal effects in the second-stage even when selection is driven by unobserved factors.

In the example of analysing health gains from the Oregon Health Insurance Experiment, the MTE-based approach addresses the unobserved confounding by modelling unobserved effects of underlying health conditions. It does so by assuming that unobserved selection-into-healthcare use is informative for underlying health conditions, assuming people with more severe underlying conditions visit the doctor more often than those without. Then it uses this information in the second-stage estimation of how much the effect goes through increased healthcare usage, estimating the ADE and AIE after controlling for this confounding.

4.1 Re-identification of CM Effects

The following assumptions are sufficient to model the correlated error terms, identifying β, γ, δ in the second-stage regression (4), and thus both the ADE and AIE.

Assumption MTE–1. Mediator monotonicity, conditional on $\boldsymbol{X}_i$.

$$\Pr\left(D_i(0) \leq D_i(1) \,|\, \boldsymbol{X}_i\right) = 1.$$

Assumption MTE–1 is the monotonicity condition first used in an IV context (Imbens & Angrist 1994). Here, it is assuming that people respond to treatment, Z_i, by consistently taking or refusing the mediator D_i (always or never-mediators), or taking the mediator D_i if and only if assigned to the treatment $Z_i = 1$ (mediator compliers). There are no mediator defiers.

The main implication of Assumption MTE–1 is that selection-into-mediator can be written as a selection model with ordered threshold crossing values that describe selection-into-D_i (Vytlacil 2002).

$$D_i(z') = \mathbb{1}\left\{V_i \leq \psi\big(z'; \boldsymbol{X}_i\big)\right\}, \quad \text{for } z' = 0, 1$$

where V_i is a latent variable with continuous distribution and conditional cumulative density function $F_V(\,.\,|\boldsymbol{X}_i)$, and $\psi(\,.\,; \boldsymbol{X}_i)$ collects observed sources of mediator selection. V_i could be assumed to follow a known distribution; the canonical Heckman selection model assumes V_i is normally distributed (a "Heckit" model). The identification strategy here applies to the general case that the distribution of V_i is unknown, without parametric restrictions.

I focus on the equivalent transformed model of Heckman & Vytlacil (2005),

$$D_i(z') = \mathbb{1}\left\{U_i \leq \pi(z'; \boldsymbol{X}_i)\right\}, \quad \text{for } z' = 0, 1$$

where $U_i \coloneqq F_V\left(V_i \mid \boldsymbol{X}_i\right)$ follows a uniform distribution, and $\pi(z'; \boldsymbol{X}_i) = F_V\big(\psi(z'; \boldsymbol{X}_i)\big) = \Pr\left(D_i = 1 \,|\, Z_i = z', \boldsymbol{X}_i\right)$ is the mediator propensity score. U_i are the unobserved mediator take-up costs. Note the maintained assumption that treatment Z_i is quasi-randomly assigned conditional on $\boldsymbol{X}_i$ implies $Z_i \perp\!\!\!\perp U_i$ conditional on $\boldsymbol{X}_i$.

This selection model setup is equivalent to the monotonicity condition, and is importing a well-known equivalence result from the IV literature to the CM setting. The main conceptual difference is not assuming Z_i is a valid instrument for identifying the D_i on Y_i effect among compliers; it is using the selection model representation to correct for selection bias. See Appendix A.5 for a validation of the general Vytlacil (2002) equivalence result in a CM setting, with conditioning covariates $\boldsymbol{X}_i$.

Assumption MTE–2. Selection on mediator benefits.

$$\text{Cov}\left(U_i,\, U_{0,i}\right), \;\text{Cov}\left(U_i,\, U_{1,i}\right) \neq 0.$$

Assumption MTE–2 is stating that unobserved selection in mediator take-up (U_i) informs second-stage confounding, when refusing or taking the mediator ($U_{0,i}$ and $U_{1,i}$). If there is unobserved confounding in Y_i, then it can be measured in D_i.

This is a relevance assumption for the MTE model. If people had been deciding to take D_i by a Roy model, then this assumption holds because $V_i = U_{C,i} - \big(U_{1,i} - U_{0,i}\big)$. Individuals could be making decisions based on other outcomes, but as long as mediator benefits guide at least part of this decision (i.e., bounded away from zero), then this assumption will hold.

For notation purposes, suppose the vector of control variables $\boldsymbol{X}_i$ has at least two entries; denote $\boldsymbol{X}_i^{\text{IV}}$ as one entry in the vector, and $\boldsymbol{X}_i^{-}$ as the remaining.

Assumption MTE–3. Mediator take-up cost instrument.

$$\boldsymbol{X}_i^{\text{IV}} \text{ satisfies } \frac{\partial}{\partial \boldsymbol{X}_i^{\text{IV}}}\Big\{\mu_1(z', \boldsymbol{X}_i) - \mu_0(z', \boldsymbol{X}_i)\Big\} = 0 < \frac{\partial}{\partial \boldsymbol{X}_i^{\text{IV}}}\Big\{\mathbb{E}\left[D_i(z') \,|\, \boldsymbol{X}_i\right]\Big\}, \text{ for } z' = 0, 1.$$

Assumption MTE–3 is requiring at least one control variable guides selection-into-D_i — an IV. It assumes an instrument exists, which satisfies an exclusion restriction (i.e., not impacting mediator gains $\mu_1 - \mu_0$), and has a non-zero influence on the mediator (i.e., strong IV first-stage). The exclusion restriction is untestable, and must be guided by domain-specific knowledge; IV first-stage strength is testable, and must be justified with data by methods common in the IV literature.

This assumption identifies the mediator propensity score separately from the direct and indirect effects, avoiding indeterminacy in the second-stage outcome equation. While not technically required for identification, it avoids relying entirely on an assumed distribution for unobserved error terms (and bias from inevitably breaking this assumption). The most compelling example of a mediator IV is using data on the cost of mediator take-up as a first-stage IV, if it varies between individuals for unrelated reasons and is strong in explaining mediator take-up.

Proposition 2. *If assumptions MTE–1, MTE–2, MTE–3 hold, then second-stage regression equation (4) is identified via the MTE-associated CFs.*

$$\begin{aligned}\mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i\right] \;&=\; \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi\big(\boldsymbol{X}_i^{-}\big) \\ &\quad + \rho_0 \left(1 - D_i\right) \lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big) + \rho_1 D_i \lambda_1\big(\pi(Z_i; \boldsymbol{X}_i)\big),\end{aligned}$$

where λ_0, λ_1 are the corresponding CFs, ρ_0, ρ_1 are linear parameters, and mediator propensity score $\pi(z'; \boldsymbol{X}_i)$ is separately identified in the first-stage (3). Proof: see Appendix A.6.

Again, this set-up required no linearity assumptions, and treatment effects vary, because Z_i, D_i are categorical and $\beta, \gamma, \delta, \varphi(\boldsymbol{X}_i)$ vary with $\boldsymbol{X}_i$. The CFs are functions which measure unobserved mediator gains, for those with unobserved mediator costs above or below a propensity score value. Following the IV notation of Kline & Walters (2019), put $\mu_V = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)\right]$, to give the following representation for the CFs:

$$\lambda_0\big(p'\big) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, p' < U_i\right],$$
$$\lambda_1\big(p'\big) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, U_i \leq p'\right] = -\lambda_0\big(p'\big)\left(\frac{1-p'}{p'}\right), \text{ for } p' \in (0,1).$$

All relevant parameters — $\alpha, \beta, \gamma, \delta, \varphi(.)$ — are identified once we control for selection bias through the CFs λ_0, λ_1, with $\pi(z'; \boldsymbol{X}_i)$ identified separately in the first-stage thanks to the instrument(s) $\boldsymbol{X}_i^{\mathrm{IV}}$. In the case that the CFs have an assumed functional form, then identification is complete. For example, in the canonical Heckman selection model, the error terms follow a normal distribution, so that λ_0, λ_1 are the inverse Mills ratio. If we do not know the distribution of $\big(U_{0,i}, U_{1,i}\big)$, then λ_0, λ_1 can be estimated separately with semi-parametric methods to avoid relying on parametric assumptions.[10]

This identification strategy is an MTE approach (Björklund & Moffitt 1987, Heckman & Vytlacil 2005) applied to a CM setting. One can see this by noting the connection to the marginal effect of the mediator,

$$\begin{aligned}&\mathbb{E}\left[Y_i(z',1) - Y_i(z',0) \,\middle|\, Z_i = z', \boldsymbol{X}_i, U_i = p'\right] \\ &= \beta + \delta z' + \underbrace{\mathbb{E}\left[U_{1,i} - U_{0,i} \,\middle|\, \boldsymbol{X}_i, U_i = p'\right]}_{=\rho_1\lambda_1(p') - \rho_0\lambda_0(p')}, \qquad \text{for } p' \in (0,1).\end{aligned}$$

The marginal effect of the mediator is identified under the MTE Approach assumptions, thanks to instrumental variation in $\boldsymbol{X}_i^{\mathrm{IV}}$. The final step uses the corresponding CFs to extrapolate from $\boldsymbol{X}_i^{\mathrm{IV}}$ compliers to mediator compliers, and thus identify the ADE and AIE.

Theorem MTE Approach. If assumptions MTE–1, MTE–2, MTE–3 hold, the ADE and AIE are identified as a function of the parameters in Proposition 2.

$$\text{ADE} = \mathbb{E}\left[\gamma + \delta D_i\right],$$
$$\text{AIE} = \mathbb{E}\left[\overline{\pi}\left(\beta + \delta Z_i + \underbrace{\left(\rho_1 - \rho_0\right)\Gamma\big(\pi(0; \boldsymbol{X}_i),\, \pi(1; \boldsymbol{X}_i)\big)}_{\text{Mediator compliers adjustment}}\right)\right]$$

where $\Gamma\left(p, p'\right) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, p < U_i \leq p'\right] = \frac{p'\lambda_1(p') - p\lambda_1(p)}{p' - p}$ is the average unobserved net gains for those with unobserved costs between $p < p'$,[11] and $\overline{\pi} = \pi(1; \boldsymbol{X}_i) - \pi(0; \boldsymbol{X}_i)$

[10] This comes at the cost of α and ρ_0, ρ_1 no longer being separately identified from λ_0, λ_1. However, this does not jeopardise identification and estimation of the ADE and AIE — see Subsection 5.2.

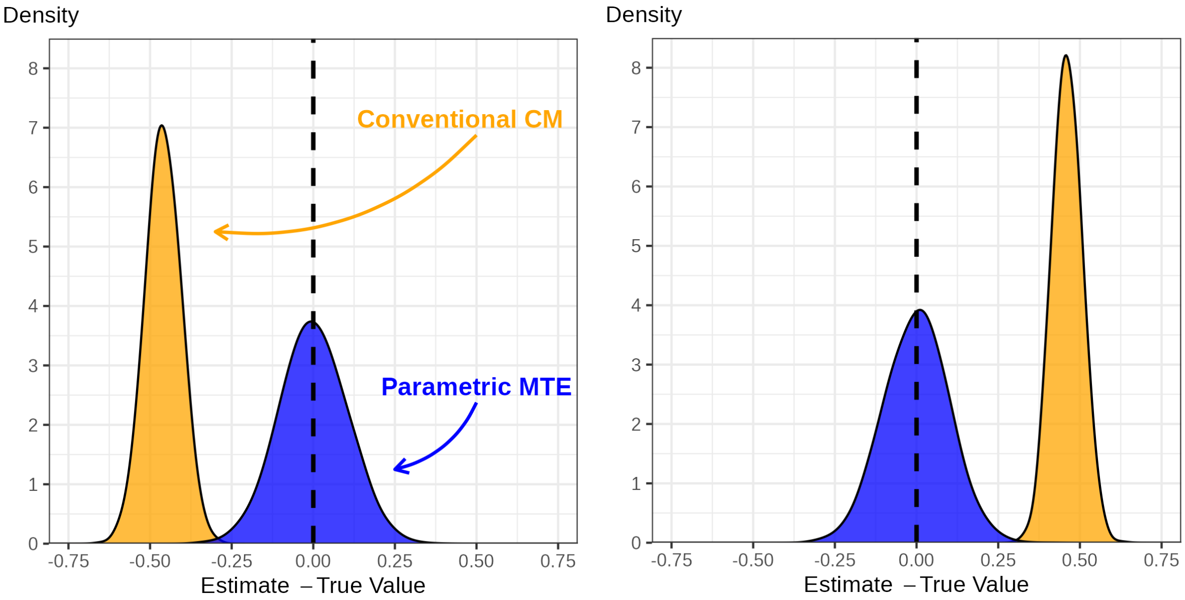
[11] The complier adjustment term was first written in this manner by Kline & Walters (2019) for an IV setting.

is the mediator complier score. Proof: see Appendix A.7.

This theorem provides a solution to the identification problem for CM effects when facing selection; rather than assuming away selection problems, it explicitly models them. The ADE is straightforward to calculate as an average of the direct effect parameters, while the AIE also includes an adjustment for unobserved complier gains to the mediator. Again, this is because the AIE only refers to individuals who were induced by treatment Z_i into taking mediator D_i (mediator compliers). The MTE approach measures both selection bias and complier differences, and thus purges these persistent bias terms to identify CM effects.

Figure 4: The MTE Approach Addresses Persistent Bias in Conventional CM Estimates.

(a) $\widehat{\text{ADE}} - \text{ADE}$. **(b)** $\widehat{\text{AIE}} - \text{AIE}$.



Note: These figures show the empirical density of point estimates minus the true average effect, for 10,000 different datasets generated from a Roy model with normally distributed error terms (with both correlation and heteroscedasticity, further described in Subsection 5.3). The black dashed line is the true value; orange is the distribution of conventional CM estimates from two-stage OLS (Imai et al. 2010), and blue estimates with a two-stage Heckman selection adjustment.

The ideal instrument $\boldsymbol{X}_i^{\text{IV}}$ for identification is continuous, and varies $\pi(z'; \boldsymbol{X}_i)$ between 0 and 1 for every possible value of $z', \boldsymbol{X}_i^-$ (identification at infinity). In practice, it is unlikely to find IV(s) that satisfy this condition. In this case, the Brinch et al. (2017) restricted approach can be used — even with a categorical instrument and no control variables. This amounts to assuming a limited specification for the respective CFs, limiting the number of parameters used to approximate λ_0, λ_1 to the number of discrete values that $\pi(z'; \boldsymbol{X}_i)$ takes minus one. E.g., if there are no control variables and $\boldsymbol{X}_i^{\text{IV}}$ is binary, then λ_0, λ_1 can only be identified up to 3 parameters each.[12] Ultimately, this changes little to the identification strategy, and little to the estimation.

In a simulation with Roy selection-into-mediator based on unobserved error terms, the MTE approach pushes conventional CM estimates back to the true value. Figure 4 shows how the MTE-based approach corrects unadjusted CM effect estimates.

5 MTE-based Estimation of CM Effects

A conventional approach to estimating CM effects involves a two-stage approach to estimating the ADE and the AIE: the first-stage (Z_i on D_i), and the second-stage (Z_i, D_i on Y_i). An MTE approach is a simple and intuitive addition to this approach: including the CF terms λ_0, λ_1 in the second-stage regression to address selection-into-mediator.

This section presents two practical estimation strategies. First, I demonstrate how to estimate CM effects with an assumed distribution of error terms, focusing on the Heckman selection model as the leading case. Second, I consider a more flexible semi-parametric approach that avoids distributional assumptions — at the cost of semi-parametrically estimating the corresponding CFs. While both methods effectively address the selection bias issues detailed in previous sections, they differ in their implementation complexity, efficiency, and underlying assumptions.

5.1 Parametric MTE

A parametric approach to MTEs solves the identification problem by assuming a distribution for the unobserved error terms in the first-stage selection model, and modelling selection based on this distribution. The Heckman selection model is the most pertinent example, assuming the normal distribution for unobserved errors (Heckman 1979). Assuming distributions other than the bivariate normal works in exactly the same manner, replacing the relevant density functions for those of an alternative distribution. As such, this section focuses exclusively on the Heckman selection model. This estimation approach is the same as the original parametric selection model definition of MTEs, in Björklund & Moffitt (1987).

The Heckman selection model assumes unobserved errors V_i follow a normal distribution,

[12] The value of 3 comes from the cases that $Z_i, \boldsymbol{X}_i^{\text{IV}}$ each could take 2 values, so $\pi(z'; \boldsymbol{X}_i)$ has 4 possible values, giving the semi-parametric identification (and estimation) of each CF only 3 degrees of freedom to work with. If the MTE-associated CFs are instead assumed to have a known distribution (i.e., parametric MTE), then those concerns do not matter.

so estimates the first-stage using a probit model.

$$\Pr\left(D_i = 1 \,|\, Z_i, \boldsymbol{X}_i\right) = \Phi\big(\theta + \overline{\pi} Z_i + \boldsymbol{\zeta}' \boldsymbol{X}_i\big),$$

where $\Phi(.)$ is the cumulative density function for the standard normal distribution, and $\theta, \overline{\pi}, \boldsymbol{\zeta}$ are parameters estimated with maximum likelihood. In the parametric case, an excluded instrument ($\boldsymbol{X}_i^{\text{IV}}$) is not technically necessary in the first-stage equation — though not including one exposes the method to bias from misspecification if the errors are not normally distributed. Thus, it is best practice to use this method with access to an instrument.

From this probit first-stage, construct the inverse Mills ratio terms to serve as the MTE-associated CFs. These terms capture the correlation between unobserved factors influencing both mediator selection and outcomes, when the errors are normally distributed.

$$\lambda_0(p') = \frac{\phi(-\Phi^{-1}(p'))}{\Phi(-\Phi^{-1}(p'))}, \quad \lambda_1(p') = \frac{\phi(\Phi^{-1}(p'))}{\Phi(\Phi^{-1}(p'))}, \quad \text{for } p' \in (0,1)$$

where $\phi(.)$ is the probability density function for the standard normal distribution.

Lastly, the second-stage is estimated with OLS, including the MTE-associated CFs with plug-in estimates of the mediator propensity score, and $\boldsymbol{\varphi}'$ a linear approximation of nuisance function $\varphi(.)$.

$$\begin{aligned} \mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i\right] \;&=\; \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \boldsymbol{\varphi}' \boldsymbol{X}_i^- \\ &\quad + \rho_0 (1 - D_i) \lambda_0\big(\widehat{\pi}(Z_i; \boldsymbol{X}_i)\big) + \rho_1 D_i \lambda_1\big(\widehat{\pi}(Z_i; \boldsymbol{X}_i)\big) + \varepsilon_i, \end{aligned}$$

where $\widehat{\pi}\big(z'; \boldsymbol{X}_i\big)$ are the predictions from the probit first-stage.

The resulting ADE and AIE estimates are composed from sample estimates of the terms in Theorem MTE Approach,

$$\widehat{\text{ADE}} = \widehat{\gamma} + \widehat{\delta}\,\overline{D}, \quad \widehat{\text{AIE}} = \widehat{\overline{\pi}} \left(\widehat{\beta} + \widehat{\delta}\,\overline{Z} + \big(\widehat{\rho}_1 - \widehat{\rho}_0\big) \frac{1}{n} \sum_{i=1}^{n} \Gamma\big(\widehat{\pi}(0; \boldsymbol{X}_i), \widehat{\pi}(1; \boldsymbol{X}_i)\big)\right)$$

where $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$, $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, $\widehat{\overline{\pi}}$ is the estimate of the mean compliance rate, and $\frac{1}{n}\sum_{i=1}^{n} \Gamma(.,.)$ is the average of the complier adjustment term as a function of λ_1 with $\widehat{\pi}\big(0; \boldsymbol{X}_i\big), \widehat{\pi}\big(1; \boldsymbol{X}_i\big)$ values plugged in.

The standard errors for estimates can be computed using the delta method. Specifically, accounting for sampling variability in both first-stage mediator propensity score estimation and second-stage causal effects estimation. This approach yields $\sqrt{n}$-consistent estimates when the underlying error terms follow a bivariate normal distribution — i.e., when $\pi(Z_i; \boldsymbol{X}_i)$ is correctly modelled by the probit first-stage. Errors can also be estimated by the bootstrap, by including estimation of both the first and second-stage within each bootstrap iteration.

In practice, a parametric MTE approach is simple to implement using standard statistical packages. The key advantage is computational simplicity and efficiency, particularly in moderate-sized samples. However, this comes at the cost of strong distributional assumptions. For example, if the error terms deviate substantially from joint normality, the estimates may be biased.[13]

5.2 Semi-parametric MTE

For settings where researchers are not comfortable specifying a specific distribution for the error terms, a semi-parametric MTE approach will nonetheless consistently estimate CM effects. This method maintains the same identification strategy but avoids assuming a specific error distribution. This estimation approach is used in the modern semi-parametric approach to estimating the distribution of MTEs, for example in Brinch et al. (2017), Heckman & Vytlacil (2007).

The semi-parametric approach begins with flexible estimation of the first-stage, estimating the mediator propensity score,

$$\Pr\left(D_i = 1 \,|\, Z_i, \boldsymbol{X}_i\right) = \pi\left(Z_i; \boldsymbol{X}_i\right),$$

where $\boldsymbol{X}_i$ must include the instrument(s) $\boldsymbol{X}_i^{\text{IV}}$. This can be estimated using flexible methods, as long as the first-stage is estimated $\sqrt{n}$-consistently.[14] An attractive option is the Klein & Spady (1993) semi-parametric binary response model, which avoids relying on an assumed distribution of first-stage errors though requires a linear specification. If it is important to avoid a linear specification, then a probability forest avoids linearity assumptions (Athey, Tibshirani & Wager 2019) — though is best used for cases with many columns in the $\boldsymbol{X}_i$ variables.

The second-stage is estimated with semi-parametric methods. Consider the subsamples of mediator refusers and takers separately,

$$\begin{aligned}
\mathbb{E}\left[Y_i \,|\, Z_i, D_i = 0, \boldsymbol{X}_i\right] &= \alpha + \gamma Z_i + \varphi_0\big(\boldsymbol{X}_i^-\big) + \rho_0 \lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big), \\
\mathbb{E}\left[Y_i \,|\, Z_i, D_i = 1, \boldsymbol{X}_i\right] &= (\alpha + \beta) + (\gamma + \delta) Z_i + \varphi_1\big(\boldsymbol{X}_i^-\big) + \rho_1 \lambda_1\big(\pi(Z_i; \boldsymbol{X}_i)\big).
\end{aligned}$$

The separated subsamples can be estimated, each individually, with semi-parametric methods.

[13] While this concern is immaterial in an IV setting estimating the LATE (Kline & Walters 2019), it is pertinent in this setting as the CF extrapolates from IV compliers to mediator compliers.

[14] If an estimate of the first-stage that is not $\sqrt{n}$-consistent is used (e.g., a modern machine learning estimator), then the resulting second-stage estimate will not be $\sqrt{n}$-consistent.

The linear parameters (including linear approximations $\boldsymbol{\varphi}'_{d'}$ of nuisance functions $\varphi_{d'}(.)$)[15] can be estimated with OLS, while $\rho_0\lambda_0$ and $\rho_1\lambda_1$ take a flexible semi-parametric specification with first-stage estimates $\widehat{\pi}(Z_i; \boldsymbol{X}_i)$ plugged in. An attractive option is a series estimator, such as a spline specification, as this estimates the function without assuming a functional form but maintains $\sqrt{n}$-consistency.

The ADE is estimated by this approach as follows. Take $\widehat{\gamma}$, the $D_i = 0$ subsample estimate of $\mathbb{E}\left[\gamma\right]$, and $(\widehat{\gamma + \delta})$, the $D_i = 1$ subsample estimate of $\mathbb{E}\left[\gamma + \delta\right]$, to give

$$\widehat{\text{ADE}}^{\text{Semi}} = (1 - \overline{D})\,\widehat{\gamma} + \overline{D}\,(\widehat{\gamma + \delta}).$$

The AIE is less simple, for two reasons that differ from the parametric MTE setting. First, the intercepts for each subsample, α and $(\alpha + \beta)$, are not separately identified from the MTE-associated CFs if the λ_0, λ_1 functions are flexibly estimated. Second, a semi-parametric specification for the CFs mean ρ_0 and λ_0 are no longer separately identified from each other (and same for ρ_1, λ_1). As such, it is not possible to directly use $\widehat{\lambda}_0, \widehat{\lambda}_1$ in estimating the complier adjustment term (as is done in the parametric case).

These problems can be avoided by estimating the AIE using its relation to the ATE. Write $\widehat{\text{ATE}}$ for the point-estimate of the ATE, and $\widehat{\delta} = (\widehat{\gamma + \delta}) - \widehat{\gamma}$ for the point estimate of $\mathbb{E}\left[\delta\right]$, to give the following estimator,

$$\widehat{\text{AIE}}^{\text{Semi}} = \widehat{\text{ATE}} - (1 - \overline{Z})\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\gamma} + \widehat{\delta}\,\widehat{\pi}(1; \boldsymbol{X}_i)\right) - \overline{Z}\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\gamma} + \widehat{\delta}\,\widehat{\pi}(0; \boldsymbol{X}_i)\right),$$

where $\frac{1}{n}\sum_{i=1}^{n} \widehat{\gamma} + \widehat{\delta}\,\widehat{\pi}(0; \boldsymbol{X}_i)$ estimates the ADE conditional on $Z_i = 0$, $\mathbb{E}\left[\gamma + \delta D_i(0)\right]$, and $\frac{1}{n}\sum_{i=1}^{n} \widehat{\gamma} + \widehat{\delta}\,\widehat{\pi}(1; \boldsymbol{X}_i)$ estimates the ADE conditional on $Z_i = 1$, $\mathbb{E}\left[\gamma + \delta D_i(1)\right]$. Appendix A.8 describes the reasoning for this estimator of the AIE, relative to estimates of the ATE and ADE, in further detail.

This semi-parametric approach achieves valid estimation of the CM effects, without specifying the distribution behind unobserved error terms, and achieves desirable properties as long as the first-stage correctly estimates the mediator propensity score, and the structural assumptions hold true. The standard errors for estimates can again be computed using the delta method, or estimated by the bootstrap — again, across both first and second-stages within each bootstrap iteration. Note that relying on propensity score estimation requires assumptions that can be found wanting in real-world settings; a common support condition for the mediator is required, and a semi-/non-parametric first-stage may become cumbersome if there are many control variables or many rows of data.

[15] Appropriate interactions between Z_i, D_i and $\boldsymbol{X}_i$ can also flexibly control for $\boldsymbol{X}_i$, again avoiding linearity assumptions.

5.3 Simulation Evidence

The following simulation gives an example to show how these methods work in practice. Suppose data observed to the researcher $Z_i, D_i, Y_i, \boldsymbol{X}_i$ are drawn from the following data generating processes, for $i = 1, \dots, N$, with $n = 5,000$ for this simulation.

$$Z_i \sim \text{Binom}\,(0.5)\,, \quad \boldsymbol{X}_i^- \sim N(4,1), \quad \boldsymbol{X}_i^{\text{IV}} \sim \text{Uniform}\,(-1,1)\,, \quad (U_{0,i}, U_{1,i}, U_{C,i}) \sim N\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right)$$

$\boldsymbol{\Sigma}$ is the matrix of parameters which controls the level of confounding from unobserved costs and benefits.[16]

Each i chooses to take mediator D_i by a Roy model, with following mean definitions for each $z', d' = 0, 1$

$$D_i(z') = \mathbb{1}\left\{C_i \leq Y_i(z', 1) - Y_i(z', 0)\right\},$$
$$\mu_{d'}\left(z'; \boldsymbol{X}_i\right) = (z' + d' + z'd') + \boldsymbol{X}_i^-, \quad \mu_C\left(z'; \boldsymbol{X}_i\right) = 3z' + \boldsymbol{X}_i^- - \boldsymbol{X}_i^{\text{IV}}.$$

Following Subsection 3.1, these data have the following first and second-stage equations:

$$D_i = \mathbb{1}\left\{U_{C,i} - \big(U_{1,i} - U_{0,i}\big) \leq -3Z_i + \boldsymbol{X}_i^- - \boldsymbol{X}_i^{\text{IV}}\right\},$$
$$Y_i = Z_i + D_i + Z_i D_i + \boldsymbol{X}_i^- + (1 - D_i)\,U_{0,i} + D_i U_{1,i}.$$

Treatment Z_i has a causal effect on outcome Y_i, and it operates partially through mediator D_i. Outcome mean $\mu_{D_i}\left(Z_i; \boldsymbol{X}_i\right)$ contains an interaction term, $Z_i D_i$, so while Z_i, D_i have constant partial effects, the ATE depends on how many i choose to take the mediator and there is treatment effect heterogeneity.

After Z_i is assigned, i chooses to take mediator D_i by considering the costs and benefits — which vary based on Z_i, demographic controls $\boldsymbol{X}_i$, and the (non-degenerate) unobserved error terms $U_{i,0}, U_{1,i}$. As a result, sequential quasi-random assignment does not hold; the mediator is not conditionally ignorable. Thus, a conventional approach to CM does not give an estimate for how much of the ATE goes through mediator D, but is contaminated by selection bias thanks to the unobserved error terms.

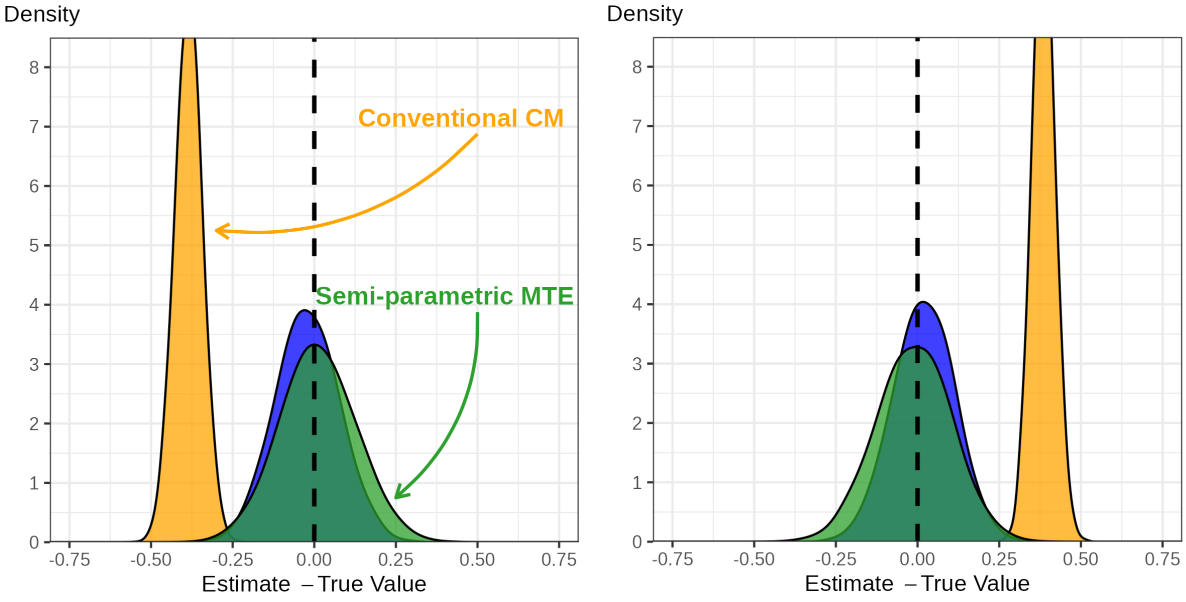
I simulate this data generating process 10,000 times, using $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} 1 & 0.75 & 0 \\ 0.75 & 2.25 & 0 \\ 0 & 0 & 0.25 \end{smallmatrix}\right)$,[17] and estimate CM effects with conventional CM methods (two-stage OLS) and the introduced MTE

[16] The correlation and relative standard deviations for $U_{0,i}, U_{1,i}$ affect how large selection bias in conventional CM estimates; correlation for these with unobserved costs $U_{C,i}$ does not particularly matter, though increased variance in unobserved costs makes estimates less precise for both OLS and MTE methods.

[17] This choice of parameters has $\text{Var}\,(U_{0,i}) = 1, \text{Var}\,(U_{1,i}) = 2.25, \text{Corr}\big(U_{0,i}, U_{1,i}\big) = 0.5$ so that unobserved errors meaningfully confound conventional CM methods, with notable heteroscedasticity. Unobserved costs are uncorrelated with $U_{0,i}, U_{1,i}$ (although non-zero correlation would not meaningfully change the results), and $\text{Var}\,(U_{C,i}) = 0.25$ maintains uncertainty in unobserved costs.

methods. In this simulation $\Pr(D_i = 1) = 0.379$, and 65.77% of the sample are mediator compliers (for whom $D_i(0) = 0$ and $D_i(1) = 1$). This gives an ATE value of 2.60, ADE 1.38, and AIE 1.22, respectively.[18]

Figure 5: Simulated Distribution of CM Effect Estimates, Semi-parametric versus OLS, Relative to True Value.



(a) $\widehat{\text{ADE}} - \text{ADE}$.

(b) $\widehat{\text{AIE}} - \text{AIE}$.

Note: These figures show the empirical density of point estimates minus the true average effect, for 10,000 different datasets generated from a Roy model with correlated uniformly distributed error terms. The black dashed line is the true value; orange is the distribution of conventional CM estimates from two-stage OLS (Imai et al. 2010), and green estimates with a two-stage semi-parametric MTE.

Figure 4 shows how these estimates perform, with a parametric MTE approach, relative to the true value. The OLS estimates' distribution do not overlap the true values for any standard level of significance; the distance between the OLS estimates and the true values are the underlying bias terms derived in Theorem 1. The parametric MTE approach perfectly reproduces the true values, as the probit first-stage correctly models the normally distributed error terms. The semi-parametric approach (not shown in Figure 4) performs similarly, with a wider distribution; this is to be expected comparing a correctly specified parametric model with a semi-parametric one.

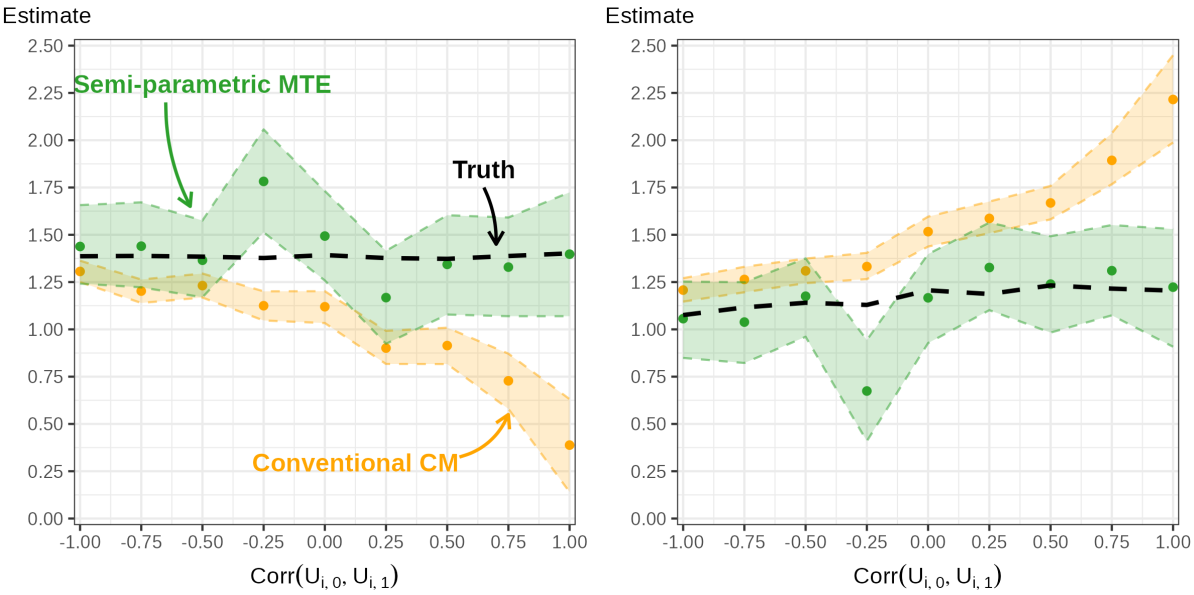
The parametric MTE may not be appropriate in setting with non-normal error terms. I simulated the same data again, but transform $U_{0,i}, U_{1,i}$ to be correlated uniform errors (with the same standard deviations as previously). Figure 5 shows the resulting distribution of

[18] Note that ATE = ADE + AIE in this setting. $\Pr(Z_i = 1) = 0.5$ ensures this equality, but it is not guaranteed in general. See Appendix A.8.

point-estimates, relative to the truth, for the parametric and semi-parametric approaches. The parametric MTE is slightly off target, showing persistent bias from incorrectly specifying the error term distribution. The semi-parametric approach is centred exactly around the truth, with a slightly higher variance (as is expected).

Figure 6: MTE-Based Estimates Work with Different Error Term Parameters.

(a) ADE. **(b)** AIE.



Note: These figures show the OLS and MTE-based point estimates of the ADE and AIE, for $n = 5{,}000$ sample size, varying $\text{Corr}\left(U_{0,i}, U_{1,i}\right)$ values with $\text{Var}\left(U_{0,i}\right) = 1, \text{Var}\left(U_{1,i}\right) = 1.5$ fixed. The black dashed line is the true value, coloured points are points estimates for the respective data generated, and shaded regions are the 95% confidence intervals from 1,000 bootstraps each. Orange represents OLS estimates, green blue the semi-parametric MTE approach.

The error terms determine the bias in OLS estimates of the ADE and AIE, so the bias varies for different values of the error-term parameters $\text{Corr}\left(U_{0,i}, U_{1,i}\right) \in [-1, 1]$ and $\text{Var}\left(U_{0,i}\right), \text{Var}\left(U_{1,i}\right) \geq 0$. The true AIE values vary, because $D_i(Z_i)$ compliers have higher average values of $U_{1,i} - U_{0,i}$ as $\text{Corr}\left(U_{0,i}, U_{1,i}\right)$ increases. Figure 6 shows MTE-based estimates against estimates calculated by standard OLS, showing 95% confidence intervals calculated from 1,000 bootstraps. The point estimates of the MTE-based do not exactly equal the true values, as they are estimates from one simulation (not averages across many generated datasets, as in Figure 5). The MTE approach improves on OLS estimates by correcting for bias, with confidence regions overlapping the true values.[19] This correction did not come for free: the standard errors are significantly greater in an MTE approach than conventional

[19] In the appendix, Figure A1 shows the same simulation while varying $\text{Var}\left(U_{1,i}\right)$, with fixed $\text{Var}\left(U_{0,i}\right) = 1, \text{Corr}\left(U_{0,i}, U_{1,i}\right) = 0.5$. The conclusion is the same as for varying the correlation coefficient, ρ, in Figure 6.

CM estimates (based on two-stage OLS). In this manner, this simulation shows the pros and cons of using the MTE approach to estimating CM effects in practice.

6 CM in the Oregon Health Insurance Experiment

In the Oregon Health Insurance Experiment, winning the wait-list lottery significantly improved subjective health and well-being among participants. This study investigates the mechanisms behind these benefits, quantifying the extent to which improvements are mediated through increased healthcare usage.

To address concerns for unobserved selection-into-healthcare, I use the respondents' regular healthcare provider before the wait-list lottery as an IV for healthcare usage. Approximately 73.2% reported visiting a healthcare provider within the past year, but rates vary notably depending on their provider type: those who reported attended hospital emergency rooms (A&E) and urgent care clinics before the wait-list lottery reported significantly lower healthcare visitation rates after the lottery (8.4 and 11 percentage points lower, respectively) than the 40% who attended private clinics.[20] The IV validity arises from differential costs faced by individuals based on their usual care provider. Private clinics generally charge through health insurance and are more expensive without coverage, while A&E and urgent care often provide costly services but rarely follow up on unpaid bills, effectively creating variation in healthcare attendance costs. Additionally, individuals' choice of provider likely depend on neighbourhood-based access.

Initial results with unadjusted CM estimates suggest almost no mediating role for healthcare usage; the unadjusted estimates of the AIE are close to zero for both outcomes, contradicting intuitive suggestive evidence. These estimates control for diagnosis of serious health conditions (before the wait-list lottery), such as kidney disease or diabetes. However, this approach would not be considering possible unobserved confounding coming from underlying health conditions. My MTE-based approach attempts to model this unobserved confounding, and applying them to these data reveals a much larger, positive AIE, restoring the mediating mechanism of healthcare usage in line with suggestive intuition. This is because a correlational estimate of health and well-being gains to healthcare visits are practically zero, while the IV estimates restore positive gains and the MTE-based methods pick this up with a larger AIE estimate. These numbers are reported in Table 1, where panel A shows the CM effects with the binary outcome of subjective health, and panel B the binary outcome of subjective well-being.

[20] The combined F statistic for the categorical variable for healthcare usual location (before the wait-list lottery) on healthcare usage (following the lottery) is 38.4.

Table 1: CM Effect Estimates for Wait-list Lottery Effects on Health and Well-being.

	First-stage	ATE	ADE	AIE	AIE / ATE
	(1)	(2)	(3)	(4)	(5)
Panel A: Health overall good?					
Unadjusted	4.700	4.500	4.700	-0.190	-0.043
	(0.880)	(0.910)	(0.910)	(0.063)	(0.018)
Parametric MTE	4.70	5.00	2.50	1.80	0.36
	(0.88)	(0.95)	(1.20)	(0.55)	(0.14)
Semi-parametric MTE	4.70	5.00	2.00	3.10	0.61
	(0.84)	(0.97)	(1.30)	(0.87)	(0.22)
Panel B: Happy overall?					
Unadjusted	4.7000	7.1000	7.0000	0.0670	0.0094
	(0.8600)	(0.9600)	(0.9600)	(0.0520)	(0.0077)
Parametric MTE	4.70	7.50	5.00	1.90	0.25
	(0.870)	(0.990)	(1.100)	(0.510)	(0.075)
Semi-parametric MTE	4.70	7.50	5.00	2.50	0.34
	(0.84)	(0.93)	(1.20)	(0.74)	(0.11)

Note: This table shows the point estimates (and SEs in brackets) of applying the proposed CM methods to replication data from the Oregon Health Insurance Experiment (Finkelstein & Baicker 2014). The first-stage column to the average effect of winning the wait-list lottery on healthcare usage (mediator first-stage), ATE average effect on surveyed health and well-being, ADE and AIE to respective CM effects through and absent healthcare usage. SEs were calculated with 1,000 bootstrap replications. The numbers are pp increases in the binary outcome, so an estimate of 4.1 in row 1 column 1 means an increase in 4.1 pp of using healthcare in the last 12 months after winning the wait-list lottery.

This reversal in conclusions highlights the importance of correcting for negative selection into healthcare usage. A conventional approach to CM fails to account for the fact that individuals with poorer underlying health tend to visit healthcare providers more frequently, generating negative selection bias that obscures the true positive AIE, and clouds inference on the mediator mechanism. By explicitly adjusting for this bias using the MTE approach, I isolate a credible positive indirect effect of healthcare usage on subjective health and well-being.

These findings offer credible evidence that improved healthcare access yields meaningful subjective health and well-being benefits, despite previous research emphasising negligible effects on objective health measures such as blood pressure (Baicker, Taubman, Allen, Bernstein, Gruber, Newhouse, Schneider, Wright, Zaslavsky & Finkelstein 2013). Subjective measures likely reflect broader psychological and financial relief associated with reduced healthcare-related anxiety and diminished risk of catastrophic medical debt, thus producing more noticeable short-term subjective improvements.

Nevertheless, this analysis is subject to notable limitations. The IV is not ideal, and potentially more important mediators (such as explicit health insurance status) would require additional IVs beyond the wait-list lottery itself, presenting a challenging identification issue. Furthermore, the 95% confidence intervals for both ADE and AIE estimates (based on boostrapped SEs) remain large, though statistically significant and excluding zero. This uncertainty underscores common challenges in applied CM analyses, where statistical precision can be limited by data constraints.

7 Summary and Concluding Remarks

This paper has studied a selection-on-observables approach to CM in a natural experiment setting. I have shown the pitfalls of using the most popular methods for estimating direct and indirect effects without a clear case for the mediator being quasi-randomly assigned. Using the Roy model as a benchmark, a mediator is unlikely to be quasi-randomly assigned in natural experiment settings, and the bias terms likely crowd out inference regarding CM effects.

This paper has also contributed to the growing CM literature in economics, connecting to MTE methods and developing a compelling way of estimating direct and indirect effects in a natural experiment setting. It has also recognised limitations in the common practice of suggestive evidence for mechanisms, and given credible CM estimates in a famous natural experiment setting well-known in the economics field. Further research could build on the approaches presented by suggesting efficiency improvements, adjustments for common statistical irregularities (say, cluster dependence), or integrating the MTE approach to the growing double robustness literature on CM (Farbmacher, Huber, Laffers, Langen & Spindler 2022, Bia, Huber & Laffers 2024).

These findings do not provide a blanket endorsement for applied researchers to use CM methods. There are strong structural assumptions for adjusting identifying CM effects despite unobserved selection-into-mediator, and inference requires an IV for mediator take-up. If these assumptions do not hold true, then selection-adjusted estimates of CM effects will also be biased, and will not improve on an unadjusted conventional approach.

Yet, there are likely settings in which the structural assumptions are credible. Mediator monotonicity aligns well with economic theory in many cases, and it is plausible for researchers to study big data settings with external variation in mediator take-up costs. In these cases, this paper opens the door to identifying mechanisms behind treatment effects in natural experiment settings.

References

Angrist, J. D. (1998), 'Estimating the labor market impact of voluntary military service using social security data on military applicants', *Econometrica* **66**(2), 249–288. https://doi.org/10.2307/2998558. 9

Angrist, J. D. & Pischke, J.-S. (2009), *Mostly harmless econometrics: An empiricist's companion*, Princeton university press. 10

Athey, S., Tibshirani, J. & Wager, S. (2019), 'Generalized random forests', *The Annals of Statistics* **47**(2), 1148–1178. https://doi.org/10.1214/18-aos1709. 22

Baicker, K., Taubman, S. L., Allen, H. L., Bernstein, M., Gruber, J. H., Newhouse, J. P., Schneider, E. C., Wright, B. J., Zaslavsky, A. M. & Finkelstein, A. N. (2013), 'The oregon experiment—effects of medicaid on clinical outcomes', *New England Journal of Medicine* **368**(18), 1713–1722. https://doi.org/10.1056/nejmsa1212321. 28

Bia, M., Huber, M. & Lafférs, L. (2024), 'Double machine learning for sample selection models', *Journal of Business & Economic Statistics* **42**(3), 958–969. https://doi.org/10.1080/07350015.2023.2271071. 29

Björklund, A. & Moffitt, R. (1987), 'The estimation of wage gains and welfare gains in self-selection models', *The Review of Economics and Statistics* pp. 42–49. https://doi.org/10.2307/1937899. 15, 18, 20

Blackwell, M., Ma, R. & Opacic, A. (2024), 'Assumption smuggling in intermediate outcome tests of causal mechanisms', *arXiv preprint arXiv:2407.07072* . https://doi.org/10.20944/preprints202411.2377.v1. 3, 6

Brinch, C. N., Mogstad, M. & Wiswall, M. (2017), 'Beyond late with a discrete instrument', *Journal of Political Economy* **125**(4), 985–1039. https://doi.org/10.1086/692712. 3, 19, 22

Cinelli, C., Forney, A. & Pearl, J. (2024), 'A crash course in good and bad controls', *Sociological Methods & Research* **53**(3), 1071–1104. https://doi.org/10.2139/ssrn.3689437. 11

Deuchert, E., Huber, M. & Schelker, M. (2019), 'Direct and indirect effects based on difference-in-differences with an application to political preferences following the vietnam draft lottery', *Journal of Business & Economic Statistics* **37**(4), 710–720. https://doi.org/10.1080/07350015.2017.1419139. 3

Ding, P. & Miratrix, L. W. (2015), 'To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias', *Journal of Causal Inference* **3**(1), 41–57. https://doi.org/10.1515/jci-2013-0021. 11

Eisenhauer, P., Heckman, J. J. & Vytlacil, E. (2015), 'The generalized roy model and the cost-benefit analysis of social programs', *Journal of Political Economy* **123**(2), 413–443. https://doi.org/10.1086/679498. 14

Farbmacher, H., Huber, M., Lafférs, L., Langen, H. & Spindler, M. (2022), 'Causal mediation analysis with double machine learning', *The Econometrics Journal* **25**(2), 277–300. https://doi.org/10.1093/ectj/utac003. 29

Finkelstein, A. & Baicker, K. (2014), 'Oregon health insurance experiment, 2007-2010'. https://doi.org/10.3886/ICPSR34314.v3. 4, 28

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K. & Group, O. H. S. (2012), 'The oregon health insurance experiment: Evidence from the first year*', *The Quarterly Journal of Economics* **127**(3), 1057–1106. https://doi.org/10.1093/qje/qjs020. 1, 4, 5, 8

Florens, J.-P., Heckman, J. J., Meghir, C. & Vytlacil, E. (2008), 'Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects', *Econometrica* **76**(5), 1191–1206. https://doi.org/10.3982/ecta5317. 3

Flores, C. A. & Flores-Lagunes, A. (2009), 'Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness', *IZA Discussion paper* . https://doi.org/10.2139/ssrn.1423353. 3

Frölich, M. & Huber, M. (2017), 'Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **79**(5), 1645–1666. https://doi.org/10.1111/rssb.12232. 3, 7

Green, D. P., Ha, S. E. & Bullock, J. G. (2010), 'Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose', *The Annals of the American Academy of Political and Social Science* **628**(1), 200–208. https://doi.org/10.1177/0002716209351526. 3, 6

Heckman, J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica: journal of the econometric society* pp. 679–694. https://doi.org/10.2307/1913937. 15

Heckman, J., Ichimura, H., Smith, J. & Todd, P. (1998), 'Characterizing selection bias using experimental data', *Econometrica* **66**(5), 1017–1098. https://doi.org/10.2307/2999630. 10

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica: Journal of the econometric society* pp. 153–161. https://doi.org/10.2307/1912352. 15, 20, 45

Heckman, J. J. & Honore, B. E. (1990), 'The empirical content of the roy model', *Econometrica: Journal of the Econometric Society* pp. 1121–1149. https://doi.org/10.2307/2938303. 14

Heckman, J. J. & Pinto, R. (2015), 'Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs', *Econometric reviews* **34**(1-2), 6–31. https://doi.org/10.3386/w19314. 3

Heckman, J. J. & Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation 1', *Econometrica* **73**(3), 669–738. https://doi.org/10.1111/j.1468-0262.2005.00594.x. 3, 12, 15, 16, 18, 42

Heckman, J. & Navarro-Lozano, S. (2004), 'Using matching, instrumental variables, and control functions to estimate economic choice models', *Review of Economics and statistics* **86**(1), 30–57. https://doi.org/10.3386/w9497. 3

Heckman, J., Pinto, R. & Savelyev, P. (2013), 'Understanding the mechanisms through which an influential early childhood program boosted adult outcomes', *American economic review* **103**(6), 2052–2086. https://doi.org/10.3386/w18581. 3

Heckman, J. & Vytlacil, E. (2007), Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments, *in* J. Heckman & E. Leamer, eds, 'Handbook of Econometrics', 1 edn, Vol. 6B, Elsevier, chapter 71. 22

Huber, M. (2020), 'Mediation analysis', *Handbook of labor, human resources and population economics* pp. 1–38. https://doi.org/10.1007/978-3-319-57365-6_162-1. 3

Huber, M., Hsu, Y.-C., Lee, Y.-Y. & Lettry, L. (2020), 'Direct and indirect effects of continuous treatments based on generalized propensity score weighting', *Journal of Applied Econometrics* **35**(7), 814–840. https://doi.org/10.1002/jae.2765. 7

Imai, K., Keele, L. & Yamamoto, T. (2010), 'Identification, inference and sensitivity analysis for causal mediation effects', *Statistical Science* pp. 51–71. https://doi.org/10.1214/10-sts321. 1, 3, 8, 9, 19, 25, 34, 39, 46, 48

Imai, K., Tingley, D. & Yamamoto, T. (2013), 'Experimental designs for identifying causal mechanisms', *Journal of the Royal Statistical Society Series A: Statistics in Society* **176**(1), 5–51. https://doi.org/10.1111/j.1467-985x.2012.01032.x. 3

Imbens, G. & Angrist, J. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475. https://doi.org/10.2307/2951620. 7, 16

Klein, R. W. & Spady, R. H. (1993), 'An efficient semiparametric estimator for binary response models', *Econometrica* pp. 387–421. https://doi.org/10.2307/2951556. 22

Kline, P. & Walters, C. R. (2019), 'On heckits, late, and numerical equivalence', *Econometrica* **87**(2), 677–696. https://doi.org/10.3982/ecta15444. 3, 18, 22, 42, 44, 45

Kwon, S. & Roth, J. (2024), 'Testing mechanisms', *ArXiv preprint* . https://doi.org/10.48550/arXiv.2404.11739. 3

Ludwig, J., Kling, J. R. & Mullainathan, S. (2011), 'Mechanism experiments and policy evaluations', *Journal of economic Perspectives* **25**(3), 17–38. https://doi.org/10.1257/jep.25.3.17. 3

R Core Team (2025), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. 48

Roy, A. D. (1951), 'Some thoughts on the distribution of earnings', *Oxford economic papers* **3**(2), 135–146. https://doi.org/10.1093/oxfordjournals.oep.a041827. 14

Słoczyński, T. (2022), 'Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights', *Review of Economics and Statistics* **104**(3), 501–509. https://doi.org/10.1162/rest_a_00953. 9

Tingley, D., Yamamoto, T., Hirose, K., Keele, L. & Imai, K. (2014), 'Mediation: R package for causal mediation analysis', *Journal of statistical software* **59**, 1–38. https://doi.org/10.18637/jss.v059.i05. 48

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**(1), 331–341. https://doi.org/10.1111/1468-0262.00277. 3, 16, 41

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019), 'Welcome to the tidyverse', *Journal of Open Source Software* **4**(43), 1686. https://doi.org/10.21105/joss.01686. 48

Wood, S., N., Pya & S"afken, B. (2016), 'Smoothing parameter and model selection for general smooth models (with discussion)', *Journal of the American Statistical Association* **111**, 1548–1575. https://doi.org/10.1080/01621459.2016.1180986. 48

A Supplementary Appendix

This section is for supplementary information, and validation of presented propositions and theorems. It is not meant for publication.

Any comments or suggestions may be sent to me at seh325@cornell.edu, or raised as an issue on the Github project, https://github.com/shoganhennessy/mediation-natural-experiment.

A.1 Identification in Causal Mediation

Imai et al. (2010, Theorem 1) states that the ADE and AIE are identified under sequential quasi-random assignment, at each level of $Z_i = 0, 1$. For $z' = 0, 1$:

$$\mathbb{E}\left[Y_i(1, D_i(z')) - Y_i(0, D_i(z'))\right] = \int \int \Big(\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i, \boldsymbol{X}_i\right]\Big) dF_{D_i \,|\, Z_i = z', \boldsymbol{X}_i} dF_{\boldsymbol{X}_i},$$

$$\mathbb{E}\left[Y_i(z', D_i(1)) - Y_i(z', D_i(0))\right] = \int \int \mathbb{E}\left[Y_i \,|\, Z_i = z', D_i, \boldsymbol{X}_i\right] \Big(dF_{D_i \,|\, Z_i = 1, \boldsymbol{X}_i} - dF_{D_i \,|\, Z_i = 0, \boldsymbol{X}_i}\Big) dF_{\boldsymbol{X}_i}.$$

I focus on the averages, which are identified by consequence of the above.

$$\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] = \mathbb{E}_{Z_i}\left[\mathbb{E}\left[Y_i(1, D_i(z')) - Y_i(0, D_i(z')) \,|\, Z_i = z'\right]\right]$$
$$\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] = \mathbb{E}_{Z_i}\left[\mathbb{E}\left[Y_i(z', D_i(1)) - Y_i(z', D_i(0)) \,|\, Z_i = z'\right]\right]$$

My estimand for the ADE is a simple rearrangement of the above. The estimand for the AIE relies on a different sequence, relying on (1) sequential quasi-random assignment, (2) conditional monotonicity. These give (1) identification equivalence of AIE local to mediator compliers conditional on $\boldsymbol{X}_i$ and AIE conditional on $\boldsymbol{X}_i$, LAIE = AIE, (2) identification of the complier score.

$$\begin{aligned}
&\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \,|\, \boldsymbol{X}_i\right] \\
&= \Pr\left(D_i(0) = 0, D_i(1) = 1 \,|\, \boldsymbol{X}_i\right) \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right] \\
&= \Pr\left(D_i(0) = 0, D_i(1) = 1 \,|\, \boldsymbol{X}_i\right) \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, \boldsymbol{X}_i\right] \\
&= \Pr\left(D_i(0) = 0, D_i(1) = 1 \,|\, \boldsymbol{X}_i\right) \Big(\mathbb{E}\left[Y_i \,|\, Z_i, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i, D_i = 0, \boldsymbol{X}_i\right]\Big) \\
&= \Big(\mathbb{E}\left[D_i \,|\, Z_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[D_i \,|\, Z_i = 0, \boldsymbol{X}_i\right]\Big) \Big(\mathbb{E}\left[Y_i \,|\, Z_i, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i, D_i = 0, \boldsymbol{X}_i\right]\Big)
\end{aligned}$$

Monotonicity is not technically required for the above. Breaking monotonicity would not change the identification in any of the above; it would be the same except replacing the complier score with a complier/defier score, $\Pr\left(D_i(0) \neq D_i(1) \,|\, \boldsymbol{X}_i\right) = \mathbb{E}\left[D_i \,|\, Z_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[D_i \,|\, Z_i = 0, \boldsymbol{X}_i\right]$.

A.2 Bias in Causal Mediation (CM) Estimands

Suppose that Z_i is ignorable conditional on $\boldsymbol{X}_i$, but D_i is not.

A.2.1 Bias in the Average Direct Effect (ADE)

To show that the conventional approach to mediation gives an estimate for the ADE with selection and group difference-bias, start with the components of the conventional estimands. This proof starts with the relevant expectations, conditional on a specific value of $\boldsymbol{X}_i$ and $d' \in \{0, 1\}$.

$$\begin{aligned}\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i = d', \boldsymbol{X}_i\right] =& \mathbb{E}\left[Y_i(1, D_i(Z_i)) \,|\, D_i(1) = d', \boldsymbol{X}_i\right], \\ \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = d', \boldsymbol{X}_i\right] =& \mathbb{E}\left[Y_i(0, D_i(Z_i)) \,|\, D_i(0) = d', \boldsymbol{X}_i\right]\end{aligned}$$

And so,

$$\begin{aligned}&\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = d', \boldsymbol{X}_i\right] \\ &\ = \mathbb{E}\left[Y_i(1, D_i(Z_i)) \,|\, D_i(1) = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(0, D_i(Z_i)) \,|\, D_i(0) = d', \boldsymbol{X}_i\right] \\ &\ = \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, D_i(1) = d', \boldsymbol{X}_i\right] \\ &\qquad + \mathbb{E}\left[Y_i(0, D_i(Z_i)) \,|\, D_i(1) = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(0, D_i(Z_i)) \,|\, D_i(0) = d', \boldsymbol{X}_i\right].\end{aligned}$$

The final term is a sum of the ADE, conditional on $D_i(1) = d'$, and a selection bias term — difference in baseline outcomes between the (partially overlapping) groups for whom $D_i(1) = d'$ and $D_i(0) = d'$.

To reach the final term, note the following.

$$\begin{aligned}&\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, \boldsymbol{X}_i\right] \\ &\ = \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, D_i(1) = d', \boldsymbol{X}_i\right] \\ &\qquad + \Big(1 - \Pr\left(D_i(1) = d' \,|\, \boldsymbol{X}_i\right)\Big) \begin{pmatrix}\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, D_i(1) = d', \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, D_i(1) = 1 - d', \boldsymbol{X}_i\right]\end{pmatrix}\end{aligned}$$

The second term is the difference between the ADE and LADE local to relevant complier groups.

Collect everything together, as follows.

$$
\begin{aligned}
&\mathbb{E}\left[Y_i \mid Z_i = 1, D_i = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \mid Z_i = 0, D_i = d', \boldsymbol{X}_i\right] \\
=& \underbrace{\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid \boldsymbol{X}_i\right]}_{\text{ADE, conditional on } \boldsymbol{X}_i} \\
&+ \underbrace{\mathbb{E}\left[Y_i(0, D_i(Z_i)) \mid D_i(1) = d', \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(0, D_i(Z_i)) \mid D_i(0) = d', \boldsymbol{X}_i\right]}_{\text{Selection bias}} \\
&+ \underbrace{\Big(1 - \Pr\left(D_i(1) = d' \mid \boldsymbol{X}_i\right)\Big) \begin{pmatrix} \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid D_i(1) = 1 - d', \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \mid D_i(1) = d', \boldsymbol{X}_i\right] \end{pmatrix}}_{\text{group difference-bias}}
\end{aligned}
$$

The proof is achieved by applying the expectation across $D_i = d'$, and $\boldsymbol{X}_i$.

A.2.2 Bias in the Average Indirect Effect (AIE)

To show that the conventional approach to mediation gives an estimate for the AIE with selection and group difference-bias, start with the definition of the ADE — the direct effect among compliers times the size of the complier group.

This proof starts with the relevant expectations, conditional on a specific value of $\boldsymbol{X}_i$.

$$
\begin{aligned}
&\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \mid \boldsymbol{X}_i\right] \\
=& \Pr\left(D_i(0) = 0, D_i(1) = 1 \mid \boldsymbol{X}_i\right) \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \mid D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right]
\end{aligned}
$$

When D_i is not ignorable, the bias comes from estimating the second term, $\mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \mid D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right]$, the indirect effect among mediator compliers.

Let $z' \in \{0, 1\}$. Again, note the mean outcomes in terms of average potential outcomes,

$$
\begin{aligned}
\mathbb{E}\left[Y_i \mid Z_i = z', D_i = 1, \boldsymbol{X}_i\right] =& \mathbb{E}\left[Y_i(z', 1) \mid D_i = 1, \boldsymbol{X}_i\right], \\
\mathbb{E}\left[Y_i \mid Z_i = z', D_i = 0, \boldsymbol{X}_i\right] =& \mathbb{E}\left[Y_i(z', 0) \mid D_i = 0, \boldsymbol{X}_i\right].
\end{aligned}
$$

Compose the selection bias term, as follows.

$$
\begin{aligned}
&\mathbb{E}\left[Y_i \mid Z_i = z', D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \mid Z_i = z', D_i = 0, \boldsymbol{X}_i\right] \\
&= \mathbb{E}\left[Y_i(z', 1) \mid D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(z', 0) \mid D_i = 0, \boldsymbol{X}_i\right] \\
&= \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \mid D_i = 1, \boldsymbol{X}_i\right] + \mathbb{E}\left[Y_i(z', 0) \mid D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(z', 0) \mid D_i = 0, \boldsymbol{X}_i\right]
\end{aligned}
$$

The final term is a sum of the AIE, among the treated group $D_i = 1$, and a selection bias term — difference in baseline potential outcomes between the groups for whom $D_i = 1$ and $D_i = 0$.

The AIE is the direct effect among compliers times the size of the complier group, so we need to compensate for the difference between the treated group $D_i = 1$ and complier group $D_i(0) = 0, D_i(1) = 1$.

Start with the difference between treated group's average and overall average.

$$\begin{aligned}
&\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i = 1, \boldsymbol{X}_i\right] \\
=&\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, \boldsymbol{X}_i\right] \\
&+ \Big(1 - \Pr\left(D_i = 1 \,|\, \boldsymbol{X}_i\right)\Big) \begin{pmatrix} \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i = 1, \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i = 0, \boldsymbol{X}_i\right] \end{pmatrix}
\end{aligned}$$

Then the difference between the compliers' average and the overall average.

$$\begin{aligned}
&\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right] \\
=&\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, \boldsymbol{X}_i\right] \\
&+ \frac{1 - \Pr\left(D_i(0) = 0, D_i(1) = 1 \,|\, \boldsymbol{X}_i\right)}{\Pr\left(D_i(0) = 0, D_i(1) = 1 \,|\, \boldsymbol{X}_i\right)} \begin{pmatrix} \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i(1) = 0 \text{ or } D_i(0) = 1, \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, \boldsymbol{X}_i\right] \end{pmatrix}
\end{aligned}$$

Collect everything together, as follows.

$$\begin{aligned}
&\mathbb{E}\left[Y_i \,|\, Z_i = z', D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = z', D_i = 0, \boldsymbol{X}_i\right] \\
=&\underbrace{\mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i(1) = 1, D_i(0) = 0, \boldsymbol{X}_i\right]}_{\text{AIE among compliers, conditional on } \boldsymbol{X}_i, Z_i = z'} \\
&+ \underbrace{\mathbb{E}\left[Y_i(z', 0) \,|\, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i(z', 0) \,|\, D_i = 0, \boldsymbol{X}_i\right]}_{\text{Selection bias}} \\
&+ \underbrace{\left[\begin{aligned} &\Big(1 - \Pr\left(D_i = 1 \,|\, \boldsymbol{X}_i\right)\Big) \begin{pmatrix} \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i = 1, \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i = 0, \boldsymbol{X}_i\right] \end{pmatrix} \\ &- \frac{1 - \Pr\left(D_i(0) = 0, D_i(1) = 1 \,|\, \boldsymbol{X}_i\right)}{\Pr\left(D_i(0) = 0, D_i(1) = 1 \,|\, \boldsymbol{X}_i\right)} \begin{pmatrix} \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, D_i(1) = 0 \text{ or } D_i(0) = 1, \boldsymbol{X}_i\right] \\ - \mathbb{E}\left[Y_i(z', 1) - Y_i(z', 0) \,|\, \boldsymbol{X}_i\right] \end{pmatrix} \end{aligned}\right]}_{\text{group difference-bias}}
\end{aligned}$$

The proof is finally achieved by multiplying by the complier score, $\Pr\left(D_i(0) = 0, D_i(1) = 1 \,|\, \boldsymbol{X}_i\right)$ $= \mathbb{E}\left[D_i \,|\, Z_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[D_i \,|\, Z_i = 0, \boldsymbol{X}_i\right]$, then applying the expectation across $Z_i = z'$, and $\boldsymbol{X}_i$.

A.3 A Regression Framework for Direct and Indirect Effects

Put $\mu_{d'}(z'; \boldsymbol{X}) = \mathbb{E}\left[Y_i(z', d') \,|\, \boldsymbol{X}\right]$ and $U_{d',i} = Y_i(z', d') - \mu_{d'}(z'; \boldsymbol{X})$ for each $z', d' = 0, 1$, so we have the following expressions:

$$Y_i(Z_i, 0) = \mu_0(Z_i; \boldsymbol{X}_i) + U_{0,i}, \;\; Y_i(Z_i, 1) = \mu_1(Z_i; \boldsymbol{X}_i) + U_{1,i}.$$

$U_{0,i}, U_{1,i}$ are error terms with unknown distributions, mean independent of Z_i, $\boldsymbol{X}_i$ by definition — but possibly correlated with D_i. Z_i is conditionally independent of potential outcomes, so that $U_{0,i}, U_{1,i} \perp\!\!\!\perp Z_i$.

The first-stage regression of $Z \to Y$ has unbiased estimates, since $Z_i \perp\!\!\!\perp D_i(.) \big| \boldsymbol{X}_i$. Put $\pi(z'; \boldsymbol{X}) = \mathbb{E}\left[D_i(z') \,|\, \boldsymbol{X}\right]$, and $\eta_{z',i} = D_i(z') - \pi(z'; \boldsymbol{X})$ the first-stage error terms.

$$\begin{aligned} D_i &= Z_i D_i(1) + (1 - Z_i) D_i(0) \\ &= D_i(0) + Z_i \left[D_i(1) - D_i(0)\right] \\ &= \quad \underbrace{\pi(0; \boldsymbol{X}_i)}_{\text{Intercept, } := \theta + \zeta(\boldsymbol{X}_i)} \quad + \underbrace{Z_i\big(\pi(1; \boldsymbol{X}_i) - \pi(0; \boldsymbol{X}_i)\big)}_{\text{Regressor, } := \overline{\pi} Z_i} + \underbrace{(1 - Z_i)\eta_{0,i} + Z_i \eta_{1,i}}_{\text{Errors, } := \eta_i} \\ \implies \; \mathbb{E}\left[D_i \,|\, Z_i, \boldsymbol{X}_i\right] &= \theta + \overline{\pi} Z_i + \zeta(\boldsymbol{X}_i). \end{aligned}$$

Since the quasi-random assignment assumption gives $\mathbb{E}\left[Z_i \eta_{z',i} \,|\, \boldsymbol{X}_i\right] = \mathbb{E}\left[Z_i \,|\, \boldsymbol{X}_i\right] \mathbb{E}\left[\eta_{z',i} \,|\, \boldsymbol{X}_i\right] = 0$, for each $z' = 0, 1$. By the same argument Z_i is also assumed independent of potential outcomes $Y_i(., .)$, so that $U_{0,i}, U_{1,i} \perp\!\!\!\perp Z_i$. Thus, the reduced form regression $Z \to Y$ also leads to unbiased estimates for the ATE.

The same cannot be said of the regression that estimates direct and indirect effects, without further assumptions.

$$\begin{aligned} Y_i &= Z_i Y_i(1, D_i(1)) + (1 - Z_i) Y_i(0, D_i(0)) \\ &= Z_i D_i Y_i(1, 1) \\ &\quad + (1 - Z_i) D_i Y_i(0, 1) \\ &\quad + Z_i (1 - D_i) Y_i(1, 0) \\ &\quad + (1 - Z_i)(1 - D_i) Y_i(0, 0) \\ &= Y_i(0, 0) \\ &\quad + Z_i \left[Y_i(1, 0) - Y_i(0, 0)\right] \\ &\quad + D_i \left[Y_i(0, 1) - Y_i(0, 0)\right] \\ &\quad + Z_i D_i \left[Y_i(1, 1) - Y_i(1, 0) - (Y_i(0, 1) - Y_i(0, 0))\right] \end{aligned}$$

And so Y_i can be written as a regression equation in terms of the observed factors and error terms.

$$
\begin{aligned}
Y_i = \; & \mu_0(0; \boldsymbol{X}_i) \\
& + D_i \left[\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i) \right] \\
& + Z_i \left[\mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i) \right] \\
& + Z_i D_i \left[\mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - \left(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i) \right) \right] \\
& + U_{0,i} + D_i \left(U_{1,i} - U_{0,i} \right) \\
= \; & \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) + \left(1 - D_i\right) U_{0,i} + D_i U_{1,i}
\end{aligned}
$$

With the following definitions:

(a) $\alpha = \mathbb{E}\left[\mu_0(0; \boldsymbol{X}_i)\right]$ and $\varphi(\boldsymbol{X}_i) = \mu_0(0; \boldsymbol{X}_i) - \alpha$ are the intercept terms.

(b) $\beta = \mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)$ is the indirect effect under $Z_i = 0$

(c) $\gamma = \mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)$ is the direct effect under $D_i = 0$.

(d) $\delta = \mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - \left(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\right)$ is the interaction effect.

(e) $\left(1 - D_i\right) U_{0,i} + D_i U_{1,i}$ is the remaining error term.

This sequence gives us the resulting regression equation:

$$
\begin{aligned}
\mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i \right] \; = \; & \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\
& + \left(1 - D_i\right) \mathbb{E}\left[U_{0,i} \,|\, D_i = 0, \boldsymbol{X}_i \right] + D_i \mathbb{E}\left[U_{1,i} \,|\, D_i = 1, \boldsymbol{X}_i \right]
\end{aligned}
$$

Taking the conditional expectation, and collecting for the expressions of the direct and indirect effects:

$$
\begin{aligned}
\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \right] &= \mathbb{E}\left[\gamma + \delta D_i \right] \\
\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \right] &= \mathbb{E}\left[\overline{\pi} \left(\beta + Z_i \delta + \widetilde{U}_i \right) \right]
\end{aligned}
$$

These equations have simpler expressions after assuming constant treatment effects in a linear framework; I have avoided this as having compliers, and controlling for observed factors $\boldsymbol{X}_i$ only makes sense in the case of heterogeneous treatment effects.

These terms are conventionally estimated in a simultaneous regression (Imai et al. 2010). If sequential quasi-random assignment does not hold, then the regression estimates from estimating the mediation equations (without adjusting for the contaminated bias term) suffer from omitted variables bias.

$$\mathbb{E}_{\boldsymbol{X}_i}\left[\mathbb{E}\left[Y_i \,|\, Z_i = D_i = 0, \boldsymbol{X}_i\right]\right] = \mathbb{E}\left[\alpha\right] + \mathbb{E}\left[U_{0,i} \,|\, D_i = 0\right]$$

$$\mathbb{E}_{\boldsymbol{X}_i}\left[\mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 0, \boldsymbol{X}_i\right]\right] = \mathbb{E}\left[\beta\right] + \left(\mathbb{E}\left[U_{1,i} \,|\, D_i = 1\right] - \mathbb{E}\left[U_{0,i} \,|\, D_i = 0\right]\right)$$

$$\mathbb{E}_{\boldsymbol{X}_i}\left[\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i = 0, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 0, \boldsymbol{X}_i\right]\right] = \mathbb{E}\left[\gamma\right] + \mathbb{E}\left[U_{0,i} \,|\, D_i = 0\right]$$

$$\mathbb{E}_{\boldsymbol{X}_i}\begin{bmatrix}\mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 1, D_i = 0, \boldsymbol{X}_i\right] \\ - \left(\mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 1, \boldsymbol{X}_i\right] - \mathbb{E}\left[Y_i \,|\, Z_i = 0, D_i = 0, \boldsymbol{X}_i\right]\right)\end{bmatrix} = \mathbb{E}\left[\delta\right]$$

And so the ADE and AIE estimates are contaminated by these bias terms. Additionally, the AIE estimates refers to gains from the mediator among $D(z)$ compliers (not the entire average), so will be biased when not accounting for $\widetilde{U}_i$, too.

A.4 Roy Model and Sequential quasi-random assignment

Proof of Proposition 1.

Suppose Z_i is ignorable, and selection-into-D_i follows a Roy model, with the definitions in Section 3. If selection-into-D_i is degenerate on $U_{0,i}, U_{1,i}$:

$$\mathbb{E}\left[D_i \,|\, Z_i, \boldsymbol{X}_i, U_{1,i} - U_{0,i} = u\right] = \mathbb{E}\left[D_i \,|\, Z_i, \boldsymbol{X}_i, U_{1,i} - U_{0,i} = u'\right], \text{ for all } u, u' \text{ in the range of } U_{1,i} - U_{0,i}.$$

In this case, the control set $\boldsymbol{X}_i$ and the costs $\mu_c, U_{c,i}$ are the only determinants of selection-into-D_i — and, $U_{0,i}, U_{1,i}$ play no role. This could be achieved by either assuming that unobserved gains are degenerate (the researcher had observed everything in $\boldsymbol{X}_i$), or selection-into-D_i had been disrupted in some fashion (e.g., by a natural experiment design for D_i).

To motivate a contraposition argument, suppose D_i is ignorable conditional on $Z_i, \boldsymbol{X}_i$. For each $z', d' = 0, 1$

$$\begin{aligned}
&D_i \perp\!\!\!\perp Y_i(z', d') \;\mid\; \boldsymbol{X}_i, Z_i = z' \\
&\implies D_i \perp\!\!\!\perp \mu_{d'}(z'; \boldsymbol{X}_i) + U_{d',i} \;\mid\; \boldsymbol{X}_i, Z_i = z' \\
&\implies D_i \perp\!\!\!\perp U_{d',i} \;\mid\; \boldsymbol{X}_i, Z_i = z' \\
&\implies D_i \perp\!\!\!\perp U_{1,i} - U_{0,i} \;\mid\; \boldsymbol{X}_i, Z_i = z' \\
&\implies \mathbb{E}\left[D_i \,|\, U_{1,i} - U_{0,i} = u', \boldsymbol{X}_i, Z_i = z'\right] = \mathbb{E}\left[D_i \,|\, \boldsymbol{X}_i, Z_i = z'\right] \\
&\qquad \text{for all } u' \text{ in the range of } U_{1,i} - U_{0,i}.
\end{aligned}$$

This final implication is that selection-into-D_i is degenerate on $U_{0,i}, U_{1,i}$. Thus, a contraposition argument has that if selection-into-D_i is non-degenerate on $U_{0,i}, U_{1,i}$, then D_i is not ignorable.

A.5 Monotonicity $\implies$ Selection Model, in a CM Setting.

Proof that (conditional) monotonicity implies a selection model representation in a CM setting. This proof is an applied example of the Vytlacil (2002) equivalence result, now including conditioning covariates $\boldsymbol{X}_i$, and is presented merely as a validation exercise.

Assume condition monotonicity MTE–1 holds, for any treatment values $z < z'$ and any covariate value $\boldsymbol{X}_i = \boldsymbol{x}$.

$$\Pr\left(D_i(z') \geq D_i(z) \,|\, \boldsymbol{x}\right) = 1.$$

For each value of $\boldsymbol{X}_i = \boldsymbol{x}$ and any treatment values $z < z'$, we first define:

- $\mathcal{A} = \{i : D_i(z) = D_i(z') = 1\}$, always-mediators
- $\mathcal{N} = \{i : D_i(z) = D_i(z') = 0\}$, never-mediators
- $\mathcal{C} = \{i : D_i(z) = 0, D_i(z') = 1\}$, mediator-compliers.

For any mediator complier $i \in \mathcal{C}$, partition the set as follows.

- $\mathcal{Z}_1(i) = \{z' : D_i(z') = 1\}$, treatment values where i takes the mediator
- $\mathcal{Z}_0(i) = \{z' : D_i(z') = 0\}$, treatment values where i doesn't take the mediator.

Note that having binary $Z_i = 0, 1$ reduces this to the simple case of $\mathcal{Z}_0(i) = \{0\}$, and $\mathcal{Z}_1(i) = \{1\}$. The equivalence result holds for continuous values of Z_i, so continue with the more general $\mathcal{Z}_0(i), \mathcal{Z}_1(i)$ notation.

By monotonicity, we have

$$\sup_{z' \in \mathcal{Z}_0(i)} \pi(z'; \boldsymbol{x}) \leq \inf_{z' \in \mathcal{Z}_1(i)} \pi(z'; \boldsymbol{x}), \quad \text{for any } i \in \mathcal{C}$$

where $\pi(z'; \boldsymbol{x}) = \Pr\left(D_i = 1 \,|\, Z_i = z', \boldsymbol{X}_i = \boldsymbol{x}\right)$ is the mediator propensity score. A simple proof by contradiction verifies this statement (Vytlacil 2002, Lemma 1).

Now we construct V_i as follows:

$$V_i = \begin{cases} 1, & \text{if } i \in \mathcal{N} \\ 0, & \text{if } i \in \mathcal{A} \\ \inf_{z' \in \mathcal{Z}_1(i)} \pi(z'; \boldsymbol{x}), & \text{if } i \in \mathcal{C}. \end{cases}$$

Define $\psi(z'; \boldsymbol{x}) = \pi(z'; \boldsymbol{x})$. Then we can represent $D_i(z')$ as a selection model,

$$D_i(z') = \mathbb{1}\left\{\psi(z'; \boldsymbol{X}_i) \geq V_i\right\}, \quad \text{for } z' = 0, 1.$$

We can verify this works:

- For $i \in \mathcal{A}$: $V_i = 0$ and $\psi(z'; \boldsymbol{x}) \geq 0$ for all z', so $D_i(z') = 1$
- For $i \in \mathcal{N}$: $V_i = 1$ and $\psi(z'; \boldsymbol{x}) \leq 1$ for all z', with $\psi(z'; \boldsymbol{x}) < 1$ for $z' \in \mathcal{Z}_0(i)$, so $D_i(z') = 0$ for $z' \in \mathcal{Z}_0(i)$
- For $i \in \mathcal{C}$: $V_i = \inf_{z' \in \mathcal{Z}_1(i)} \pi(z'; \boldsymbol{x})$
  - When $z' \in \mathcal{Z}_1(i)$: $\psi(z'; \boldsymbol{x}) \geq \inf_{z'' \in \mathcal{Z}_1(i)} \pi(z''; \boldsymbol{x}) = V_i$, so $D_i(z') = 1$
  - When $z' \in \mathcal{Z}_0(i)$: $\psi(z'; \boldsymbol{x}) < \inf_{z'' \in \mathcal{Z}_1(i)} \pi(z''; \boldsymbol{x}) = V_i$, so $D_i(z') = 0$.

Therefore, the construction $D_i(z') = \mathbb{1}\left\{\psi(z'; \boldsymbol{X}_i) \geq V_i\right\}$ is a valid representation of the selection process under monotonicity.

This selection model can be transformed to one with a uniform distribution, to get the general selection model of Heckman & Vytlacil (2005). Let $F_V\big(. \,\big|\, \boldsymbol{X}_i\big)$ be the conditional cumulative density function of V_i given $\boldsymbol{X}_i$. Define

$$U_i = F_V\left(V_i \mid \boldsymbol{X}_i\right)$$
$$\pi(z'; \boldsymbol{X}_i) = F_V\left(\psi(z'; \boldsymbol{X}_i) \mid \boldsymbol{X}_i\right) = \Pr\left(D_i = 1 \,|\, Z_i = z', \boldsymbol{X}_i\right)$$

We can then equivalently represent the mediator choice as the transformed selection model

$$D_i(z') = \mathbb{1}\left\{\pi(z'; \boldsymbol{X}_i) \geq U_i\right\}, \quad \text{for } z' = 0, 1$$

where $U_i \,|\, \boldsymbol{X}_i \sim \text{Uniform}(0,1)$ by the probability integral transformation.

A.6 MTE Identification of the Second-stage

Proof of Proposition 2. This proof relies heavily on the notation and reasoning of Kline & Walters (2019) for an IV setting.

By Assumption MTE–1 (mediator monotonicity), selection-into-mediator can be represented as a threshold-crossing selection model.

$$D_i(z') = \mathbb{1}\left\{\pi(z'; \boldsymbol{X}_i) \geq U_i\right\}, \text{ for } z' = 0, 1$$

where $U_i = F_V\left(V_i \mid \boldsymbol{X}_i\right)$ follows a uniform distribution on $[0,1]$, and $\pi(z'; \boldsymbol{X}_i) = \mathbb{E}\left[D_i \,|\, Z_i = z', \boldsymbol{X}_i\right]$ is the mediator propensity score.

The threshold crossing selection model represents individuals who refuse the mediator as follows:

$$D_i = 0 \implies \pi(Z_i; \boldsymbol{X}_i) < U_i$$

Our objective is to determine $\mathbb{E}\left[U_{0,i} \,|\, D_i = 0, Z_i, \boldsymbol{X}_i\right]$, which can then be written as

$$\mathbb{E}\left[U_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i, Z_i, \boldsymbol{X}_i\right].$$

Since Z_i is ignorable, we have:

$$\mathbb{E}\left[U_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i, Z_i, \boldsymbol{X}_i\right] = \mathbb{E}\left[U_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right]$$

Assumption MTE–2 has $\text{Cov}(U_i, U_{0,i}) \neq 0$. This non-zero covariance implies statistical dependence between the selection error and outcome error. This dependence allows us to represent $U_{0,i}$ using a linear projection. We use $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$ rather than U_i directly in the projection to allow for flexibility in how the selection error affects outcomes. The linear projection can be written as follows

$$U_{0,i} = \rho_0\big(F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V\big) + \varepsilon_{0,i},$$

where

- $\mu_V = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)\right]$ is the mean of $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$
- $\rho_0 = \frac{\text{Cov}\left(U_{0,i}, F_V^{-1}(U_i|\boldsymbol{X}_i)\right)}{\text{Var}\left(F_V^{-1}(U_i|\boldsymbol{X}_i)\right)}$ is the projection coefficient
- $\varepsilon_{0,i}$ is a residual with $\mathbb{E}\left[\varepsilon_{0,i} \,\middle|\, F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)\right] = 0$.

The coefficient ρ_0 is the slope in the best linear predictor of $U_{0,i}$ given $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$, and is chosen to ensure that the residual $\varepsilon_{0,i}$ is uncorrelated with $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$. This property is crucial for the identification strategy, as it isolates the component of U_i that is related to selection-into-D_i.

The non-zero covariance condition in MTE–2 ensures $\rho_0 \neq 0$, so is relevant. Since U_i and $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$ are related by a monotonic transformation (the inverse cumulative density function), the covariance $\text{Cov}(U_i, U_{0,i}) \neq 0$ implies $\text{Cov}(F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right), U_{0,i}) \neq 0$.

Given the linear projection of $U_{0,i}$ onto $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$, we can compute the conditional expectation:

$$\mathbb{E}\left[U_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right] = \mathbb{E}\left[\rho_0\big(F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V\big) + \varepsilon_{0,i} \,\middle|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right]$$

Since $\mathbb{E}\left[\varepsilon_{0,i} \,\middle|\, F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)\right] = 0$ by construction, and U_i is a function of $F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right)$, we have

$$\mathbb{E}\left[\varepsilon_{0,i} \,|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right] = 0.$$

Therefore:

$$\mathbb{E}\left[U_{0,i} \,\middle|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right] = \rho_0 \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, \pi(Z_i; \boldsymbol{X}_i) < U_i\right].$$

This gives us the control function representation:

$$\mathbb{E}\left[U_{0,i} \,\middle|\, D_i = 0, Z_i, \boldsymbol{X}_i\right] = \rho_0 \lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big)$$

where $\lambda_0\left(p'\right) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, p' < U_i\right]$. The control function $\lambda_0\left(p'\right)$ captures the expected value of the transformed selection term, conditional on being above the threshold $p' \in (0,1)$.

The same sequence of steps for mediator takers, $D_i = 1$, gives the other mte:

$$\mathbb{E}\left[U_{1,i} \,\middle|\, D_i = 1, Z_i, \boldsymbol{X}_i\right] = \rho_1 \lambda_1\big(\pi(Z_i; \boldsymbol{X}_i)\big),$$

where $\lambda_1\left(p'\right) = \mathbb{E}\left[F_V^{-1}\left(U_i \mid \boldsymbol{X}_i\right) - \mu_V \,\middle|\, U_i \leq p'\right]$ for $p' \in (0,1)$, and $\rho_1 = \frac{\operatorname{Cov}\left(U_{1,i}, F_V^{-1}(U_i \mid \boldsymbol{X}_i)\right)}{\operatorname{Var}\left(F_V^{-1}(U_i \mid \boldsymbol{X}_i)\right)}$ is the corresponding projection coefficient.

The relationship between $\lambda_0(p')$ and $\lambda_1(p')$ can be derived as:

$$\lambda_1\left(p'\right) = -\lambda_0\left(p'\right)\left(\frac{1-p'}{p'}\right), \text{ for } p' \in (0,1).$$

This relationship ensures consistency in the CF approach across the $D_i = 0$ and $D_i = 1$ groups (Kline & Walters 2019).

Assumption MTE–3 (mediator take-up cost instrument $\boldsymbol{X}_i^{\text{IV}}$) ensures identification of the propensity score function $\pi(z'; \boldsymbol{X}_i)$ in the first stage by providing valid instrumental variation. This variation allows us to identify the propensity score, and consequently the control functions λ_0 and λ_1.

Combining all elements, the conditional expectation of Y_i given $Z_i, D_i, \boldsymbol{X}_i$ is

$$\begin{aligned}
\mathbb{E}\left[Y_i \,\middle|\, Z_i, D_i, \boldsymbol{X}_i\right] &= \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\
&\quad + (1 - D_i)\, \mathbb{E}\left[U_{0,i} \,\middle|\, D_i = 0\right] + D_i \mathbb{E}\left[U_{1,i} \,\middle|\, D_i = 1\right].
\end{aligned}$$

Substitute the CFs,

$$\begin{aligned}
&(1 - D_i)\mathbb{E}\left[U_{0,i} \,\middle|\, Z_i, D_i = 0, \boldsymbol{X}_i\right] + D_i \mathbb{E}\left[U_{1,i} \,\middle|\, Z_i, D_i = 1, \boldsymbol{X}_i\right] \\
&= (1 - D_i)\rho_0 \lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big) + D_i \rho_1 \lambda_1\big(\pi(Z_i; \boldsymbol{X}_i)\big).
\end{aligned}$$

This gives the final result,

$$\begin{aligned}\mathbb{E}\left[Y_i \,|\, Z_i, D_i, \boldsymbol{X}_i\right] \;&=\; \alpha + \beta D_i + \gamma Z_i + \delta Z_i D_i + \varphi(\boldsymbol{X}_i) \\ &\quad + \rho_0 \left(1 - D_i\right) \lambda_0\big(\pi(Z_i; \boldsymbol{X}_i)\big) + \rho_1 D_i \lambda_1\big(\pi(Z_i; \boldsymbol{X}_i)\big).\end{aligned}$$

All parameters — $\alpha, \beta, \gamma, \delta, \varphi(.), \rho_0, \rho_1$ — are identified once we control for selection bias through the CFs λ_0, λ_1, with $\pi(z'; \boldsymbol{X}_i)$ identified separately in the first-stage. λ_0, λ_1 can be assumed to be certain functions (say, the inverse Mills ratio in Heckman 1979), or treated as non-parametric parameters to be estimated — at cost of the constant and ρ_0, ρ_1 no longer being separately identified from λ_0, λ_1, see Appendix A.8.

A.7 MTE-associated CF Identification of the ADE and AIE

Proof of Theorem MTE Approach.

Assume MTE–1, MTE–2, MTE–3 hold. Then Proposition 2 has $\alpha, \beta, \gamma, \delta, \varphi(.), \rho_0, \rho_1$ identified in a regression. The following composes the ADE and AIE from these parameters.

For the ADE,

$$\begin{aligned}\mathbb{E}\left[\gamma + \delta D_i\right] &= \mathbb{E}\left[\Big(\mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\Big) + D_i\Big(\mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - \big(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\big)\Big)\right] \\ &= \mathbb{E}\left[D_i\Big(\mu_1(1; \boldsymbol{X}_i) - \mu_1(0; \boldsymbol{X}_i)\Big) + (1 - D_i)\Big(\mu_0(1; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\Big)\right] \\ &= \mathbb{E}\left[D_i\Big(Y_i(1,1) - U_{1,i} - \big(Y_i(0,1) - U_{1,i}\big)\Big) + (1 - D_i)\Big(Y_i(1,0) - U_{0,i} - \big(Y_i(0,0) - U_{0,i}\big)\Big)\right] \\ &= \mathbb{E}\left[D_i\Big(Y_i(1,1) - Y_i(0,1)\Big) + (1 - D_i)\Big(Y_i(1,0) - Y_i(0,0)\Big)\right] \\ &= \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] \\ &= \text{ADE}.\end{aligned}$$

Identification is similar for the AIE, but also involves the complier adjustment term.

$$\begin{aligned}(\rho_1 - \rho_0)\,\Gamma\big(\pi(0; \boldsymbol{X}_i),\, \pi(1; \boldsymbol{X}_i)\big) &= (\rho_1 - \rho_0)\,\frac{\pi(1; \boldsymbol{X}_i)\lambda_1(\pi(1; \boldsymbol{X}_i)) - \pi(0; \boldsymbol{X}_i)\lambda_1(\pi(0; \boldsymbol{X}_i))}{\pi(1; \boldsymbol{X}_i) - \pi(0; \boldsymbol{X}_i)} \\ &= (\rho_1 - \rho_0)\,\mathbb{E}\left[F_V^{-1}(U_i|\boldsymbol{X}_i) - \mu_V \,\middle|\, \pi(0; \boldsymbol{X}_i) < U_i \leq \pi(1; \boldsymbol{X}_i), \boldsymbol{X}_i\right] \\ &= (\rho_1 - \rho_0)\,\mathbb{E}\left[F_V^{-1}(U_i|\boldsymbol{X}_i) - \mu_V \,\middle|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right] \\ &= \mathbb{E}\left[\rho_1\big(F_V^{-1}(U_i|\boldsymbol{X}_i) - \mu_V\big) \,\middle|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right] \\ &\quad - \mathbb{E}\left[\rho_0\big(F_V^{-1}(U_i|\boldsymbol{X}_i) - \mu_V\big) \,\middle|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right] \\ &= \mathbb{E}\left[U_{1,i} - U_{0,i} \,|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right].\end{aligned}$$

This complier adjustment was first presented for an IV setting by Kline & Walters (2019).

Collecting for the AIE,

$$\begin{aligned}
&\mathbb{E}\left[\overline{\pi}\left(\beta + \delta Z_i + (\rho_1 - \rho_0)\Gamma\big(\pi(0; \boldsymbol{X}_i),\, \pi(1; \boldsymbol{X}_i)\big)\right)\right] \\
&= \mathbb{E}\left[\overline{\pi}\left(\Big(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\Big) + Z_i\Big(\mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i) - \big(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\big)\Big)\right)\right] \\
&\quad + \mathbb{E}\Big[\overline{\pi}\, \mathbb{E}\left[U_{1,i} - U_{0,i} \,|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right]\Big] \\
&= \mathbb{E}\left[\overline{\pi}\left(Z_i\Big(\mu_1(1; \boldsymbol{X}_i) - \mu_0(1; \boldsymbol{X}_i)\Big) + (1 - Z_i)\Big(\mu_1(0; \boldsymbol{X}_i) - \mu_0(0; \boldsymbol{X}_i)\Big)\right)\right] \\
&\quad + \mathbb{E}\Big[\overline{\pi}\, \mathbb{E}\left[U_{1,i} - U_{0,i} \,|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right]\Big] \\
&= \mathbb{E}\left[\overline{\pi}\left(\mu_1(Z_i, \boldsymbol{X}_i) - \mu_0(Z_i, \boldsymbol{X}_i) + \mathbb{E}\left[U_{1,i} - U_{0,i} \,|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right]\right)\right] \\
&= \mathbb{E}\left[\overline{\pi}\, \mathbb{E}\left[\mu_1(Z_i, \boldsymbol{X}_i) - \mu_0(Z_i, \boldsymbol{X}_i) + U_{1,i} - U_{0,i} \,|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[D_i(1) - D_i(0) \,|\, \boldsymbol{X}_i\right]\ \mathbb{E}\left[Y_i(Z_i, 1) - Y_i(Z_i, 0) \,|\, D_i(0) = 0, D_i(1) = 1, \boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \,|\, \boldsymbol{X}_i\right]\right] \\
&= \mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0))\right] \\
&= \text{AIE}.
\end{aligned}$$

A.8 Semi-parametric Estimation of the AIE

It is difficult to directly use the CFs to compose estimates of the complier adjustment term, because various intercepts lose identification, but also because trusting semi-parametric estimates at individual points across the $\widehat{\lambda}_0(p'), \widehat{\lambda}_1(p')$ functions would increase variation more than is necessary.

This can be avoided by noting the relation between the ATE and the conditional ADE and conditional AIE. The following showing how to identify the AIE via relation to the ATE and conditional ADE, and omits the conditional on $\boldsymbol{X}_i$ for brevity.

A simple algebraic rearrangement has the following (as first noted in Imai et al. 2010, Section 3.1),

$$\begin{aligned}
\text{ATE} &= \mathbb{E}\left[Y_i(1, D_i(1)) - Y_i(1, D_i(1))\right] \\
&= \mathbb{E}\left[Y_i(1, D_i(1)) - Y_i(0, D_i(1))\right] + \mathbb{E}\left[Y_i(0, D_i(1)) - Y_i(0, D_i(0))\right] \\
&= \underbrace{\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, Z_i = 1\right]}_{\text{ADE conditional on } Z_i=1} + \underbrace{\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \,|\, Z_i = 0\right]}_{\text{AIE conditional on } Z_i=0}.
\end{aligned}$$

A similar re-arrangement also has the following,

$$\text{ATE} = \underbrace{\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \,|\, Z_i = 1\right]}_{\text{AIE conditional on } Z_i=1} + \underbrace{\mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i)) \,|\, Z_i = 0\right]}_{\text{ADE conditional on } Z_i=0}.$$

Reverting to the regression notation, to show how the ADE conditional on Z_i is identified:

$$\begin{aligned}
\text{ADE} &= \mathbb{E}\left[Y_i(1, D_i(Z_i)) - Y_i(0, D_i(Z_i))\right] \\
&= \mathbb{E}\left[\gamma + \delta D_i(Z_i)\right]. \\
\implies \text{ADE conditional on } Z_i = 0 &= \mathbb{E}\left[\gamma + \delta D_i(Z_i) \,|\, Z_i = 0\right] \\
&= \mathbb{E}\left[\gamma + \delta D_i(0)\right]. \\
\text{ADE conditional on } Z_i = 1 &= \mathbb{E}\left[\gamma + \delta D_i(Z_i) \,|\, Z_i = 1\right] \\
&= \mathbb{E}\left[\gamma + \delta D_i(1)\right].
\end{aligned}$$

Finally achieve identification of the AIE via the ATE and conditional ADE, as follows,

$$\begin{aligned}
\text{AIE} = &\Pr\left(Z_i = 0\right) \underbrace{\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \,|\, Z_i = 0\right]}_{\text{AIE conditional on } Z_i=0} \\
&+ \Pr\left(Z_i = 1\right) \underbrace{\mathbb{E}\left[Y_i(Z_i, D_i(1)) - Y_i(Z_i, D_i(0)) \,|\, Z_i = 1\right]}_{\text{AIE conditional on } Z_i=1} \\
= &\Pr\left(Z_i = 0\right) \Big[\text{ATE} - \left(\text{ADE conditional on } Z_i = 1\right)\Big] \\
&+ \Pr\left(Z_i = 1\right) \Big[\text{ATE} - \left(\text{ADE conditional on } Z_i = 0\right)\Big] \\
= &\,\text{ATE} - \Pr\left(Z_i = 0\right) \mathbb{E}\left[\gamma + \delta D_i(1)\right] - \Pr\left(Z_i = 1\right) \mathbb{E}\left[\gamma + \delta D_i(0)\right].
\end{aligned}$$

The semi-parametric AIE estimate then uses this representation, avoiding directly interacting with the estimated CFs, by plugging in estimates $\widehat{\Pr}(Z_i = 1) = \overline{Z}$, $\widehat{\text{ATE}}$, and the estimates from each side of the $D_i = 0, 1$ separated samples $\widehat{\gamma}, \widehat{\delta}$.

$$\widehat{\text{AIE}}^{\text{CF}} = \widehat{\text{ATE}} - (1 - \overline{Z}) \left(\widehat{\gamma} + \frac{1}{n}\sum_{i=1}^{N} \widehat{\delta}\, \widehat{\pi}(1; \boldsymbol{X}_i)\right) - \overline{Z} \left(\widehat{\gamma} + \frac{1}{n}\sum_{i=1}^{N} \widehat{\delta}\, \widehat{\pi}(0; \boldsymbol{X}_i)\right),$$

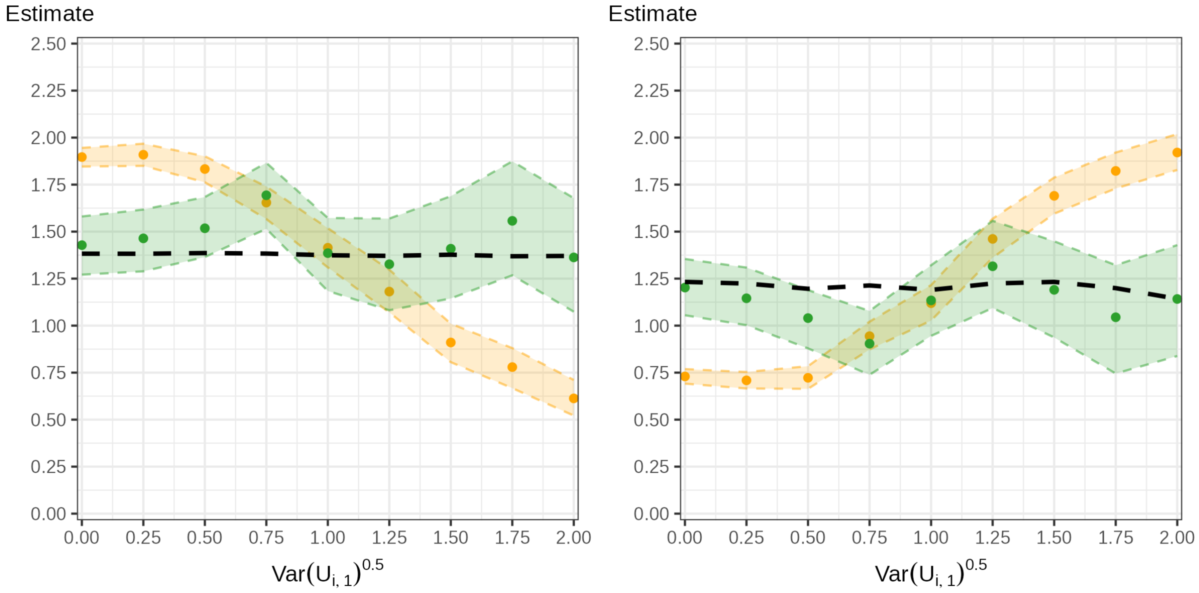
where $\frac{1}{n}\sum_{i=1}^{N} \widehat{\delta}\, \widehat{\pi}(0; \boldsymbol{X}_i)$ estimates $\mathbb{E}\left[\delta D_i(0)\right]$, and $\frac{1}{n}\sum_{i=1}^{N} \widehat{\delta}\, \widehat{\pi}(0; \boldsymbol{X}_i)$ estimates $\mathbb{E}\left[\delta D_i(1)\right]$. Everything involved is a standard point estimate, so their composition will converge to a normal distribution, too. Standard error computation can be achieved by a bootstrap procedure.

A.9 Implementation and Further Simulation Evidence

A number of statistical packages, for the R language (R Core Team 2025), made the simulation analysis for this paper possible.

- *Tidyverse* (Wickham, Averick, Bryan, Chang, McGowan, François, Grolemund, Hayes, Henry, Hester, Kuhn, Pedersen, Miller, Bache, Müller, Ooms, Robinson, Seidel, Spinu, Takahashi, Vaughan, Wilke, Woo & Yutani 2019) collected tools for data analysis in the R language.
- *Mgcv* (Wood, N., Pya & S"afken 2016) allows semi-parametric estimation, using splines, in the R language.
- *Mediate* (Tingley, Yamamoto, Hirose, Keele & Imai 2014) automates the sequential quasi-random assignment estimates of CM effects (Imai et al. 2010) in the R language.

Figure A1: OLS versus MTE-based Estimates of CM Effects, varying Var $(U_{1,i})$ relative to Var $(U_{0,i}) = 1$.



(a) ADE.

(b) AIE.

Note: These figures show the OLS and MTE-based estimates of the ADE and AIE, for $n = 5,000$ sample size. The black dashed line is the true value, points are points estimates from data simulated with a given Corr$\left(U_{0,i}, U_{1,i}\right) = 0.5$, Var $(U_{0,i}) = 1$, and Var $(U_{1,i})^{\frac{1}{2}}$ varied across $[0, 2]$. Shaded regions are the 95% confidence intervals; orange are the OLS estimates, green the semi-parametric MTE-based approach.