NBER WORKING PAPER SERIES

RECONCILING SEEMINGLY CONTRADICTORY RESULTS FROM THE OREGON HEALTH INSURANCE EXPERIMENT AND THE MASSACHUSETTS HEALTH REFORM

Amanda E. Kowalski

Working Paper 24647
http://www.nber.org/papers/w24647

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2018, Revised February 2019

Previously circulated as "Extrapolation using Selection and Moral Hazard Heterogeneity from within the Oregon Health Insurance Experiment." This paper includes material from Kowalski (2016). I co-developed the Stata command mtemore to accompany Kowalski (2016) and the Stata command mtebinary to accompany this paper. I thank Saumya Chatrath, Neil Christy, Aigerim Kabdiyeva, Samuel Moy, Srajal Nayak, Ljubica Ristovska, Sukanya Stravasti, and Matthew Tauzer for excellent research assistance. Joseph Altonji, John Asker, Steve Berry, Christian Brinch, Lasse Brune, Pedro Carneiro, Raj Chetty, Joseph Doyle, Mark Duggan, Caroline Hoxby, Liran Einav, Amy Finkelstein, Matthew Gentzkow, Jonathan Gruber, John Ham, Guido Imbens, Dean Karlan, Larry Katz, Pat Kline, Michal Kolesar, Jonathan Levin, Rebecca McKibbin, Sarah Miller, Costas Meghir, Magne Mogstad, Mark Rosenzweig, Joseph Shapiro, Orie Shelef, Ashley Swanson, Eva Vivalt, Ed Vytlacil, David Wilson, and seminar participants at Academia Sinica, AEA Annual Meeting, Annual Health Econometrics Workshop, Berkeley, BU/MIT/Harvard Health Economics, CHES, Chicago Harris, Dartmouth, Duke Fuqua, IFS, LSE, Michigan, NBER Summer Institute, Northwestern, Ohio State, Princeton, Rand, Santa Clara, SMU, Stanford, Stanford GSB, Stanford SITE, Stockholm, UBC, UC Davis, UC Irvine, UConn Development Conference, UCLA Anderson, USC, UT Austin, Yale, Wharton, Wisconsin, and WEAI provided helpful comments. NSF CAREER Award 1350132 and the Stanford Institute for Economic Policy Research (SIEPR) provided support. This project uses data from the Oregon Health Insurance Experiment, AEARCTR-0000028. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform
Amanda E. Kowalski
NBER Working Paper No. 24647
May 2018, Revised February 2019
JEL No. C1,H75,I10,I13

ABSTRACT

I aim to shed light on why emergency room (ER) utilization increased following the Oregon Health Insurance Experiment but decreased following a Massachusetts policy. To do so, I unite the literatures on insurance and treatment effects. Under an MTE model that assumes no more than the LATE assumptions, comparisons across always takers, compliers, and never takers can inform the impact of polices that expand and contract coverage. Starting from the Oregon experiment as the "gold standard," I make comparisons within Oregon and extrapolate my findings to Massachusetts. Within Oregon, I find adverse selection and heterogeneous moral hazard. Although previous enrollees increased their ER utilization, evidence suggests that subsequent enrollees will be healthier, and they will decrease their ER utilization. Accordingly, I can reconcile the Oregon and Massachusetts results because the Massachusetts policy expanded coverage from a higher baseline, and new enrollees reported better health.

Amanda E. Kowalski
Department of Economics
University of Michigan
611 Tappan Ave.
Lorch Hall 213
Ann Arbor, MI 48109-1220
and NBER
aekowals@umich.edu

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/28

1 Introduction

Findings from the Oregon Health Insurance Experiment are considered the "gold standard" for evidence in health economics because they are based on a randomized lottery. The state of Oregon conducted the lottery in 2008 as a fair way to expand eligibility for its Medicaid health insurance program to a limited number of uninsured individuals. The lottery also effectively created a randomized experiment that facilitated evaluation of the impact of expanding health insurance coverage.

A headline finding from the Oregon experiment is that health insurance coverage increased emergency room (ER) utilization (Taubman et al., 2014). Legislation requires that emergency rooms see all patients regardless of coverage, so the uninsured often access the healthcare system through the ER. There was hope that coverage would *decrease* ER utilization, either because of substitution toward primary care or because of improved health. However, it is plausible that coverage increased ER utilization because formerly uninsured individuals could visit the ER at lower personal cost after gaining coverage. The sign and magnitude of the treatment effect of insurance coverage on ER utilization are important for policy evaluation because care provided in the ER is expensive, but the insured do not necessarily value additional ER care at its cost.

The finding that ER utilization increased in Oregon was particularly surprising because previous evidence from an expansion of insurance coverage due to the Massachusetts health reform of 2006 showed that ER utilization decreased or stayed the same (Chen et al., 2011; Smulowitz et al., 2011; Kolstad and Kowalski, 2012; Miller, 2012). Unlike the Oregon policy, the Massachusetts reform was a natural experiment that did not involve randomization. Therefore, it is tempting to dismiss results based on the Massachusetts reform and to focus on results from Oregon as the definitive answer. Discussion of the Oregon experiment and the Massachusetts reform in the *New York Times* has done just that (Tavernise, 2014).

I start from the premise that when results from two experiments give different answers, it need not be the case that one experiment must be flawed. Instead, it could be the case that each experiment yields a different local average treatment effect (LATE), in the terminology of Imbens and Angrist (1994). If each LATE is derived from the same underlying marginal treatment effect (MTE) function, as introduced by Björklund and Moffitt (1987) and developed by Heckman and Vytlacil (1999, 2001, 2005), Carneiro et al. (2011), and Brinch et al. (2017), then it could be possible to use that MTE function to recover the two different LATEs, thereby reconciling the results. Although the MTE literature generally focuses on a single context, I aim to use treatment effect heterogeneity that I find within the Oregon context to reconcile results across the Oregon and Massachusetts contexts.

To do so, I begin with an MTE model shown by Vytlacil (2002) to assume no more than the LATE assumptions of independence and monotonicity proposed by Imbens and Angrist (1994). In my exposition of the model, I emphasize the link between the MTE and always takers, compliers, and never takers, using the terminology of Angrist et al. (1996). In that terminology, the LATE is the average treatment effect on "compliers" who gain coverage if and only if they win the lottery. However, the MTE function also characterizes the treatment effects on "always takers" who gain coverage regardless of the lottery outcome and "never takers" who do not gain coverage regardless of the lottery outcome.

I use simple figures derived from the MTE model to make clear that the LATE assumptions imply an ordering from always takers to compliers to never takers, originally shown by Vytlacil (2002). The intuition behind the ordering is simple. Always takers are individuals that are already eligible for coverage under the existing policy, compliers are individuals that become eligible for coverage if they win the lottery implemented by the new policy, and never takers are remaining individuals do not become eligible for coverage if they win the lottery implemented by the new policy. Future policies that expand coverage could enroll never takers, and future policies that contract coverage could disenroll always takers. Therefore, even though the treatment effect on compliers is relevant for the policy implemented by the experiment, treatment effects on always and never takers could be relevant for future policies. Treatment effects on always and never takers could also be relevant for policies in other contexts.

To reconcile the LATE from the Oregon context with the LATE from the Massachusetts context, I proceed in three steps. First, starting with the Oregon experiment as the "gold standard," I assess whether I find heterogeneity across the unobservable that separates always takers, compliers, and never takers. Second, I use evidence from Massachusetts to assess whether heterogeneity across the unobservable within Oregon can reconcile the Oregon and Massachusetts LATEs. Third, I assess whether observables alone can explain heterogeneity and reconcile the Oregon and Massachusetts LATEs.

As the first step, I estimate the model using publicly-available data from the Oregon experiment (Finkelstein, 2013).[1] Within my analysis sample, I replicate a positive LATE, which shows that the average treatment effect of insurance on ER utilization is positive for compliers. However, only 26% of individuals are compliers, while 15% are always takers and 59% are never takers. By making comparisons across these groups under the MTE model that assumes no more than the LATE assumptions, I find heterogeneous selection into coverage. Specifically, compliers are adversely selected relative to never takers in the sense that they use the ER more in the absence of insurance. Under transparent ancillary

[1]Publicly available data are rare in health economics, because many sources of data are proprietary and confidential. I am grateful to the investigators of the Oregon Health Insurance Experiment for making their data available. By using publicly available data, I encourage replication and future work.

assumptions, I find treatment effect heterogeneity. Specifically, I find a downward-sloping MTE function. The downward slope indicates that the treatment effect of insurance on ER utilization decreases as enrollment increases. It is so pronounced that even though the average treatment effect on compliers is positive, the average treatment effect on never takers is negative.

As the second step in my reconciliation of the Oregon and Massachusetts LATEs, I bring in evidence from Massachusetts. Recasting my previous work on the Massachusetts reform from Hackmann et al. (2015) in terms of the MTE model with ancillary assumptions, I show that the MTE function within Massachusetts is also downward-sloping. Given that I find downward-sloping MTE functions within Oregon and Massachusetts, I use data from Kolstad and Kowalski (2012) to characterize the Massachusetts reform as an expansion of coverage along the Oregon MTE. Because enrollment levels were high in Massachusetts before the reform, I predict that Massachusetts compliers respond to insurance like a subset of Oregon never takers. By re-weighting the Oregon MTE to attain a Massachusetts LATE, I predict a decrease in ER utilization in Massachusetts of the same order of magnitude as the decrease found by Miller (2012). MTE-reweighting thus offers a plausible pathway to reconcile the increase in ER utilization found in Oregon with the decrease in ER utilization found in Massachusetts.

As the third step, I examine observables to assess whether I can reconcile the Oregon and Massachusetts LATEs using observables alone. I begin by examining self-reported health, which is elicited as excellent, very good, good, fair, or poor. Finkelstein et al. (2012) shows that individuals who won the lottery reported better self-reported health, so I only compare the self-reported health of groups without coverage: compliers who lost the lottery and never takers. I find that 55% of Oregon compliers who lost the lottery report fair or poor health, while only 34% of Oregon never takers report fair or poor health. The difference is statistically different from zero, indicating adverse selection on self-reported health. I also find suggestive evidence of adverse selection on self-reported health within Massachusetts. However, the difference between Massachusetts and Oregon is even more striking than the difference within Massachusetts: only 21% of Massachusetts compliers report fair or poor health, which suggests that Massachusetts compliers are healthier than Oregon compliers. These comparisons suggest an important mechanism for my findings—individuals in worse health gain coverage in early expansions and increase their ER utilization upon gaining coverage, but individuals in better health gain coverage in later expansions and decrease their ER utilization upon gaining coverage. However, I cannot directly test this mechanism by including self-reported heath in the Oregon MTE because self-reported health is only observed with coverage for always takers.

Therefore, I turn to a different observable, ER utilization from before the lottery took place, which is correlated with self-reported health and available for all individuals within the Oregon data. Before the lottery took place, always takers visited the ER more than compliers, who visited the ER more than never takers, indicating adverse selection. When I include previous ER utilization in the Oregon MTE, I can explain all of the heterogeneity in the treatment effect. Therefore, differences in previous ER utilization between Oregon compliers and Massachusetts compliers could explain the entire difference between the positive Oregon LATE and the negative Massachusetts LATE. Unfortunately, I do not observe previous ER utilization in my Massachusetts data, so I cannot use it directly to reconcile the Oregon and Massachusetts LATEs.

Finally, I turn to the three common observables available in the Oregon and Massachusetts data – age, gender, and English-speaking status – and explore whether I can use them to reconcile the Oregon and Massachusetts LATEs. I cannot reconcile the LATEs using LATE-reweighting following Angrist and Fernandez-Val (2013) and Hotz et al. (2005). This result is not surprising. LATE-reweighting compares compliers with different values of the common observables, but my analysis of the Oregon MTE shows that the meaningful treatment effect heterogeneity is across the unobservable that separates always takers from compliers from never takers, not across compliers with different values of the common observables. MTE-reweighting effectively allows me to extrapolate from Oregon to Massachusetts using an unobservable that captures previous ER utilization and health, as well as the common observables. With MTE-reweighting, I can reconcile the positive LATE from Oregon with the negative LATE from Massachusetts, and I obtain an extrapolated Massachusetts LATE that is comparable in magnitude to the estimate from Miller (2012).

2 Model

I begin with a model shown by Vytlacil (2002) to assume no more than the LATE assumptions. To ensure that I do not introduce additional assumptions, I follow the exposition from Heckman and Vytlacil (2005) closely. However, I adapt the model for my empirical context, and I try to build intuition using simple figures.

2.1 First Stage: Enrollment

Let the observed binary variable D represent enrollment in Medicaid, which is the "treatment" offered by the Oregon Health Insurance Experiment. Let V_T represent potential utility in the treated state (enrolled in Medicaid, $D = 1$), and let V_U represent potential utility in the untreated state (not enrolled in Medicaid, $D = 0$). The following definition relates realized utility V to the potential utilities:

$$V = V_U + (V_T - V_U)D. \tag{1}$$

I specify the net benefit of treatment in terms of the potential utilities as follows:

$$V_T - V_U = \mu_D(Z, X) - \nu_D, \tag{2}$$

where $\mu_D(\cdot)$ is an unspecified function, Z is an observed binary instrument, X is an optional observed vector of covariates, and ν_D is an unobserved term with an unspecified distribution. In the Oregon context, Z represents the outcome of the randomized lottery. Individuals with $Z = 0$ are lottery losers. I refer to them as the "control group." Individuals with $Z = 1$ are lottery winners. I refer to them as the "intervention group" because they receive the intervention, an opportunity to be eligible for Medicaid. I need different terminology for the intervention group ($Z = 1$) and the treated group ($D = 1$) because not all Oregon lottery winners enroll in Medicaid. To derive an equation for treatment as a function of the lottery outcome, I assume

A.1. (Continuity) The cumulative distribution function of ν_D conditional on X, which I denote with $F(\cdot \mid X)$, is absolutely continuous with respect to the Lebesgue measure.

A.2. (Independence) The random vectors (ν_D,γ_T) and (ν_D,γ_U) are independent of Z conditional on X, where γ_T and γ_U are unobserved terms introduced in the second stage.

A.3. (Instrument Relevance) $\mu_D(Z, X)$ is a nondegenerate random variable conditional on X.

Under A.1, the transformation of ν_D by the function $F(\cdot \mid X)$ is a normalization that yields $U_D = F(\nu_D \mid X)$, which is uniformly distributed between 0 and 1, as I show for completeness in Appendix A. Since ν_D enters negatively into the net benefit of treatment, I interpret U_D as the normalized "unobserved net cost of treatment." The further imposition of A.2 implies the following treatment equation, which states that individuals are treated if their unobserved net cost of treatment is weakly less than a threshold:

$$D = 1\{U_D \leq \mathrm{P}(D = 1 \mid Z = z, X)\}. \tag{3}$$

I show the derivation in Appendix B for completeness. Under A.3, the threshold is different for lottery winners and losers with the same vector of covariates X, which yields the following two special cases:

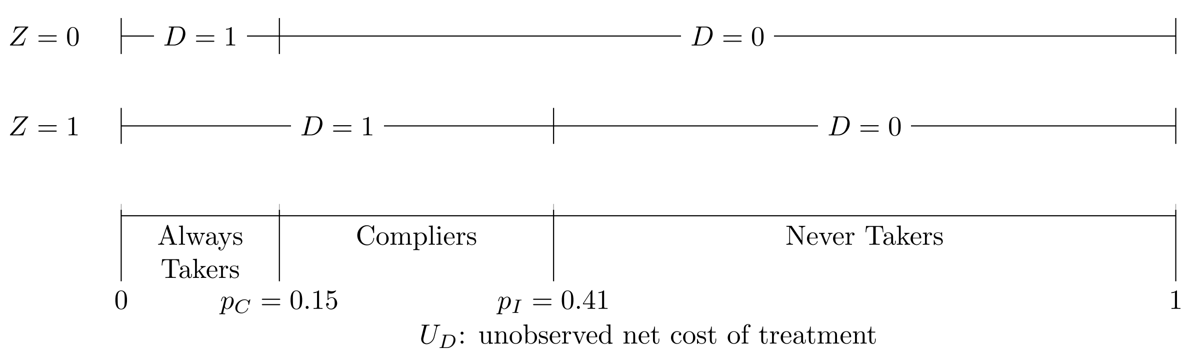
$$D = 1\{U_D \leq p_{CX}\} \qquad \text{where } p_{CX} = P\big(D = 1 \mid Z = 0, X\big), \tag{4}$$

$$D = 1\{U_D \leq p_{IX}\} \qquad \text{where } p_{IX} = P\big(D = 1 \mid Z = 1, X\big). \tag{5}$$

where p_{CX} is the probability of treatment in the control group conditional on X, and p_{IX} is the probability of treatment in the intervention group conditional on X.

As I show in Figure 1, these two special cases of the treatment equation allow me to identify three distinct ranges of the unobserved net cost of treatment, U_D. As originally shown by Vytlacil (2002), the three ranges of U_D correspond to ranges for always takers, compliers, and never takers. Within my analysis sample from the Oregon experiment, 15% of lottery losers enroll and 41% of lottery winners enroll. Accordingly, in Figure 1, I depict $p_C = 0.15$ and $p_I = 0.41$, suppressing X to emphasize that these quantities are averages in the full analysis sample, not in a sample conditional on X. In the top line of Figure 1, I depict the lottery losers. By (4), I infer that treated enrolled lottery losers have $0 \leq U_D \leq 0.15$. Treated lottery losers must be always takers because always takers are treated regardless of the lottery outcome. In the middle line of Figure 1, I depict the lottery winners. By (5), I infer that the untreated lottery winners have $0.41 < U_D \leq 1$. Untreated lottery winners must be never takers because never takers are untreated regardless of the lottery outcome. In the bottom line of Figure 1, I depict U_D for lottery losers and winners on the same axis, and I label the implied ranges of U_D for always and never takers. Individuals with values of U_D in the middle range, $0.15 < U_D \leq 0.41$, enroll in Medicaid if they win the lottery, but they do not enroll if they lose the lottery. These individuals must be compliers because compliers receive treatment if and only if they win the lottery.

Figure 1: Ranges of U_D for Always Takers, Compliers, and Never Takers



I emphasize that the ordering from always takers to compliers to never takers along U_D is an ordering across an important margin: the margin of enrollment in Medicaid. As Medicaid enrollment expands, always takers enroll first, followed by compliers, followed by never takers.

There could be several mechanisms for this ordering, and all of those mechanisms are captured by the unobserved term U_D. In the Oregon experiment, individuals entered the experiment by joining a waitlist for Medicaid, but they were only required to provide eligibility documentation if they won. Therefore, some individuals who were already eligible for Medicaid signed up for the lottery, perhaps because they were not aware of their eligibility, and these individuals could become always takers. On the other side of the spectrum, some individuals did not enroll in Medicaid even if they won, either because they were ineligible or because they did not submit eligibility information in the required timeframe. Therefore, U_D could reflect eligibility, the submission of eligibility information, or other correlated factors. However, the model does not require me to specify what is included in U_D. Instead, it gives me a framework to think about and examine empirically what factors separate always takers from compliers and never takers. As part of that framework, I can consider their ER utilization in the second stage.

2.2 Second Stage: ER Utilization

I relate Medicaid enrollment D to realized ER utilization Y as follows:

$$Y = Y_U + (Y_T - Y_U)D, \tag{6}$$

where I specify potential ER utilization with Medicaid Y_T and without Medicaid Y_U as follows:

$$Y_T = g_T(X, U_D, \gamma_T) \tag{7}$$

$$Y_U = g_U(X, U_D, \gamma_U), \tag{8}$$

where $g_U(\cdot)$ and $g_T(\cdot)$ are unspecified functions that need not be additively separable in their observed and unobserved components,[2] X is the same optional vector of observed covariates from the first stage, U_D is the normalized unobserved net cost of treatment from the first stage, and γ_T and γ_U represent additional unobserved terms with unspecified distributions in the second stage. To make sure that average treated and untreated potential outcomes are defined for each X, I assume:

A.4. (Treated and Untreated) $0 < \mathrm{P}(D = 1 \mid X) < 1$.

A.5. (Second Stage Technical Assumption) The values of $E[Y_T]$ and $E[Y_U]$ are finite.

[2] Vytlacil (2002) shows that the additive separability of the observed and unobserved components of (2) implies the LATE monotonicity assumption of Imbens and Angrist (1994) in the first stage. The LATE assumptions do not include a similar monotonicity assumption in the second stage.

As a whole, because I have only made stylistic changes to the model presented by Heckman and Vytlacil (2005), by the proof of Vytlacil (2002), the model, given by the utility equations (1) and (2), the treatment equations (3)–(5), the potential outcome equations (6)–(8), and assumptions A.2–A.5, assumes no more than the LATE assumptions.

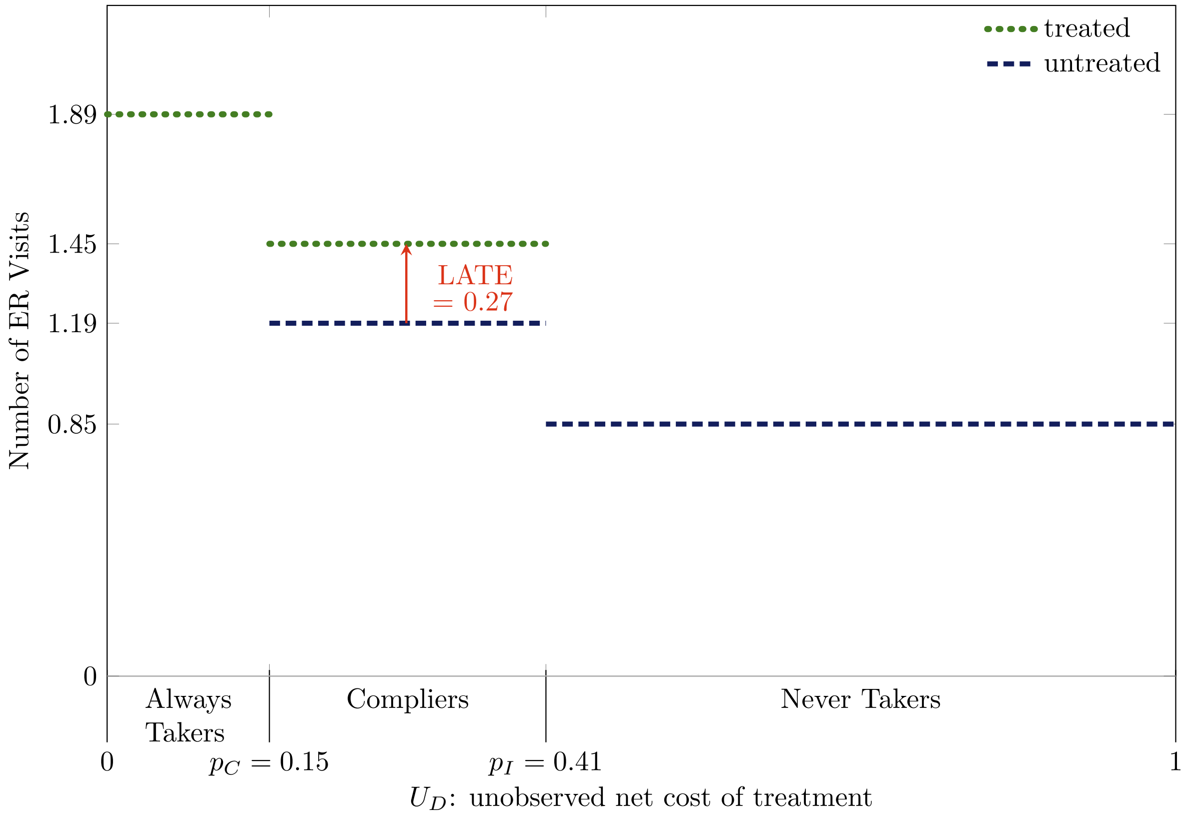
Under the model and the equivalent LATE assumptions, it is not possible to identify any individual as a complier, but it is possible to derive the average treated and untreated outcomes of compliers. It is also possible to derive the average treated outcome for always takers and the average untreated outcome for never takers. However, it is not possible to derive the average untreated outcome for always takers or the average treated outcome for never takers without further assumptions because always takers are always treated and never takers are never treated within the experiment. In Appendix C, I use the model to derive the average treated outcomes for always takers and compliers, and average untreated outcomes for compliers and never takers.[3] My derivation is consistent with the derivations of Imbens and Rubin (1997), Katz et al. (2001), Abadie (2002), and Abadie (2003), which rely on the LATE assumptions.

I use the average treated and untreated outcomes that I derive from the Oregon experiment to illustrate the implications of the model graphically in Figure 2. Along the vertical axis, I depict average ER utilization after the lottery took place from March 10, 2008 to September 30, 2009. I show that during that period, always takers visited the ER 1.89 times, compliers visited 1.45 times if enrolled and 1.19 times if not, and never takers visited 0.85 times. The difference in visits between treated and untreated compliers is equal to the LATE, as shown by Imbens and Rubin (1997). I depict the LATE with an arrow to indicate that it has magnitude and direction. The positive LATE of 0.27 is consistent the headline finding of Taubman et al. (2014), who show that insurance increases ER utilization for compliers.[4]

[3]The derivation relies on average ER utilization for four observed groups: lottery losers with Medicaid (always takers only), lottery winners with Medicaid (always takers and compliers with Medicaid), lottery losers without Medicaid (never takers and compliers without Medicaid), and lottery winners without Medicaid (never takers only). Because of randomization, average ER utilization of lottery losers with Medicaid identifies average ER utilization with Medicaid for all always takers, even the lottery winners. Similarly, average ER utilization of lottery winners without Medicaid identifies average ER utilization without Medicaid for all never takers. Furthermore, the fraction of always takers among lottery losers and the fraction of never takers among lottery winners identify the respective fractions in the full sample. Using these fractions and average ER utilization for always takers with Medicaid and never takers without Medicaid, it is straightforward to back out average ER utilization for compliers with and without Medicaid from the average ER utilization for lottery winners with Medicaid and lottery losers without Medicaid. (It is not possible to calculate average ER utilization for always takers *without* Medicaid or never takers *with* Medicaid without ancillary assumptions because these groups do not change their enrollment based on the lottery.)

[4]I am able to replicate the LATE of 0.41 reported by Taubman et al. (2014), almost exactly, limited only by minor changes made to the publicly available data to hinder identification of individuals with large and uncommon numbers of ER visits. However, that LATE is obtained from a regression that includes controls for previous ER utilization as well as the number of lottery entrants from a household. It would not be valid

Figure 2: Number of ER Visits for Always Takers, Compliers, and Never Takers



Note. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. p_C is the probability of treatment in the control group, and p_I is the probability of treatment in the intervention group. Some differences between statistics might not appear internally consistent because of rounding.

Figure 2 provides more information than the LATE alone. As originally shown by Angrist (1990) and Angrist and Krueger (1992), the calculation of the LATE does not require the ability to calculate the average treated and untreated outcomes of compliers depicted in Figure 2. Using the Wald (1940) approach, the reduced form $\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]$ is equal to 0.07, and the first stage $\mathrm{E}[D|Z=1] - \mathrm{E}[D|Z=0]$ is equal to 0.26. Dividing the reduced form by the first stage yields a LATE of 0.27 visits, which is equal to the LATE reported in Figure 2. However, Figure 2 also includes average outcomes for always and never

to obtain a LATE without any control for the number of lottery entrants because the probability of winning the lottery was only random conditional on the number of entrants. Therefore, I control for the number of lottery entrants nonparametrically by restricting my analysis sample to the 19,643 individuals that were the only members of their household to enter the lottery from the full sample of 24,646 individuals with administrative data on their visits to the ER. By doing so and excluding controls for previous ER utilization for simplicity, I obtain a smaller, but still positive, LATE. The focus of my work is on reconciling a positive LATE in Oregon with a negative LATE in Massachusetts, not on evaluating the Oregon experiment or previous analysis of it, which has been discussed in Baicker et al. (2013, 2014); Taubman et al. (2014), and Finkelstein et al. (2016).

takers, which are not required to calculate the LATE. If these outcomes are different from the comparable outcomes for compliers, then there could be reason to question whether the LATE applies to always and never takers. Such differences could reflect selection or treatment effect heterogeneity.

2.3 Definitions of Selection and Treatment Effect Heterogeneity

I define selection and treatment effect heterogeneity along U_D using the following functions:

$$\begin{aligned}
\text{Selection Heterogeneity:} \quad & \mathrm{MUO}(x,p) = \mathrm{E}\left[Y_U \mid X = x, U_D = p\right] \\
\text{Selection + Treatment Effect Heterogeneity:} \quad & \mathrm{MTO}(x,p) = \mathrm{E}\left[Y_T \mid X = x, U_D = p\right] \\
\text{Treatment Effect Heterogeneity:} \quad & \mathrm{MTE}(x,p) = \mathrm{E}\left[Y_T - Y_U \mid X = x, U_D = p\right],
\end{aligned}$$

where x is a realization of the covariate vector X and p is a realization of the unobserved net cost of treatment U_D.

I refer to the first function as the "marginal untreated outcome (MUO)" function, and I use it to define "selection heterogeneity," a term that I use to capture positive and negative selection, also referred to as "adverse" and "advantageous" selection in the insurance literature. The MTE literature uses the MUO function as an intermediate function in the derivation of the third function, the "marginal treatment effect (MTE)" function of Heckman and Vytlacil (1999, 2001, 2005). However, the literature does not use the MUO function to define selection heterogeneity (see Carneiro and Lee, 2009; Brinch et al., 2017). Instead, the MTE literature and the LATE literature focus on the following definition of "selection bias" (see Heckman et al., 1998; Angrist, 1998):

$$\text{Selection Bias: } \mathrm{E}[Y_U \mid D = 1] - \mathrm{E}[Y_U \mid D = 0]. \tag{9}$$

By expressing (9) as the following weighted integral of the MUO function:

$$\begin{aligned}
\int_0^1 \Big[& \frac{1}{\mathrm{P}(D=1)} \Big\{ \mathrm{P}(Z=0)\, p_C\, \omega(p, 0, p_c) + \mathrm{P}(Z=1)\, p_I\, \omega(p, 0, p_I) \Big\} \\
& - \frac{1}{\mathrm{P}(D=0)} \Big\{ \mathrm{P}(Z=0)\, (1-p_C)\, \omega(p, p_c, 1) + \mathrm{P}(Z=1)\, (1-p_I)\ \omega(p, p_I, 1) \Big\} \Big] \mathrm{MUO}(p)\, \mathrm{d}p.
\end{aligned}$$

where the weights are $\omega(p, p_L, p_H) = 1\{p_L \leq p < p_H\}/(p_H - p_L)$, I demonstrate that selection heterogeneity generalizes selection bias. The weighted integral also shows that selection bias is a function of the fraction of lottery winners, $\mathrm{P}(Z = 1)$, unlike selection heterogeneity. To the extent that selection bias is intended to capture a real-world phenomenon, it is undesirable for it to be an explicit function of the experimental design used to estimate it. Furthermore, selection bias is not identified without ancillary assumptions because the untreated outcome for always takers is not observed. However, I show that a different policy-relevant special case of selection heterogeneity is identified.

Turning to the next function, which I refer to as the "marginal treated outcome (MTO)" function, I emphasize that there is a meaningful distinction between the MTO function and the MUO function. The literature focuses on the MTO and MUO functions as intermediate inputs used to derive the MTE function. Mechanically, the MTE function is equal to the MTO function minus the MUO function. I emphasize that because the MTE function defines treatment effect heterogeneity and the MUO function defines selection heterogeneity, the MTO function defines the sum of selection heterogeneity plus treatment effect heterogeneity. It is tempting to assert that there should be no meaningful distinction between the MTO function and the MUO function because it would be possible to rename the treated as the untreated and vice versa. However, the treatment effect is defined relative to the untreated outcome, so changing the definition of the treatment would also change the definition of the treatment effect, preserving the asymmetry between the MTO and the MUO. The treatment effect has magnitude *and* direction: it is equal to $Y_T - Y_U$, not $|Y_T - Y_U|$, so the distinction between treated and untreated matters.

3 Findings

I have three main findings. First, I find selection and treatment effect heterogeneity within Oregon along the unobservable that separates always takers from compliers from never takers. Heterogeneity in the treatment effect is such that even though compliers increase their ER utilization upon gaining coverage, never takers would decrease their ER utilization upon gaining coverage. Second, I find that the heterogeneity within Oregon can reconcile the positive LATE in Oregon with the negative LATE in Massachusetts because the Massachusetts compliers are comparable to a subset of the Oregon never takers. Third, I find a nuanced role for observables in explaining the reconciliation. Self-reported health and previous ER-utilization can potentially explain the heterogeneity within Oregon and reconcile the Oregon and Massachusetts LATEs. However, those observables are not available in the Massachusetts data, so they are effectively part of the unobservable in the MTE function. Thus, LATE-reweighting that relies only on the common observables available in both contexts cannot reconcile the results, while MTE-reweighting can.

3.1 I Find Heterogeneity within Oregon

3.1.1 Selection Heterogeneity

Under the model that assumes no more than the LATE assumptions, I identify a special case of selection heterogeneity using a test that I refer to as the "untreated outcome test." The test statistic for this test is equal to the average untreated outcome of compliers minus the average untreated outcome of never takers. I derive both of these quantities in Appendix C. The untreated outcome test is similar or equivalent to tests proposed by Bertanha and Imbens (2014), Guo et al. (2014), and Black et al. (2017), which are generalized by Mogstad et al. (2018). Relative to the literature, my innovation with respect to the untreated outcome test is that I show that it identifies selection heterogeneity without any assumptions beyond the LATE assumptions.[5] This follows because having defined selection heterogeneity via the MUO function in an MTE model that assumes no more than the LATE assumptions, I express the untreated outcome test statistic as the following weighted integral of the MUO function:

$$\mathrm{E}[Y_U \mid p_C < U_D \leq p_I] - \mathrm{E}[Y_U \mid p_I < U_D \leq 1] = \int_0^1 (\omega(p, p_C, p_I) - \omega(p, p_I, 1))\,\mathrm{MUO}(p)\,\mathrm{d}p,$$

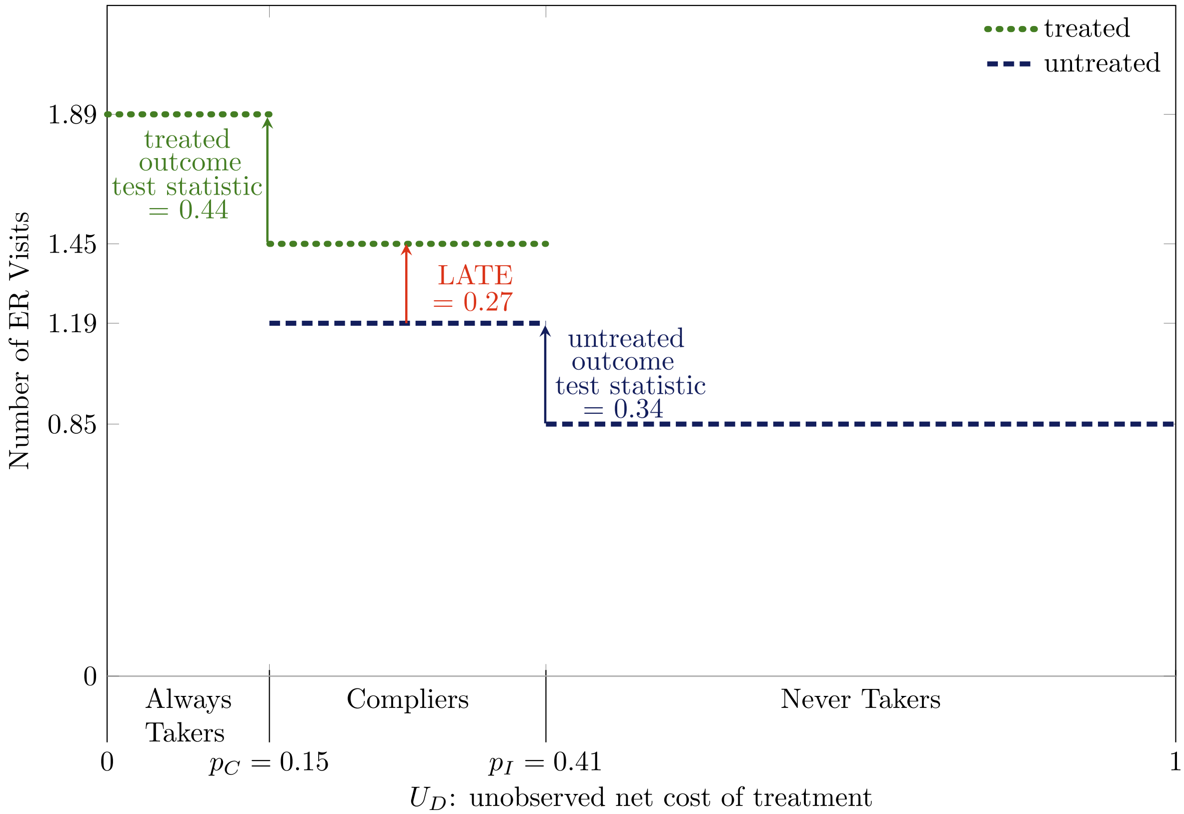
with weights $\omega(p, p_L, p_H) = 1\{p_L \leq p < p_H\}/(p_H - p_L)$, where the first term represents the average untreated outcome of compliers ($p_C < U_D \leq p_I$) and the second term represents the average untreated outcome of never takers ($p_I < U_D \leq 1$).

Applying the untreated outcome test to my analysis sample from the Oregon experiment, I reject the null hypothesis of selection homogeneity. As shown in Figure 3, when they are not enrolled in Medicaid, compliers visit the ER an average of 1.19 times, while never takers visit 0.85 times. The difference of 0.34 visits, reported as the untreated outcome test statistic in Table 1, is statistically different from zero. Under the model, compliers enroll in Medicaid before never takers, so the selection heterogeneity that I find indicates what the insurance literature refers to as "adverse selection" from compliers to never takers.

Without further assumptions, the untreated outcome test is not informative about selection heterogeneity from always takers to compliers because untreated outcomes are not observed for always takers. However, treated outcomes are observed for always takers. A test that I refer to as the "treated outcome test" has a test statistic that is equal to the average treated outcome of always takers minus the average treated outcome of never takers. I derive both of these quantities in Appendix C. The econometric literature that proposes tests related to the untreated outcome test also proposes tests related to the treated out-

[5] I refer to the Bertanha and Imbens (2014) test as "similar" to the untreated outcome test because the authors develop it for a regression discontinuity context, but it is effectively equivalent. However, the authors do not interpret it as a test of selection heterogeneity; instead, they interpret it as one component of a test for external validity. Guo et al. (2014) propose a test that is equivalent to the untreated outcome test as one component of a test for unmeasured confounding, but they also do not discuss it as a test for selection heterogeneity. Black et al. (2017) propose a test that is equivalent to the untreated outcome test as a test for selection bias on the untreated outcome, which they define with the outcome test statistic. They do not discuss how their definition of selection bias relates to the MUO function or to the definition of selection bias from the literature.

Figure 3: Number of ER Visits for Always Takers, Compliers, and Never Takers in the Oregon Experiment



Note. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. Treatment represents enrollment in Medicaid. p_C is the probability of treatment in the control group, and p_I is the probability of treatment in the intervention group. Some differences between statistics might not appear internally consistent because of rounding.

come test (Bertanha and Imbens, 2014; Guo et al., 2014; Black et al., 2017). Relative to the literature, I emphasize that a rejection of the treated outcome test identifies selection heterogeneity, treatment effect heterogeneity, or a combination of selection and treatment effect heterogeneity. Recall that MTO function is the sum of the MUO function, which defines selection heterogeneity, and the MTO function, which defines treatment effect heterogeneity. Therefore, I show that the treated outcome test identifies the sum of selection heterogeneity plus treatment effect heterogeneity by expressing the treated outcome test statistic as the following weighted integral of the MTO function:

$$\mathrm{E}[Y_T \mid 0 \leq U_D \leq p_C] - \mathrm{E}[Y_T \mid p_C < U_D \leq p_I] = \int_0^1 (\omega(p, 0, p_C) - \omega(p, p_C, p_I)) \,\mathrm{MTO}(p)\,\mathrm{d}p,$$

with weights $\omega(p, p_L, p_H) = 1\{p_L \leq p < p_H\}/(p_H - p_L)$, where the first term represents the average treated outcome of always takers ($0 \leq U_D \leq p_C$) and the second term represents the

Table 1: Number of ER Visits for Always Takers, Compliers, and Never Takers

	Mean				
	(1) Always Takers	(2) Compliers	(3) Never Takers	Untreated Outcome Test (2) - (3)	Treated Outcome Test (1) - (2)
Number of ER Visits					
Treated	1.89	1.45	0.55		0.44
	(0.08)	(0.11)	(0.45)		(0.17)
Untreated	1.35	1.19	0.85	0.34	
	(0.17)	(0.11)	(0.03)	(0.13)	
Treatment Effect	0.54	0.27	-0.29		
(Treated - Untreated)	(0.19)	(0.15)	(0.45)		

Note. Bootstrapped standard errors are in parentheses. The shaded cells report extrapolated values from MTE-reweighting via (10)–(12) for treated individuals (N=4,725) and untreated individuals (N=14,897). The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. Treatment represents enrollment in Medicaid. Some differences between statistics might not appear internally consistent because of rounding.

average untreated outcome of compliers ($p_C < U_D \leq p_I$).

Applying the treated outcome test to my analysis sample from the Oregon experiment, I reject the null hypothesis that treatment effect heterogeneity and selection heterogeneity sum to zero. As shown in Figure 3, always takers visit the ER an average of 1.89 times when enrolled in Medicaid, while compliers visit an average of 1.45 times. The average difference of 0.44 visits, reported as the treated outcome test statistic in Table 1, is statistically different from zero. Under the model, always takers enroll in Medicaid before compliers, so their greater visits with Medicaid must reflect either adverse selection, or a decrease in the treatment effect from always takers to compliers, or both. In pursuit of reconciling the Oregon LATE with the Massachusetts LATE, I am particularly interested in whether there is treatment effect heterogeneity within the Oregon experiment. Although the treated outcome test indicates that there could be treatment effect heterogeneity, I cannot separate it from selection heterogeneity without an ancillary assumption.

3.1.2 Treatment Effect Heterogeneity

To identify treatment effect heterogeneity, I make a transparent ancillary assumption beyond the model that assumes no more than the LATE assumptions:

AA.1. (Linear Selection Heterogeneity and Linear Treatment Effect Heterogeneity) In (7) and (8), for $k \in \{T, U\}$, specify $g_k(X, U_D, \gamma_k) = \alpha_k + \beta_k U_D + \gamma_k$, where $\mathrm{E}[\gamma_k \mid U_D = p] = 0$. Therefore,

$$\mathrm{MTO}(p) = \mathrm{E}[Y_T \mid U_D = p] = \alpha_T + \beta_T p$$

$$\begin{aligned}\mathrm{MUO}(p) &= \mathrm{E}\left[Y_U \mid U_D = p\right] = \alpha_U + \beta_U p\\ \mathrm{MTE}(p) &= \mathrm{E}\left[Y_T - Y_U \mid U_D = p\right] = (\alpha_T - \alpha_U) + (\beta_T - \beta_U)\, p.\end{aligned}$$

This assumption requires that any selection heterogeneity is linear in U_D and that any treatment effect heterogeneity is linear in U_D, but it allows for the possibility that there is no selection or treatment effect heterogeneity. In that case, the MUO slope coefficient β_U and the MTE slope coefficient $(\beta_T - \beta_U)$ will both be equal to zero. Brinch et al. (2017) impose the same assumption to examine the impact of family size on child outcomes; Olsen (1980) imposes linearity of the MTO function to examine the impact of family size on maternal outcomes; and several other papers impose linearity of the MTE function in other applications (see Moffitt, 2008; French and Song, 2014). Applied work that extrapolates to other policies using the LATE also makes a stronger, implicit assumption that the MTE function is linear and has zero slope.

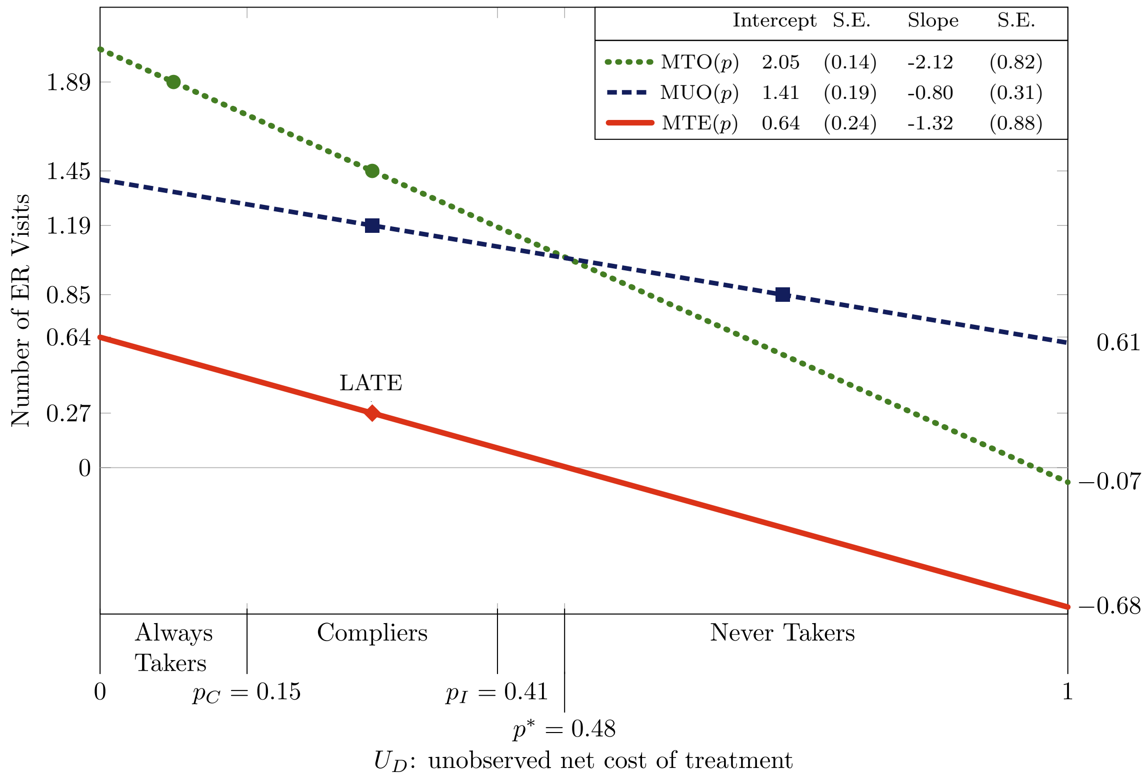
Figure 4 depicts the MTO, MUO, and MTE functions within the Oregon experiment under AA.1. On the vertical axis, the two points labeled with circular markers indicate the average outcomes of always takers and treated compliers, which fall at the median of the support for each group on the horizontal axis. These two points identify the intercept and slope of the MTO function, depicted with a dotted line. The two points labeled with square markers identify the intercept and slope of the MUO function, depicted with a dashed line. I depict the MTE function, the vertical difference between the MTO and MUO functions, with a solid line. As shown, MTE function is positive for low levels of enrollment and negative for high levels of enrollment, even though the LATE is positive. The downward slope of the MTE function indicates that the treatment effect of insurance on ER utilization decreases as enrollment increases.

3.2 Oregon Heterogeneity Can Reconcile Oregon and Massachusetts LATEs

3.2.1 Massachusetts MTE(p) Also Slopes Downward

My goal is to reconcile the Oregon and Massachusetts LATEs using the Oregon MTE function, since estimates from the Oregon experiment are considered the "gold standard." Before assuming that the Oregon MTE function is the same as the Massachusetts MTE function, however, I assess whether such an assumption is plausible. I acknowledge that many factors differed between the Massachusetts and Oregon contexts. For example, treatment in the Oregon context only captures enrollment in Medicaid, while treatment in Massachusetts also captures enrollment in other types of health insurance coverage. However, given my interest in reconciling the LATEs across contexts, I am ultimately only interested in factors that lead to empirical differences in treatment effects across contexts, and such differences should be captured in differences the MTE functions across contexts.

Figure 4: MTO(p), MUO(p), and MTE(p)



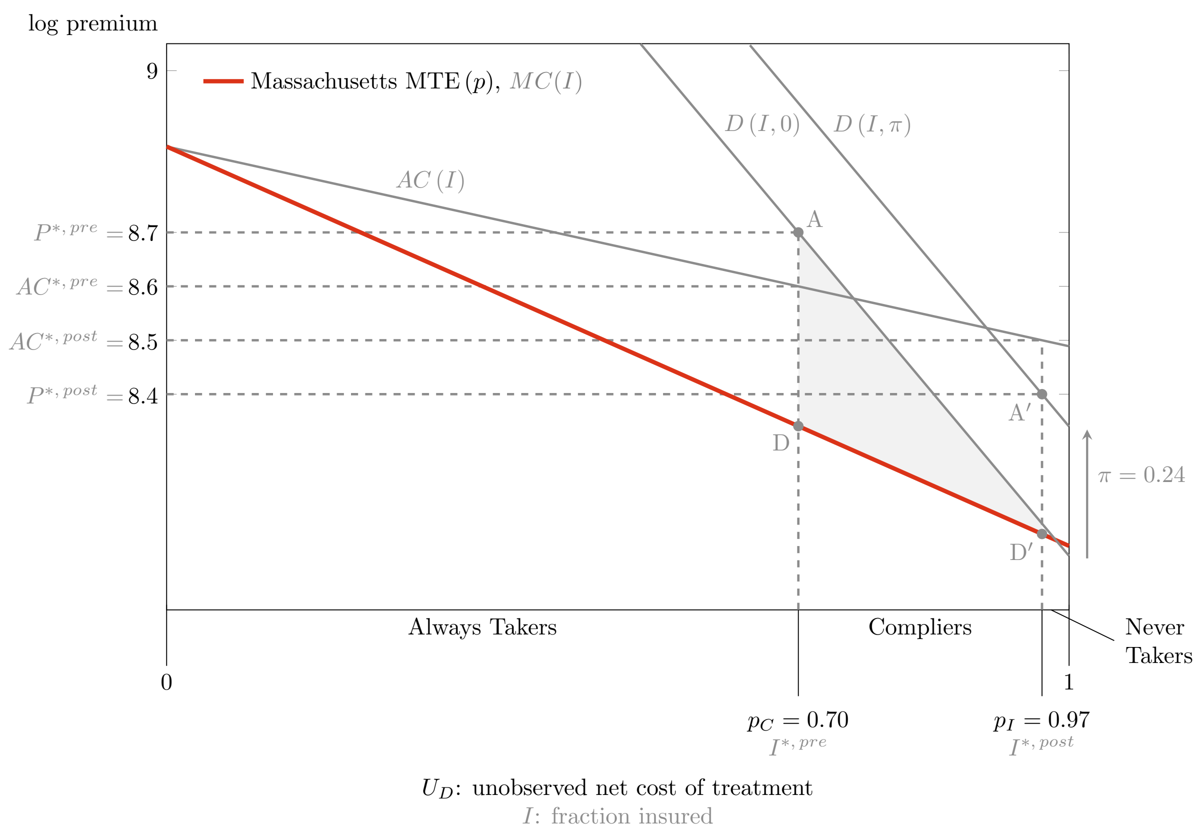
Note. Bootstrapped standard errors are in parentheses. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. Treatment represents enrollment in Medicaid. p_C is the probability of treatment in the control group, and p_I is the probability of treatment in the intervention group. Some differences between statistics might not appear internally consistent because of rounding.

It would be straightforward to estimate a Massachusetts MTE function using individual-level data on insurance coverage and ER utilization from a representative sample of individuals in Massachusetts before and after the reform. With such data, I could define the instrument as an indicator for after the reform, the treatment as an indicator for insurance coverage, and the outcome as the number of visits to the ER. However, none of the studies that examine the impact of the Massachusetts reform on ER utilization use such data.[6] Therefore, I need to be more creative to estimate a Massachusetts MTE function.

[6] Chen et al. (2011) use data on ER visits from the Massachusetts Division of Health Care Finance and Policy aggregated to the quarter level, but they do not use data on insurance. Miller (2012) uses the same data aggregated to the county-quarter level, matched to county-level data on insurance before the reform, but the individual-level data on ER utilization and insurance coverage before and after the reform are not available. In Kolstad and Kowalski (2012), I use data from the Behavioral Risk Factor Surveillance System (BRFSS), which contains all of the necessary elements except ER utilization. I also use the Healthcare Cost and Utilization Project (HCUP) National Inpatient Sample (NIS), which contains the necessary elements on the individual level, but it is restricted to individuals who were admitted to the hospital. The data from Smulowitz et al. (2011) are even more restricted because they only include individuals who visited the ER at a convenience sample of 11 Massachusetts hospitals.

To estimate a Massachusetts MTE function, I recast results from my previous work on the Massachusetts reform from Hackmann et al. (2015) in terms of the MTE model with ancillary assumption AA.1. Although I do not observe ER spending or visits in the Hackmann et al. (2015) data, I do observe total health care spending, and evidence from the Oregon experiment shows that ER spending and total health care spending are complements (Taubman et al., 2014). To show the Hackmann et al. (2015) estimates and MTE function, I reproduce Figure 8 from Hackmann et al. (2015) using notation consistent with the MTE model while preserving notation from the original figure in lighter typeface in Figure 5. The marginal cost function estimated in Hackmann et al. (2015) represents a marginal treatment effect function because it represents the difference between marginal costs to insurers on behalf of insured individuals and uninsured individuals. This Massachusetts MTE function, like the Oregon MTE function, is downward sloping, indicating that in both contexts, the treatment effect of insurance on utilization decreases as insurance enrollment increases.

Figure 5: Figure 8 from Hackmann et al. (2015) Recast as Massachusetts MTE(p)



3.2.2 MTE-Reweighting from Oregon to Massachusetts Can Reconcile LATEs

Given that I find a downward-sloping MTE function for total health care spending in Massachusetts, I am more confident in assuming that the MTE function for ER utilization in Massachusetts is the same as the MTE function for ER utilization in Oregon. Under this assumption, I can re-weight the Oregon MTE function to obtain a LATE for the Massachusetts reform.

I re-weight the Oregon MTE function and its component MTO and MUO functions over a general range of the enrollment margin $p_L < U_D \leq p_H$ as follows:

$$\mathrm{E}\left[Y_T \mid p_L < U_D \leq p_H\right] = \int_0^1 \omega(p, p_L, p_H)\mathrm{MTO}(p)\,\mathrm{d}p \tag{10}$$

$$\mathrm{E}\left[Y_U \mid p_L < U_D \leq p_H\right] = \int_0^1 \omega(p, p_L, p_H)\mathrm{MUO}(p)\,\mathrm{d}p \tag{11}$$

$$\mathrm{E}\left[Y_T - Y_U \mid p_L < U_D \leq p_H\right] = \int_0^1 \omega(p, p_L, p_H)\mathrm{MTE}(p)\,\mathrm{d}p, \tag{12}$$

using weights $\omega(p, p_L, p_H) = 1\{p_L < p \leq p_H\}/(p_H - p_L)$. These weights are special cases of general weights for MTE-reweighting given by Heckman and Vytlacil (2007). Unlike the weights used by Brinch et al. (2017), these weights allow me to recover exact values of observed average outcomes for always takers ($0 \leq U_D \leq p_C$), compliers ($p_C < U_D \leq p_I$), and never takers ($p_I < U_D \leq 1$).

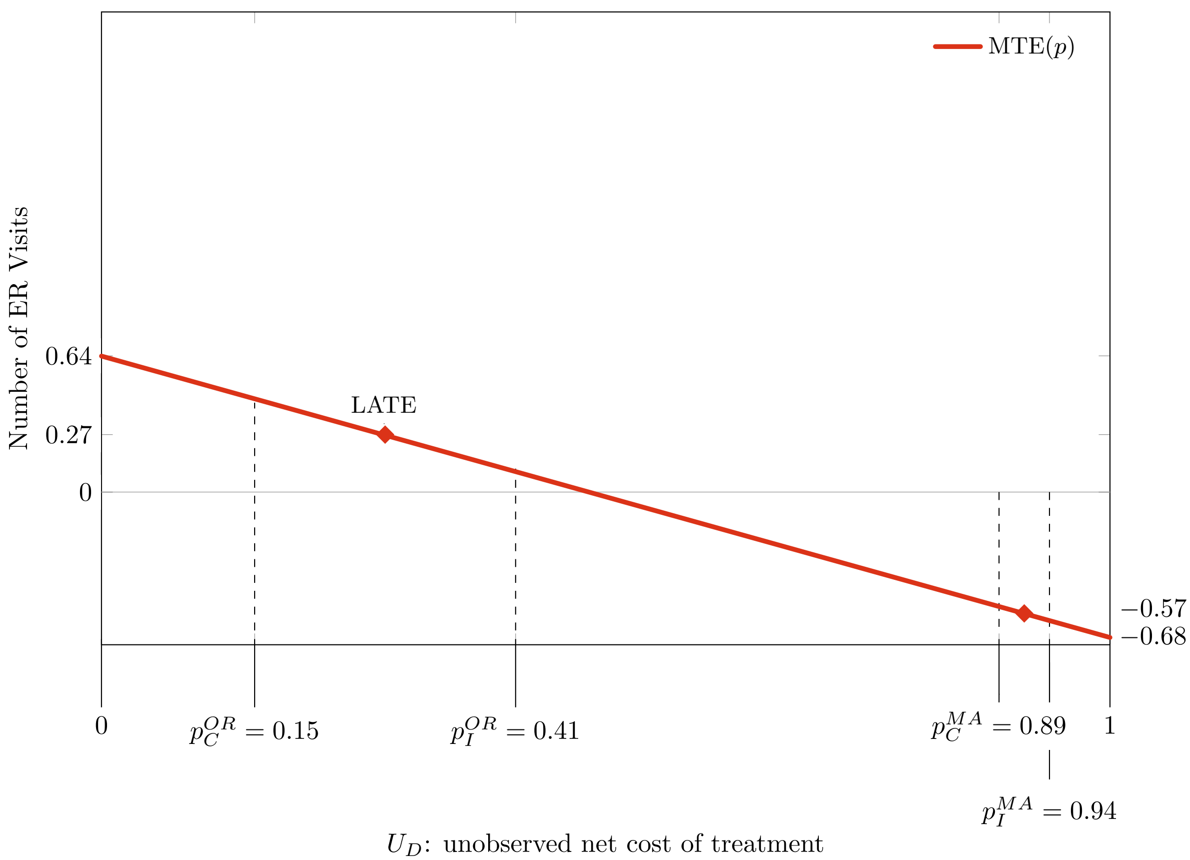
I demonstrate the results of reweighting to obtain estimates for Oregon always and never takers in the shaded cells of Table 1. I only observe always takers when enrolled in Medicaid, and they visit the ER 1.89 times. Reweighting indicates that if always takers were not enrolled in Medicaid, they would visit the ER 1.35 times, such that the average treatment effect for always takers is an increase of 0.54 visits. In contrast, reweighting indicates that the average treatment effect for never takers is a *decrease* of 0.29 visits.

I reweight the Oregon MTE to obtain estimates for Massachusetts compliers using the same approach. I demonstrate the approach graphically in Figure 6, in which I reproduce the Oregon MTE. I label the probability of health insurance coverage in Massachusetts before the reform as $p_C^{MA} = 0.89$ and after the reform as $p_I^{MA} = 0.94$. I obtain these values from the Behavioral Risk Factor Surveillance System (BRFSS) data that I used to study the Massachusetts reform in Kolstad and Kowalski (2012). Unlike the Hackmann et al. (2015) data, which only capture enrollment in the individual health insurance market, the BRFSS data capture enrollment in the entire state. It is important to capture enrollment in the entire state for comparison to the literature on the impact of the Massachusetts reform on ER utilization (Chen et al., 2011; Kolstad and Kowalski, 2012; Miller, 2012; Smulowitz et al., 2011).

As shown, enrollment levels before and after the Massachusetts reform would entail enrollment of a subset of never takers in Oregon. Therefore, application of the Oregon MTE to Massachusetts implies that Massachusetts compliers are comparable to a subset of Oregon never takers in terms of their unobserved net cost of treatment U_D. There is a case to be made that the Oregon sample is actually a subset of the Massachusetts sample along the lower range of U_D because all individuals in the Oregon sample entered a lottery for Medicaid, so they should all have low unobserved net costs of Medicaid relative to individuals in Massachusetts. Therefore, it is likely conservative to compare Massachusetts compliers to this particular subset of Oregon never takers.

MTE-reweighting the Oregon MTE via (12) over the range from $p_C^{MA} = 0.89$ to $p_I^{MA} = 0.94$, I predict that the Massachusetts reform should have decreased ER visits by an average

Figure 6: Extrapolation of MTE(p) to Massachusetts



Note. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. Treatment represents enrollment in Medicaid. p_C^{OR} is the probability of treatment in the control group in Oregon, p_I^{OR} the probability of treatment in the intervention group in Oregon, p_C^{MA} the probability of treatment in the control group in the Massachusetts reform, and p_I^{MA} the probability of treatment in the intervention group in the Massachusetts reform.

of 0.57 visits among Massachusetts compliers. Miller (2012) finds that insurance enrollment induced by the Massachusetts reform decreased ER visits by 0.67 to 1.28 visits per person per year, depending on the empirical strategy.[7] The decrease that I find over the 19 months from March 10, 2008 to September 30, 2009 translates into a decrease of 0.36 visits per person per year (=(0.57/19)*12), which is smaller than her estimates but still comparable. Therefore, my extrapolations can reconcile the increase in ER utilization in Oregon with the decrease in ER utilization in Massachusetts using only variation in the unobserved net cost of treatment U_D.

3.3 Self-Reported Health and Previous ER Utilization Explain Heterogeneity, but Common Observables Do Not

3.3.1 Self-Reported Health Could Reconcile LATEs

To explore mechanisms for why the impact of coverage on ER utilization is positive for some groups but negative for others, I examine observables. I begin by examining self-reported health. I observe self-reported health for almost all individuals in the Massachusetts BRFSS data from Kolstad and Kowalski (2012), and I observe self-reported health for a subset of individuals in the Oregon administrative data who were surveyed. Using the Oregon data, Finkelstein et al. (2012) shows that Medicaid improved self-reported health, so I only compare the self-reported health of groups without Medicaid: compliers who lost the lottery and never takers. I obtain the average probability that individuals in these groups reported fair or poor health as I describe in Appendix C.

As shown in Table 2, within Oregon and Massachusetts, I find that never takers are less likely to be in fair or poor health than compliers who are not enrolled in Medicaid, consistent with adverse selection via the untreated outcome test. However, differences in self-reported health are more striking across both contexts than they are within each context. As I show in Table 2, 55% of Oregon compliers report fair or poor health, while only 34% of Oregon never takers report fair or poor health. In stark contrast, only 21% of Massachusetts compliers report fair or poor health. These comparisons suggest an important mechanism for heterogeneity in the treatment effect. Upon gaining coverage, individuals in worse health (Oregon compliers) increase their ER utilization, while individuals in better health (Oregon never takers and Massachusetts compliers) decrease their ER utilization.

[7] Other estimates from the literature are not directly comparable. Chen et al. (2011) does not provide an estimate but reports no change in ER utilization based on figures that compare ER utilization in Massachusetts, New Hampshire, and Vermont over time. The Kolstad and Kowalski (2012) estimate shows that hospital admissions from the ER decreased by 2.02 percentage points on a base of 38.7% after the reform relative to before the reform in Massachusetts relative to other states. The Smulowitz et al. (2011) estimate shows that low-severity visits to the ER decreased by 1.8% after the reform relative to before the reform for publicly-subsidized and uninsured patients relative to insured and Medicare patients.

Table 2: Always Takers, Compliers, and Never Takers: Oregon vs. Massachusetts

	Means				Difference in Means	
		(1)	(2)	(3)		
	All	Always Takers	Compliers	Never Takers	(1) - (2)	(2) - (3)
Oregon Health Insurance Experiment of 2008						
Fair or Poor Health, Untreated[a]	0.42	-	0.55	0.34	-	0.20
	(0.01)		(0.03)	(0.01)		(0.04)
Number of Pre-period ER Visits	0.87	1.36	0.88	0.73	0.48	0.15
	(0.01)	(0.05)	(0.07)	(0.03)	(0.09)	(0.09)
Common Observables						
Age	40.69	39.45	42.41	40.25	-2.96	2.16
	(0.09)	(0.29)	(0.41)	(0.19)	(0.53)	(0.57)
Female	0.56	0.72	0.53	0.53	0.19	0.003
	(0.003)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)
English	0.91	0.90	0.92	0.91	-0.02	0.01
	(0.002)	(0.01)	(0.01)	(0.004)	(0.01)	(0.01)
N	19,643	2,986	5,092	11,565		
Massachusetts Health Reform of 2006						
Fair or Poor Health, Untreated[a]	0.19	-	0.21	0.18	-	0.03
	(0.02)		(0.03)	(0.01)		(0.04)
Common Observables						
Age	42.00	42.15	42.42	38.98	-0.26	3.43
	(0.086)	(0.12)	(1.41)	(0.49)	(1.49)	(1.57)
Female	0.51	0.52	0.43	0.38	0.10	0.04
	(0.003)	(0.004)	(0.05)	(0.02)	(0.05)	(0.06)
English	0.96	0.98	0.86	0.81	0.12	0.05
	(0.001)	(0.001)	(0.02)	(0.02)	(0.02)	(0.04)
N	62,456	55,966	3,175	3,314		

Note. Bootstrapped standard errors are in parentheses. Data for the Massachusetts health reform are taken from pooled annual samples of the Behavioral Risk Factor Surveillance System (BRFSS) from years 2004–2009 and restricted to ages 21–64 (the age range of the Oregon sample). For the Massachusetts health reform, treatment is an indicator that equals one for individuals with any form of health insurance ("Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare?"). The instrument is an indicator that equals one in the post-period of the expansion on and after July 2007. "Age" is measured in year 2008 for the Oregon Health Insurance Experiment and in year 2006 for the Massachusetts health reform. "Female" is a binary indicator for the gender of the respondent. "English" is a binary indicator that equals one for individuals in the Oregon Health Insurance Experiment who requested materials in English and that equals one for individuals in the BRFSS who completed the interview in English. The number of pre-period visits is measured before the study period from January 1, 2007 to March 9, 2008. "Fair or Poor Health" equals one when individuals self-report having fair or poor health on a 5-point scale. [a]Number of observations in the Oregon Health Insurance Experiment with nonmissing self-reported health: 5,833. Number of observations in the BRFSS with nonmissing self-reported health: 62,161. Some differences between statistics might not appear internally consistent because of rounding.

3.3.2 Previous ER Utilization Explains Heterogeneity within Oregon

To quantify how much heterogeneity in the treatment effect observables can explain, I incorporate observables into the MTE function. To do so, I use a shape restriction commonly used in the MTE literature (see Brinch et al., 2017; Carneiro and Lee, 2009; Carneiro et al., 2011; Maestas et al., 2013). In my context, the shape restriction requires that included observables X and the remaining unobserved net cost of treatment U_D have additively-separable impacts on ER utilization with and without Medicaid. I incorporate the shape restriction into AA.1 to obtain the following alternative ancillary assumption:

AA.2. (Linear Selection Heterogeneity and Linear Treatment Effect Heterogeneity with Covariate Shape Restriction) In (7) and (8), for $k \in \{T, U\}$, specify $g_k(X, U_D, \gamma_k) = \delta_k' X + \lambda_k U_D + \xi_k$, where $\mathrm{E}[\gamma_k \mid X = x, U_D = p] = 0$. Therefore,

$$\begin{aligned}
\mathrm{MTO}(p) &= \mathrm{E}[Y_T \mid X = x, U_D = p] = \delta_T' x + \lambda_T p \\
\mathrm{MUO}(p) &= \mathrm{E}[Y_U \mid X = x, U_D = p] = \delta_U' x + \lambda_U p \\
\mathrm{MTE}(p) &= \mathrm{E}[Y_T - Y_U \mid X = x, U_D = p] = (\delta_T - \delta_U)' x + (\lambda_T - \lambda_U) p.
\end{aligned}$$

I present an algorithm for estimation of these functions that simplifies the Heckman et al. (2006) algorithm in Appendix D.[8] I reweight these functions using the same approach that I use in (10)–(12).

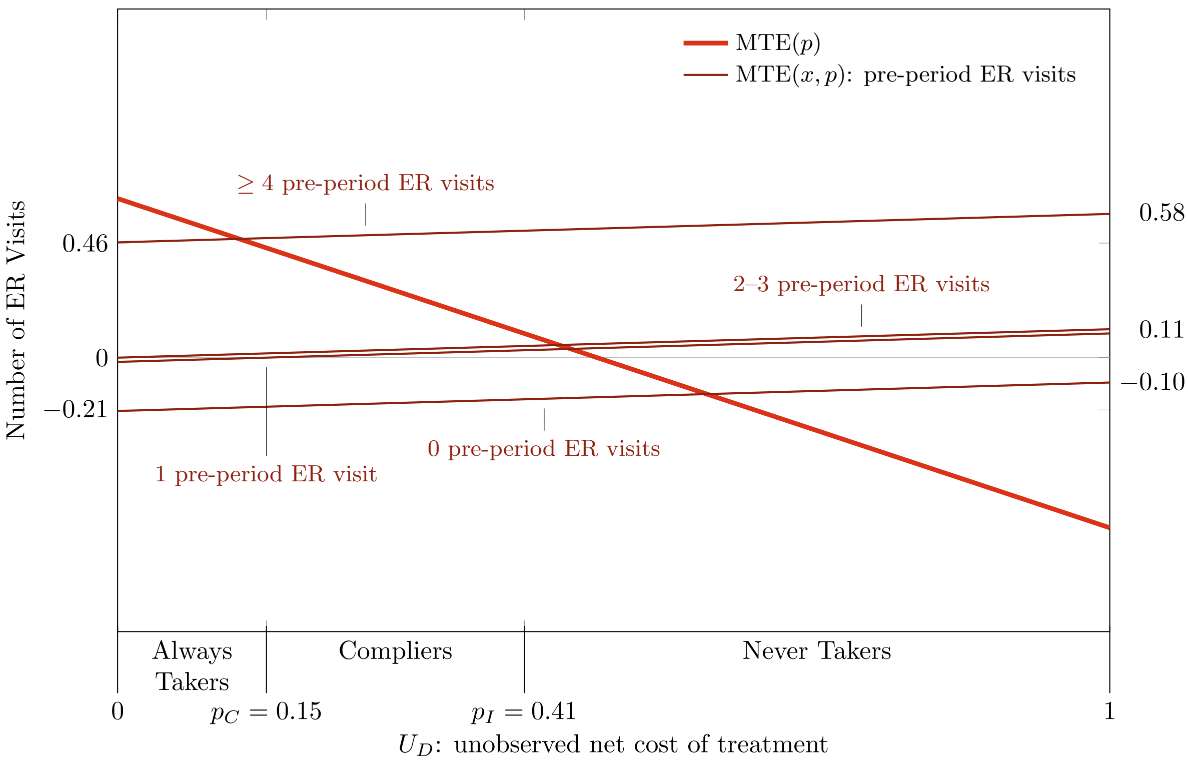
I do not incorporate self-reported health into the MTE function because evidence from Oregon shows that recorded self-reported health is an outcome and not merely a covariate for treated individuals (Finkelstein et al., 2012). However, I do observe previous ER-utilization from before the lottery took place for all individuals, and ER utilization from before the lottery took place is correlated with self-reported health for untreated individuals. Specifically, for each individual in the Oregon administrative data, I observe the total number of pre-period ER visits from January 1, 2007, to March 9, 2008. I report the average number of per-period ER visits for always takers, compliers, and never takers in Table 2, calculated as described in Appendix C. Always takers visited the ER an average of 1.36 times, while compliers visited an average of 0.88 times, and never takers visited an average of 0.73 times. The monotonic relationship in previous ER utilization across these groups indicates adverse selection on previous ER utilization: individuals with larger previous ER utilization are more likely to enroll in Medicaid.

Incorporating previous ER utilization into the MTE via AA.2, I find that previous ER utilization can explain the entire decrease in treatment effect from always takers to compliers

[8] For inference, I bootstrap using 200 replications, and I report the standard deviation as the standard error or the 2.5 and 97.5 percentiles as the 95% confidence interval.

to never takers. There is substantial variation in pre-period ER utilization: 66% of individuals have zero visits, 17% have one visit, 11% have 2 to 3 visits, and 6% have 4 or more visits in the pre-period. By incorporating observables for each of these visit ranges into the MTE, I obtain a separate MTE(x, p) for each range. As depicted in Figure 7, the MTE(p) function, which does not incorporate observables, has a pronounced downward slope, indicating substantial unexplained heterogeneity in treatment effect. However, when I incorporate controls for previous ER utilization into the MTE(x, p) function, the negative slope disappears, and the slope becomes negligible and slightly positive.

Figure 7: MTE(x, p) with Previous ER Utilization



Note. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. Pre-period ER visits refers to a group of indicators for visiting the ER 0 times, 1 time, 2–3 times, and 4 or more times during the pre-period from January 1, 2007 to March 9, 2008. Treatment represents enrollment in Medicaid. p_C is the probability of treatment in the control group, and p_I is the probability of treatment in the intervention group. In this figure, the function for 1 pre-period ER visits has been shifted downward slightly to make it easier to discern from the function for 2–3 pre-period ER visits.

The remaining slope in the MTE with observables is not meaningful. Looking beyond the slope of the MTE(x, p) function to its level at various values of pre-period ER visits reveals a clear monotonic relationship between pre-period ER visits and the treatment effect of Medicaid enrollment on subsequent ER visits. As depicted in Figure 7, the MTE(x, p) for individuals with 4 or more pre-period visits is always positive, and the MTE(x, p) for individuals with zero pre-period visits is always negative. This figure demonstrates that individuals with high numbers of ER visits in the pre-period increase their ER utilization upon gaining coverage, while individuals with zero ER visits in the pre-period decrease their ER utilization upon gaining coverage.

The finding that previous ER utilization can explain all of the treatment effect heterogeneity captured by the Oregon MTE suggests that when no observables are included in the Oregon MTE, the unobservable U_D captures previous ER utilization. It is plausible that U_D captures previous ER utilization or even ER utilization after the lottery took place because Medicaid allows hospitals to facilitate enrollment of eligibles. Because hospitals can facilitate enrollment, it is possible that some individuals became always takers precisely because they showed up at the ER to receive care, and the ER facilitated their enrollment. This mechanism could explain why always takers signed up for a lottery for Medicaid even though they were already eligible – they did not know that they were eligible until they showed up at the ER.

3.3.3 LATE-Reweighting with Common Observables Cannot Reconcile LATEs

Although self-reported health and previous ER utilization provide promising mechanisms to reconcile the Oregon and Massachusetts LATEs, neither are available for all individuals in the Oregon and Massachusetts data. Therefore, I consider whether it would be possible to reconcile the Oregon and Massachusetts LATEs using LATE-reweighting and the three common observables for all individuals in the Massachusetts BRFSS data and the Oregon administrative data: age, gender, and an indicator for communications in English. In Table 2, I present summary statistics on the common observables in both samples.

To examine variation in the common observables available for LATE-reweighting, I use each common observable to divide the sample into two subgroups, and I report LATEs within each subgroup in Table 3. As shown, the LATEs within each subgroup are all positive, with the exception of the LATE within the group that requested communication in a language other than English. Taubman et al. (2014) report LATEs within a wide variety of observable subgroups and also find that almost all are positive. Because LATEs within each subgroup are almost all positive, LATE-reweighting based on any of the common observables yields a positive treatment effect for almost any weights. Therefore, LATE-reweighting using only the common observables cannot reconcile the positive treatment effect in Oregon with the negative treatment effect in Massachusetts.

It is not surprising that LATE-reweighting with common observables cannot explain treatment effect heterogeneity across Oregon and Massachusetts because the common observables cannot explain treatment effect heterogeneity within Oregon. To demonstrate, I estimate an MTE within each subgroup, and I report the slope and intercept in Table 3.

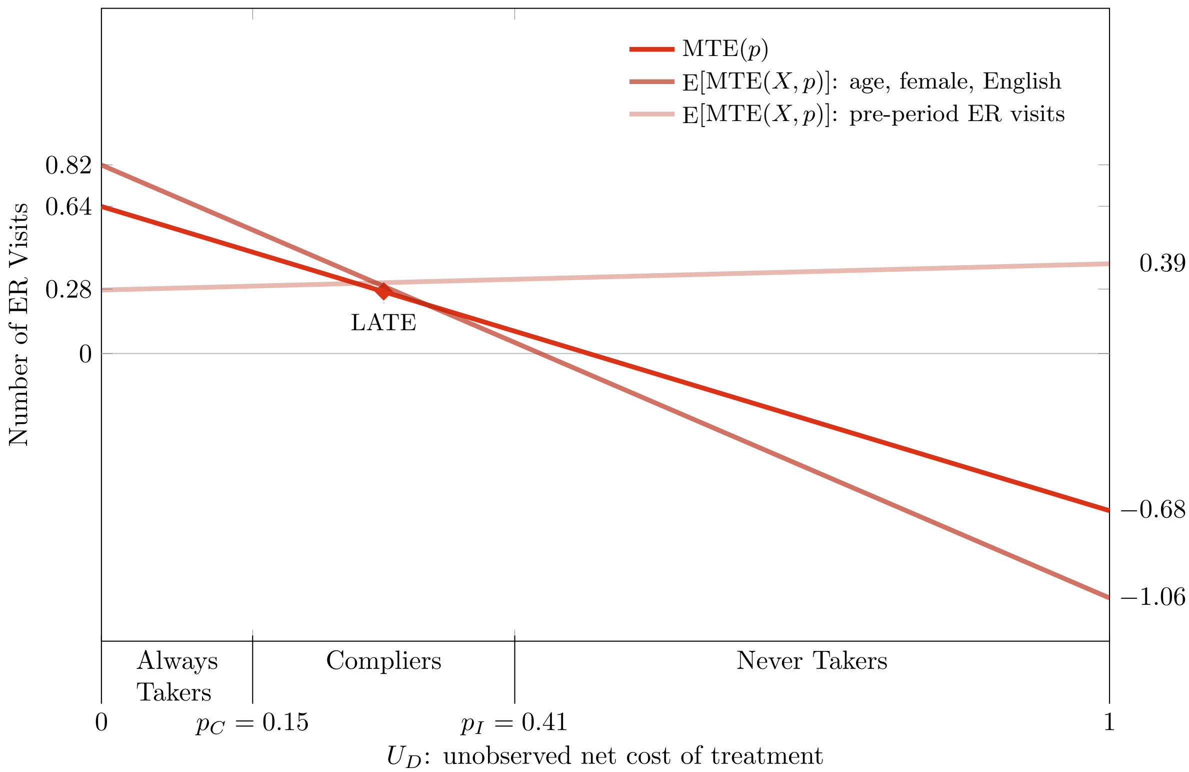
Table 3: Subgroup Analysis of Common Observables with LATE and MTE(p)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	All	Age ≥ median[a]	Age < median[a]	Female	Male	English	Non-English
Oregon Health Insurance Experiment of 2008							
LATE	0.27	0.14	0.44	0.14	0.39	0.30	-0.15
	(0.15)	(0.18)	(0.25)	(0.21)	(0.21)	(0.16)	(0.34)
p_C	0.15	0.13	0.17	0.20	0.10	0.15	0.16
	(0.003)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.01)
p_I	0.41	0.43	0.39	0.43	0.38	0.41	0.38
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
MTE intercept	0.64	0.98	0.31	0.48	0.92	0.72	0.14
	(0.24)	(0.28)	(0.39)	(0.32)	(0.33)	(0.25)	(0.47)
MTE slope	-1.32	-3.01	0.48	-1.06	-2.20	-1.51	-1.07
	(0.88)	(1.04)	(1.49)	(1.08)	(1.40)	(0.92)	(2.07)
p*	0.48	0.33	-0.63	0.45	0.42	0.48	0.13
	(2.84)	(0.85)	(10.37)	(1.49)	(3.47)	(4.53)	(11.99)
N	19,622	9,816	9,806	10,932	8,690	17,871	1,751
Massachusetts Health Reform of 2006							
p_C	0.90	0.93	0.87	0.92	0.87	0.91	0.55
	(0.003)	(0.003)	(0.005)	(0.003)	(0.005)	(0.003)	(0.02)
p_I	0.95	0.96	0.93	0.96	0.93	0.96	0.74
	(0.002)	(0.002)	(0.004)	(0.002)	(0.004)	(0.002)	(0.02)
N	62,456	40,492	21,964	38,808	23,648	59,233	3,223

Note. Bootstrapped standard errors are in parentheses. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. Treatment represents enrollment in Medicaid. The value p^* indicates the share of the sample with positive treatment effects when the MTE(p) curve slopes downward and the share of the sample with negative treatment effects when the MTE(p) curve slopes upward. When $p^* \geq 1$, this share is 100% of the sample, and when $p^* \leq 0$, this share is 0% of the sample. "Age" is measured in year 2008 for the Oregon Health Insurance Experiment and in year 2006 for the Massachusetts health reform. "English" is an indicator variable for individuals in the Oregon Health Insurance Experiment who requested materials in English and that equals one for individuals in the BRFSS who completed the interview in English. "Non-English" is the complement of "English." [a]The median age in the Oregon Health Insurance Experiment is 41. I use the same age to construct the Massachusetts subgroups.

In almost all subgroups, the MTE slopes downward. When the MTE slopes downward, the horizontal intercept p^* gives the fraction of individuals predicted to have positive treatment effects. In all but one subgroup, even though the LATEs are positive, the MTEs predict that the majority of individuals have *negative* treatment effects, indicating that the common observables leave substantial heterogeneity unexplained.

Figure 8: MTE(x, p) with Previous ER Utilization vs. MTE(x, p) with Common Observables



Note. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. Pre-period ER visits refers to a group of indicators for visiting the ER 0 times, 1 time, 2–3 times, and 4 or more times during the pre-period from January 1, 2007 to March 9, 2008. Treatment represents enrollment in Medicaid. "Age" is measured in year 2008. "Female" is a binary indicator for the gender of the respondent. "English" is a binary indicator that equals one for individuals who requested materials in English. The specification with common covariates (age, female, English) includes all two-way interactions. p_C is the probability of treatment in the control group, and p_I is the probability of treatment in the intervention group.

Furthermore, when I include all of the common observables as well as their two-way interactions in the MTE, substantial heterogeneity remains unexplained. I emphasize the comparison of unexplained heterogeneity across various MTE functions in Figure 8. To do so, I present E[MTE(x, p)] functions, which average included observed heterogeneity across all individuals. Consistent with the depiction in Figure 7, the inclusion of pre-period ER visits in MTE(x, p) results in a function that is flatter than MTE(p). Therefore, the inclusion of pre-period ER visits decreases unexplained heterogeneity in the treatment effect. In contrast, the inclusion of the common observables in $\text{MTE}(x, p)$ results in a function that is steeper than $\text{MTE}(p)$. Therefore, the inclusion of common observables *increases* unexplained heterogeneity in the treatment effect.

3.3.4 MTE-Reweighting with Common Observables Can Reconcile LATEs

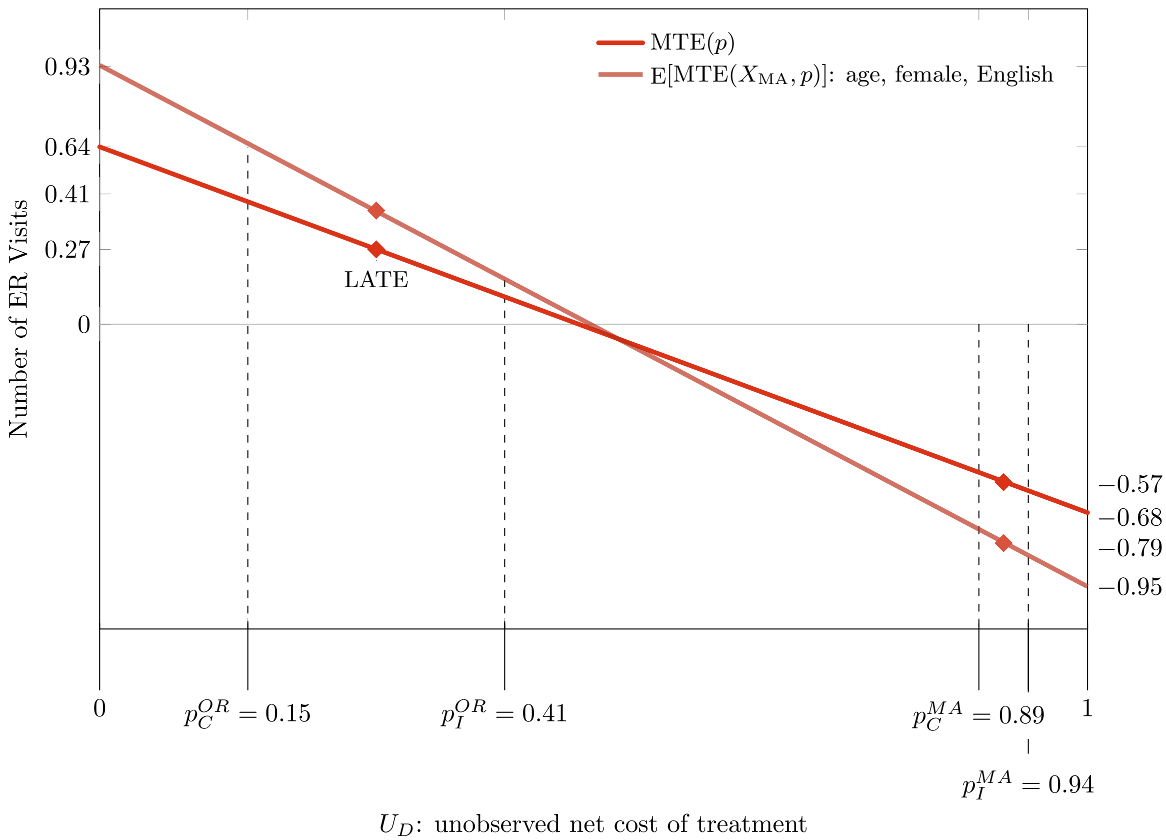
MTE-reweighting with observables can still proceed if there is unexplained heterogeneity in the treatment effect. To obtain a LATE for the Massachusetts reform by reweighting the Oregon MTE with common observables, I estimate the average $\text{MTE}(x, p)$ for compliers in Massachusetts and construct $\text{E}[\text{MTE}(X_{\text{MA}}, p)]$. In Figure 9, I plot $\text{E}[\text{MTE}(X_{\text{MA}}, p)]$. Reweighting the Oregon MTE to predict the impact of the Massachusetts reform on ER utilization, I apply (12) using the pre-reform level of coverage in Massachusetts p_C^{MA} and the post-reform level of coverage in Massachusetts p_I^{MA}. I predict that the Massachusetts reform will decrease emergency room utilization for compliers by 0.79 visits over an approximately 19-month period. Translating this decrease into an annual decrease, I predict a decrease of 0.50 visits per person per year (=(0.79/19)*12). This prediction is even closer to the Miller (2012) estimates of 0.67 to 1.28 than the decrease that I predict without incorporating common observables using $\text{MTE}(p)$, which I also plot for comparison.

Figure 9 illustrates that accounting for differences in the unobservable U_D between Oregon and Massachusetts has a much larger impact than accounting for differences in common observables between Oregon and Massachusetts. If I account for the observables of Massachusetts compliers with $\text{E}[\text{MTE}(X_{\text{MA}}, p)]$, but do not account for range of U_D for Massachusetts compliers, then I predict a Massachusetts LATE of 0.41, which is even more positive than the LATE of 0.27 estimated in Oregon. Such an approach, which can be considered a form of LATE-reweighting, does not reconcile the positive LATE in Oregon with the negative LATE in Massachusetts, given that common observables do not explain treatment effect heterogeneity across U_D in Oregon. This finding demonstrates that the power of LATE-weighting to reconcile results across contexts is limited by the common observables available for reweighting. However, MTE-reweighting with the common observables can still reconcile the positive treatment effect induced by the Oregon experiment with the negative treatment effect induced by the Massachusetts reform.

4 Conclusion

I aim to shed light on why emergency room (ER) utilization increased following the Oregon Health Insurance Experiment but decreased following the Massachusetts reform. Starting from the Oregon Health Insurance Experiment as the "gold standard," I find treatment effect heterogeneity across the unobservable that separates compliers from other groups: although Oregon compliers increase their ER utilization upon gaining coverage, Oregon never

Figure 9: Extrapolation of MTE(x, p) to Massachusetts



Note. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. "Age" is measured in year 2008 for the Oregon Health Insurance Experiment and in year 2006 for the Massachusetts health reform. "English" is an indicator variable for individuals in the Oregon Health Insurance Experiment who requested materials in English and that equals one for individuals in the BRFSS who completed the interview in English. The specification with common covariates (age, female, English) includes all two-way interactions. p_C^{OR} is the probability of treatment in the control group in Oregon, p_I^{OR} the probability of treatment in the intervention group in Oregon, p_C^{MA} the probability of treatment in the control group in the Massachusetts reform, and p_I^{MA} the probability of treatment in the intervention group in the Massachusetts reform.

takers would decrease their ER utilization upon gaining coverage. I also find heterogeneous selection: Oregon never takers report better health than Oregon compliers.

I extrapolate my findings from within the Oregon experiment to the Massachusetts reform. Given higher levels of coverage in Massachusetts, Massachusetts compliers are comparable to a subset of Oregon never takers. Like Oregon never takers, Massachusetts compliers report better health than Oregon compliers. Therefore, even though the results seem contradictory, I can reconcile the increase in ER utilization induced by the Oregon Health Insurance Experiment with the decrease in ER utilization induced by the Massachusetts reform.

Appendix

Appendix A   Proof that U_D is uniformly distributed between 0 and 1

Per the "probability integral transformation" (see Casella and Berger (2002, page 54)), the cumulative distribution function of any random variable applied to itself must be distributed uniformly between 0 and 1. Therefore, the uniformity of U_D is not a separate assumption of the model. A random variable Y is distributed uniformly between 0 and 1 if and only if $F_Y(c) = c$ for $0 \leq c \leq 1$. Therefore, the following shows that U_D is distributed uniformly between 0 and 1, where I omit conditioning on X for simplicity:

$$\begin{aligned}
F_{U_D}(u) &= P(U_D \leq u) \\
&= P(F(\nu_D) \leq u) \\
&= P(\nu_D \leq F^{-1}(u)) \\
&= F(F^{-1}(u)) = u. && (F \text{ absolutely continuous under A.1})
\end{aligned}$$

■

Appendix B   Derivation of the Treatment Equation

Medicaid enrollment D is given by

$$\begin{aligned}
D &= 1\{0 \leq V_T - V_U\} \\
&= 1\{0 \leq \mu_D(Z, X) - \nu_D\} \\
&= 1\{\nu_D \leq \mu_D(Z, X)\} \\
&= 1\{F(\nu_D \mid X) \leq F(\mu_D(Z, X) \mid X)\} && (\text{definition of } F(\cdot \mid X) \text{ from A.1}) \\
&= 1\{U_D \leq F(\mu_D(Z, X) \mid X)\} && (U_D = F(\nu_D \mid X) \text{ by definition}) \\
&= 1\{U_D \leq \mathrm{P}(D = 1 \mid Z = z, X)\},
\end{aligned}$$

where the last equality follows from

$$\begin{aligned}
F(\mu_D(Z, X) \mid X) &= \mathrm{P}(\nu_D \leq \mu_D(Z, X) \mid X) \\
&= \mathrm{P}(\nu_D \leq \mu_D(z, X) \mid Z = z, X) && (\nu_D \perp Z \mid X \text{ by A.2}) \\
&= \mathrm{P}(0 \leq \mu_D(z, X) - \nu_D \mid Z = z, X) \\
&= \mathrm{P}(0 \leq V_T - V_U \mid Z = z, X) \\
&= \mathrm{P}(D = 1 \mid Z = z, X).
\end{aligned}$$

■

Appendix C Derivation of Average Outcomes and Observables

Imbens and Rubin (1997), Katz et al. (2001), Abadie (2002), and Abadie (2003) rely on the LATE assumptions to calculate average outcomes and observables of always takers, compliers, and never takers. For consistency with my exposition, I perform the same calculations using the MTE model that assumes no more than the LATE assumptions. I build intuition with a graphical illustration that follows from the model.

I identify the expected value of Y_T for always takers as follows, supressing X for simplicity:

$$\begin{aligned}
\mathrm{E}[Y \mid D=1, Z=0] &= \mathrm{E}[Y_U + D(Y_T - Y_U) \mid D=1, Z=0] && \text{(by (6))} \\
&= \mathrm{E}[Y_T \mid D=1, Z=0] \\
&= \mathrm{E}[Y_T \mid 0 \le U_D \le p_C, Z=0] && \text{(by (5), where } p_C = P(D=1|Z=0)) \\
&= \mathrm{E}[g_T(U_D, \gamma_T) \mid 0 \le U_D \le p_C, Z=0] && \text{(by (7))} \\
&= \mathrm{E}[g_T(U_D, \gamma_T) \mid 0 \le U_D \le p_C] && (Z \perp (U_D, \gamma_T) \text{ by (A.2)}) \\
&= \mathrm{E}[Y_T \mid 0 \le U_D \le p_C].
\end{aligned}$$

I use similar steps to calculate the expected value of Y_T for lottery winners enrolled in Medicaid $\mathrm{E}[Y_T \mid 0 \le U_D \le p_I] = \mathrm{E}[Y \mid D=1, Z=1]$, the expected value of Y_U for never takers $\mathrm{E}[Y_U \mid p_I < U_D \le 1] = \mathrm{E}[Y \mid D=0, Z=1]$, and the expected value of Y_U for lottery losers not enrolled in Medicaid $\mathrm{E}[Y_U \mid p_C < U_D \le 1] = \mathrm{E}[Y \mid D=0, Z=0]$. I then use the four resulting values to calculate the expected value of Y_T for compliers enrolled in Medicaid:

$$\begin{aligned}
\mathrm{E}[Y_T \mid p_C < U_D \le p_I] &= \frac{p_I}{p_I - p_C}\mathrm{E}[Y_T \mid 0 \le U_D \le p_I] - \frac{p_C}{p_I - p_C}\mathrm{E}[Y_T \mid 0 \le U_D \le p_C] \\
&= \frac{p_I}{p_I - p_C}\mathrm{E}[Y_T \mid D=1, Z=1] - \frac{p_C}{p_I - p_C}\mathrm{E}[Y_T \mid D=1, Z=0].
\end{aligned}$$

and the expected value of Y_U for compliers not enrolled in Medicaid:

$$\begin{aligned}
\mathrm{E}[Y_U \mid p_C < U_D \le p_I] &= \frac{1-p_C}{p_I - p_C}\mathrm{E}[Y_U \mid p_C < U_D \le 1] - \frac{1-p_I}{p_I - p_C}\mathrm{E}[Y_U \mid p_I < U_D \le 1] \\
&= \frac{1-p_C}{p_I - p_C}\mathrm{E}[Y_U \mid D=0, Z=0] - \frac{1-p_I}{p_I - p_C}\mathrm{E}[Y_U \mid D=0, Z=1]
\end{aligned}$$

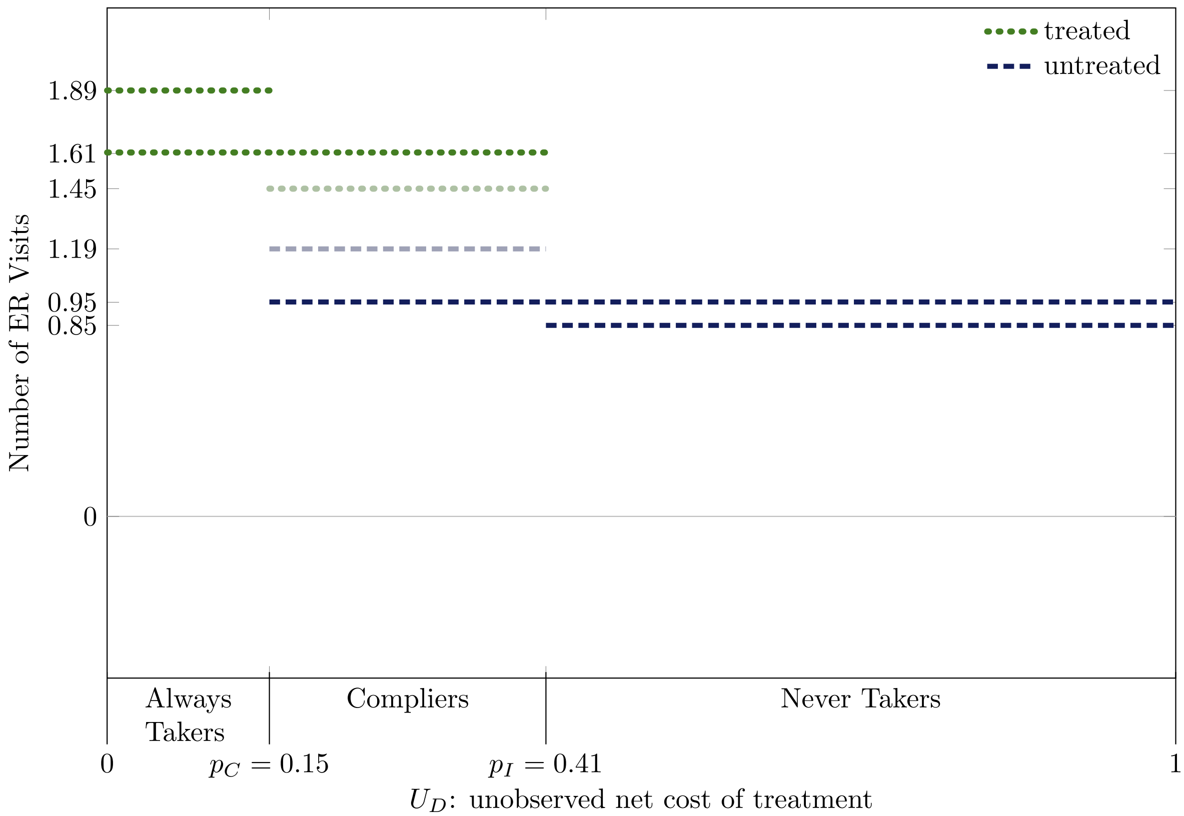
I illustrate the calculations graphically using values from Oregon data in Figure C1. I use bolded dotted lines to depict average ER utilization when enrolled in Medicaid, Y_T, for two observed groups: lottery losers enrolled in Medicaid ($0 \le U_D \le p_C$) and lottery winners enrolled in Medicaid ($0 \le U_D \le p_I$). I use bolded dashed lines to depict average ER utilization when not enrolled in Medicaid, Y_U, for two observed groups: lottery losers not enrolled in Medicaid ($p_C < U_D \le 1$) and lottery winners not enrolled in Medicaid

($p_I < U_D \leq 1$). I depict the calculated outcomes for compliers with lighter shading.

To calculate the average observable X for each group, I begin with the same approach. Even though average *outcomes* of compliers should depend on whether they win or lose the lottery, average *observables* of compliers should not. Therefore, I weight the average observables of compliers who win and lose the lottery by their respective probabilities:

$$\begin{aligned} \mathrm{E}[X \mid p_C < U_D \leq p_I] = \mathrm{P}(Z=1)\Big[\frac{p_I}{p_I - p_C}\mathrm{E}[X \mid D=1, Z=1] - \frac{p_C}{p_I - p_C}\mathrm{E}[X \mid D=1, Z=0]\Big] \\ +\mathrm{P}(Z=0)\Big[\frac{1-p_C}{p_I - p_C}\mathrm{E}[X \mid D=0, Z=0] - \frac{1-p_I}{p_I - p_C}\mathrm{E}[X \mid D=0, Z=1]\Big]. \end{aligned}$$

Figure C1: Average Treated and Untreated ER Visits for Compliers (Ligher Shading)



Note. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. Treatment represents enrollment in Medicaid. p_C is the probability of treatment in the control group, and p_I is the probability of treatment in the intervention group. Some differences between statistics might not appear internally consistent because of rounding.

Appendix D Estimating MTO(x,p), MUO(x,p), and MTE(x,p).

The steps below estimate the functions MTO(x,p), MUO(x,p), and MTE(x,p) of the form

$$\begin{aligned}
\mathrm{MTO}(x,p) &= \delta_T' x + \lambda_T p \\
\mathrm{MUO}(x,p) &= \delta_U' x + \lambda_U p \\
\mathrm{MTE}(x,p) &= \mathrm{MTO}(x,p) - \mathrm{MUO}(x,p) \\
&= (\delta_T' - \delta_U')\, x + (\lambda_T - \lambda_U)\, p.
\end{aligned}$$

1. Estimate propensity scores, $\widehat{p}$, for all individuals in the sample by fitting

$$D = \phi_0 + \phi_1 Z + \phi_2' X + \phi_3'(X'Z) + \varepsilon$$

   and using $\widehat{\phi}_0$, $\widehat{\phi}_1$, $\widehat{\phi}_2$, and $\widehat{\phi}_3$ to predict D conditional on Z and observables X.

2. The MTO function can be derived from the average treated outcome (ATO) function, defined as follows:

$$\begin{aligned}
\mathrm{ATO}(x,p) &= \mathrm{E}\left[Y_T \mid X = x, 0 \le U_D \le p\right] \\
&= \widetilde{\delta}_T' x + \widetilde{\lambda}_T p.
\end{aligned}$$

   The ATO function can be estimated directly by conditioning the sample on treated individuals ($D = 1$) and using OLS to estimate:

$$Y = \widetilde{\delta}_T' x + \widetilde{\lambda}_T \widehat{p} + \zeta_T.$$

   To recover the parameters of the MTO function from the estimated parameters of the ATO function, note that:

$$\mathrm{MTO}(x,p) = \frac{\mathrm{d}\left[p\mathrm{ATO}(x,p)\right]}{\mathrm{d}p}.$$

   Therefore,

$$\begin{aligned}
\mathrm{MTO}(x,p) &= \widetilde{\delta}_T' x + 2\widetilde{\lambda}_T p \\
&= \delta_T' x + \lambda_T p.
\end{aligned}$$

   So, estimates of the MTO parameters can be constructed as follows: $\delta_T = \widetilde{\delta}_T$ and $\lambda_T = 2\widetilde{\lambda}_T$.

3. The MUO function can be derived from the average untreated outcome (AUO) function, defined as follows:

$$\begin{aligned}\mathrm{AUO}(x,p) &= \mathrm{E}\left[Y_U \mid X=x, p<U_D\leq 1\right] \\ &= \widetilde{\delta}'_U x + \widetilde{\lambda}_U p.\end{aligned}$$

The AUO function can be estimated directly by conditioning the sample on untreated individuals ($D=0$) and using OLS to estimate:

$$Y = \widetilde{\delta}'_U x + \widetilde{\lambda}_U \widehat{p} + \zeta_U.$$

To recover the parameters of the MUO function from the estimated parameters of the AUO function, note that

$$\mathrm{MUO}(x,p) = \frac{\mathrm{d}\left[(1-p)\mathrm{AUO}(x,p)\right]}{\mathrm{d}(1-p)}.$$

Therefore,

$$\begin{aligned}\mathrm{MUO}(x,p) &= \widetilde{\delta}'_U x - \widetilde{\lambda}_U + 2\widetilde{\lambda}_U p \\ &= \delta'_U x + \lambda_U p.\end{aligned}$$

So, an estimate for λ_U can be constructed as $\lambda_U = 2\widetilde{\lambda}_U$, while the estimate for δ_U is equal to the estimated $\widetilde{\delta}_U$ with its constant coefficient shifted down by $\widetilde{\lambda}_U$.

4. Construct the estimate for $\mathrm{MTE}(x,p)$ using the estimated parameters of $\mathrm{MTO}(x,p)$ and $\mathrm{MUO}(x,p)$:

$$\mathrm{MTE}(x,p) = \mathrm{MTO}(x,p) - \mathrm{MUO}(x,p) = (\delta_T - \delta_U)'x + (\lambda_T - \lambda_U)p.$$

References

Alberto Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American statistical Association*, 97(457):284–292, 2002.

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263, 2003.

JD Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288, 1998.

Joshua D Angrist. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pages 313–336, 1990.

Joshua D Angrist and Ivan Fernandez-Val. ExtrapoLATE-ing: External validity and. In *Advances in Economics and Econometrics: Volume 3, Econometrics: Tenth World Congress*, volume 51, page 401. Cambridge University Press, 2013.

Joshua D Angrist and Alan B Krueger. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American statistical Association*, 87(418):328–336, 1992.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Katherine Baicker, Sarah Taubman, Heidi L. Allen, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Eric C. Schneider, Bill J. Wright, Alan M. Zaslavsky, and Amy N. Finkelstein. The Oregon experiment - effects of medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18):1713–1722, 2013.

Katherine Baicker, Amy Finkelstein, Jae Song, and Sarah Taubman. The impact of medicaid on labor market activity and program participation: Evidence from the Oregon health insurance experiment. *American Economic Review*, 104(5):322–28, 2014.

Marinho Bertanha and Guido W. Imbens. External validity in fuzzy regression discontinuity designs. Working Paper 20773, National Bureau of Economic Research, December 2014.

Anders Björklund and Robert Moffitt. The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, pages 42–49, 1987.

Dan A Black, Joonhwi Joo, Robert LaLonde, Jeffrey A Smith, and Evan J Taylor. Simple tests for selection bias: Learning more from instrumental variables. Working Paper 6932, CESifo, March 2017. URL https://www.cesifo-group.de/DocDL/cesifo1_wp6392.pdf.

Christian N Brinch, Magne Mogstad, and Matthew Wiswall. Beyond LATE with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039, 2017.

Pedro Carneiro and Sokbae Lee. Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, 149(2):191–208, 2009.

Pedro Carneiro, James J. Heckman, and Edward J. Vytlacil. Estimating marginal returns to education. *American Economic Review*, 101(6):2754–81, October 2011.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Christopher Chen, Gabriel Scheffler, and Amitabh Chandra. Massachusetts' health care reform and emergency department utilization. *New England Journal of Medicine*, 365 (12):e25, 2011.

Amy Finkelstein. Oregon health insurance experiment public use data. 2013. URL http://www.nber.org/oregon/4.data.html.

Amy F. Finkelstein, Sarah Taubman, Bill J. Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi L. Allen, Katherine Baicker, and the Oregon Health Study Group. The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106, 2012.

Amy N. Finkelstein, Sarah L. Taubman, Heidi L. Allen, Bill J. Wright, and Katherine Baicker. Effect of medicaid coverage on ed use — further evidence from Oregon's experiment. *New England Journal of Medicine*, 375(16):1505–1507, 2016. PMID: 27797307.

Eric French and Jae Song. The effect of disability insurance receipt on labor supply. *American Economic Journal: Economic Policy*, 6(2):291–337, 2014.

Zijian Guo, Jing Cheng, Scott A Lorch, and Dylan S Small. Using an instrumental variable to test for unmeasured confounding. *Statistics in medicine*, 33(20):3528–3546, 2014.

Martin B. Hackmann, Jonathan T. Kolstad, and Amanda E. Kowalski. Adverse selection and an individual mandate: When theory meets practice. *American Economic Review*, 105(3):1030–66, 2015.

James Heckman, Sergio Urzua, and Edward Vytlacil. Estimation of treatment effects under essential heterogeneity. 2006.

James J. Heckman and Edward Vytlacil. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73(3):669–738, 05 2005.

James J. Heckman and Edward J. Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, 96(8):4730–4734, 1999.

James J. Heckman and Edward J. Vytlacil. Local instrumental variables. In Cheng Hsiao, Kimio Morimune, and James L. Powell, editors, *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, pages 1–46. Cambridge University Press, 2001.

James J Heckman and Edward J Vytlacil. 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. *Handbook of econometrics*, 6:4875–5143, 2007.

James J. Heckman, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098, 1998.

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

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–75, 1994.

Guido W Imbens and Donald B Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997.

Lawrence F Katz, Jeffrey R Kling, Jeffrey B Liebman, et al. Moving to opportunity in boston: Early results of a randomized mobility experiment. *The Quarterly Journal of Economics*, 116(2):607–654, 2001.

Jonathan T. Kolstad and Amanda E. Kowalski. The impact of health care reform on hospital and preventive care: Evidence from massachusetts. *Journal of Public Economics*, 96:909–929, December 2012.

Amanda Kowalski. Doing more when you're running LATE: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments. Working Paper 22362, National Bureau of Economic Research, June 2016. URL http://www.nber.org/papers/w22362.

Amanda Kowalski, Yen Tran, and Ljubica Ristovska. MTEBINARY: Stata module to compute Marginal Treatment Effects (MTE) With a Binary Instrument. Statistical Software Components, Boston College Department of Economics, December 2016. URL https://ideas.repec.org/c/boc/bocode/s458285.html.

Amanda Kowalski, Yen Tran, and Ljubica Ristovska. MTEMORE: Stata module to compute Marginal Treatment Effects (MTE) With a Binary Instrument. Statistical Software Components, Boston College Department of Economics, July 2018. URL https://ideas.repec.org/c/boc/bocode/s458503.html.

Nicole Maestas, Kathleen J Mullen, and Alexander Strand. Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *The American Economic Review*, 103(5):1797–1829, 2013.

Sarah Miller. The effect of insurance on emergency room visits: an analysis of the 2006 massachusetts health reform. *Journal of Public Economics*, 96(11):893–908, 2012.

Robert Moffitt. Estimating marginal treatment effects in heterogeneous populations. *Annales d'Economie et de Statistique*, pages 239–261, 2008.

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using instrumental variables for inference about policy relevant treatment effects. *Econometrica*, 86(5):1589–1619, 2018.

Randall J Olsen. A least squares correction for selectivity bias. *Econometrica: Journal of the Econometric Society*, pages 1815–1820, 1980.

Peter B. Smulowitz, Robert Lipton, J. Frank Wharam, Leon Adelman, Scott G. Weiner, Laura Burke, Christopher W. Baugh, Jeremiah D. Schuur, Shan W. Liu, Meghan E. McGrath, Bella Liu, Assaad Sayah, Mary C. Burke, J. Hector Pope, and Bruce E. Landon. Emergency department utilization after the implementation of massachusetts health reform. *Annals of Emergency Medicine*, 58(3):225 – 234.e1, 2011.

Sarah L. Taubman, Heidi L. Allen, Bill J. Wright, Katherine Baicker, and Amy N. Finkelstein. Medicaid increases emergency-department use: Evidence from Oregon's health insurance experiment. *Science*, 343(6168):263–268, 2014.

Sarah Tavernise. Emergency visits seen increasing with health law. *New York Times*, 2014. URL http://www.nytimes.com/2014/01/03/health/access-to-health-care-may-increase-er-visits-study-suggests.html.

Edward Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341, 2002.

Abraham Wald. The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.*, 11(3):284–300, 09 1940.