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 01238
May 2018, Revised March 2020

Previously circulated as "Extrapolation using Selection and Moral Hazard Heterogeneity from within the Oregon Health Insurance Experiment." This paper includes material from Kowalski (2016). Other material from Kowalski (2016) involving weak monotonicity assumptions appears in Kowalski (2018). 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, Simon Essig Aberg, 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, Edward Nor- ton, 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. I am grateful to the investigators of the Oregon Health Insurance Experiment for making their data publicly available. By using publicly available data, I encourage replication and future work. 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 March 2020
JEL No. C1,H75,I10,I13

ABSTRACT

A headline result from the Oregon Health Insurance Experiment is that emergency room (ER) utilization increased. A seemingly contradictory result from the Massachusetts health reform is that ER utilization decreased. I reconcile both results by identifying treatment effect heterogeneity within the Oregon experiment and then extrapolating it to Massachusetts. I find that even though Oregon compliers increased their ER utilization, they were sicker than Oregon never takers, who would have decreased their ER utilization. Massachusetts expanded coverage from a higher level to healthier compliers. Therefore, Massachusetts compliers are comparable to a subset of Oregon never takers, which can reconcile the results.

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 a 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 may value additional ER care below 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 for how insurance expansions affect ER utilization. Discussion of the Oregon experiment and the Massachusetts reform in the *New York Times* has done just that (Tavernise, 2014).

However, when results from two experiments give different answers, it need not be that one experiment must be flawed. Instead, it could be 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 further developed by Heckman and Vytlacil (1999, 2001, 2005), Carneiro et al. (2011), Brinch et al. (2017), and Cornelissen et al. (2018), among others, then it is possible to use the MTE function to recover the two different LATEs, thereby reconciling the results. Although the MTE literature generally focuses on a single context, I use treatment effect heterogeneity that I find within Oregon to reconcile results across the Oregon and Massachusetts contexts.

To do so, I use a model. To be transparent about the assumptions of the model, I relate it to the MTE literature and to the LATE assumptions of Imbens and Angrist (1994). In the terminology of Angrist et al. (1996), 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 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. Future policies that expand coverage could enroll never takers, and future policies that contract coverage could disenroll always takers. Even though the treatment effect on compliers is relevant for the policy implemented by the experiment, the model informs how treatment effects on always and never takers could be relevant for other policies.

I contribute to the literature by presenting my results using simple figures derived from the model. These figures emphasize that the LATE assumptions imply an ordering from always takers to compliers to never takers, originally shown by Imbens and Rubin (1997) and Vytlacil (2002). I also contribute to the literature by using this ordering along with additional data on always takers, compliers, and never takers to motivate ancillary assumptions of the model beyond the LATE assumptions. I also provide an interpretation for these assumptions in terms of selection and treatment effect heterogeneity. However, my main contribution is that I use the model to reconcile two seemingly contradictory results from two important experiments.

To reconcile the positive LATE from the Oregon experiment with the negative LATE from the Massachusetts reform, I proceed in three steps. First, starting with the Oregon experiment as the "gold standard," I find that health improves from always takers to compliers to never takers in terms of ER utilization, previous ER utilization, and self-reported health, indicating adverse selection. I also find that the treatment effect on ER utilization decreases as health improves such that it is positive for Oregon compliers but negative for Oregon never takers. Second, I document similar patterns in adverse selection and treatment effect heterogeneity within Massachusetts. Third, I compare observables across states and extrapolate from Oregon to Massachusetts. Age, sex, and English-speaking status do not explain treatment effect heterogeneity in Oregon and cannot reconcile the results. However, I demonstrate that Massachusetts compliers are healthier than Oregon compliers and have self-reported health comparable to a subset of Oregon never takers, which justifies extrapolation from Oregon never takers to Massachusetts compliers with the Oregon MTE function. Reweighting the Oregon MTE function to yield an average treatment effect for Massachusetts compliers, I predict a decrease in ER utilization for Massachusetts compliers of the same order of magnitude as the decrease found by Miller (2012), thereby reconciling the positive LATE in Oregon with the negative LATE in Massachusetts.

I do not claim that my approach is the only potential way to reconcile the Oregon and Massachusetts findings. However, it is plausible, especially since it is consistent with differences in self-reported health between the sicker compliers in Oregon and the healthier compliers in Massachusetts. Importantly, however, the reconciliation does not depend only on observables, which vary in their ability to explain treatment effect heterogeneity and their availability in both settings. 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, which is related to observable differences in self-reported health but not related to differences in other available observables. By being explicit about how the MTE and LATE frameworks inform reconciliation, I build firmer footing for reconciliation of findings from different experiments.

2 Model

I begin with a model shown by Vytlacil (2002) to assume no more than the LATE assumptions, and then I add ancillary assumptions. To ensure that I do not introduce any ancillary assumptions before doing so explicitly, I follow the exposition from Heckman and Vytlacil (2005) closely. However, I adapt the model to my empirical context, and I build intuition using 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:

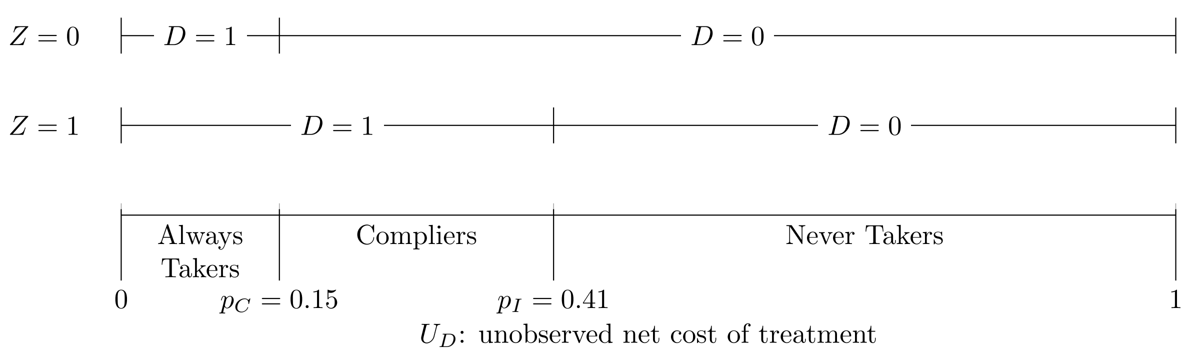
$$D = 1\{U_D \leq p_{CX}\} \qquad \text{where } p_{CX} = P\big(D = 1 \mid Z = 0, X\big), \tag{4}$$

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

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 Imbens and Rubin (1997) and 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 $p_I = 0.41$, omitting the X subscript. In the top line of Figure 1, I depict the lottery losers. By (4), treated enrolled lottery losers have $0 \leq U_D \leq 0.15$. Treated lottery losers are always takers. In the middle line of Figure 1, I depict the lottery winners. By (5), the untreated lottery winners have $0.41 < U_D \leq 1$. Untreated lottery winners are never takers. 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 are compliers.

Figure 1: Ranges of U_D Show Ordering from Always Takers to Compliers to Never Takers



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, all of which 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 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 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, 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.

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. 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 publicly available data 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 with the headline finding of Taubman et al. (2014), who show that insurance increases ER utilization for compliers.[1] 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. Figure 2 also includes average outcomes for always and never 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 on Y along U_D using the following functions:

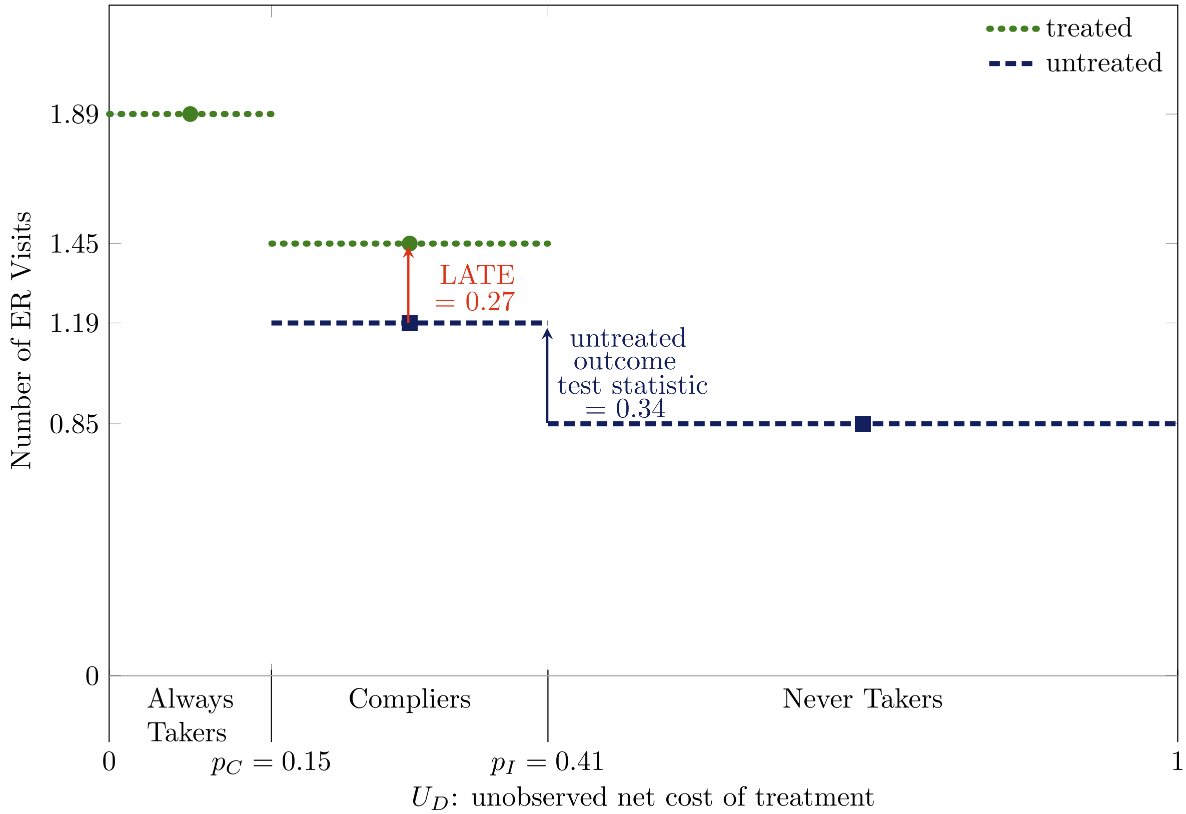
$$\begin{aligned}
\text{Selection Heterogeneity:} \quad & \text{MUO}(x,p) = \mathrm{E}\left[Y_U \mid X = x, U_D = p\right] \\
\text{Treatment Effect Heterogeneity:} \quad & \text{MTE}(x,p) = \mathrm{E}\left[Y_T - Y_U \mid X = x, U_D = p\right] \\
\text{Selection + Treatment Effect Heterogeneity:} \quad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

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 second 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

[1] 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). 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 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.

Figure 2: Untreated Outcome Test Shows Adverse Selection on ER Utilization:
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.

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:

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

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 without ancillary assumptions.

Turning to the last function, which I refer to as the "marginal treated outcome (MTO)" function, I emphasize that it defines the *sum* of selection heterogeneity plus treatment effect heterogeneity. The literature considers the MTO and MUO functions as intermediate inputs in the derivation of the MTE function, but it does not make a meaningful distinction between them (see Brinch et al., 2017). It is tempting to assert that there should be no meaningful distinction between the MTO and MUO functions because it should 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, creating a different asymmetry between the MTO and 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 can be important.

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. Selection heterogeneity is adverse in terms of ER utilization and in terms of observed measures of health. Treatment effect heterogeneity 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 a similar pattern in selection and treatment effect heterogeneity within Massachusetts. Third, I find that treatment effect heterogeneity within Oregon can reconcile the positive LATE in Oregon with the negative LATE in Massachusetts because Massachusetts compliers are comparable to a subset of Oregon never takers.

3.1 Heterogeneity Within Oregon

3.1.1 Adverse Selection on ER Utilization

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).[2] Relative to the literature, my innovation with respect to the untreated outcome test is that I show

[2] 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 their 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.

Table 1: Untreated Outcome Test Shows Adverse Selection on ER Utilization: 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)
Number of ER Visits				
Treated	1.89	1.45	0.55	
	(0.08)	(0.11)	(0.45)	
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.

that it identifies selection heterogeneity without any assumptions beyond the LATE assumptions. 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 2 and Table 1, 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, is statistically different from zero, according to the bootstrapped standard error computed as the standard deviation of the statistic over 200 replications. Under the model, compliers enroll in Medicaid before never takers, so the selection heterogeneity that I find indicates adverse selection along the unobservable that separates compliers from never takers.

3.1.2 Adverse Selection on Self-Reported Health and Previous ER Utilization

Next, I characterize adverse selection in terms of observables. To do so, I define selection heterogeneity on X along U_D:

$$\text{Selection Heterogeneity on } X: \quad \mathrm{E}\left[X \mid U_D = p\right],$$

which captures how the covariate vector X changes with the unobserved net cost of treatment U_D. I identify selection heterogeneity on X by comparing the average covariate vector of always takers, compliers, and never takers, obtained as described in Appendix C. Just like selection heterogeneity on Y, selection heterogeneity on a single covariate can be either positive or negative between compliers and never takers. It can also be either positive or negative between always takers and compliers.

I begin by examining self-reported health. I only observe self-reported health for a subset of the individuals in the Oregon administrative data who were surveyed. Self-reported health was elicited after randomization, and Finkelstein et al. (2012) shows that Medicaid improved self-reported health. That is, there is a treatment effect on self-reported health. To ensure that my examination of selection heterogeneity is not contaminated by treatment effect heterogeneity, I only compare the self-reported health of untreated individuals: compliers who lost the lottery and never takers. I obtain the average probability that individuals in these groups reported fair or poor health as described in Appendix C.

As shown in Table 2, within Oregon, I find that never takers are less likely to be in fair or poor health than untreated compliers, and the difference is statistically significant. Therefore, I find adverse selection on self-reported health that is consistent with the adverse selection on ER utilization that I find using the untreated outcome test. This adverse selection indicates that never takers are healthier than compliers, which may be one reason they visit the ER less frequently. It would also be interesting to examine whether there is adverse selection from always takers to compliers, such that compliers are healthier than always takers. However, I do not observe untreated self-reported health for always takers.

To that end, I turn to previous ER utilization as an alternative proxy for health because I observe it for all individuals, including always takers. 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 pre-period ER utilization for always takers, compliers, and never takers by cumulative visits in Table 2. As reported, 45% of always takers, 35% of compliers, and 31% of never takers had at least one previous ER visit, and the differences across groups are statistically significant. Therefore, I find adverse selection on previous ER utilization, not just from compliers to never takers, but also from always takers to compliers.

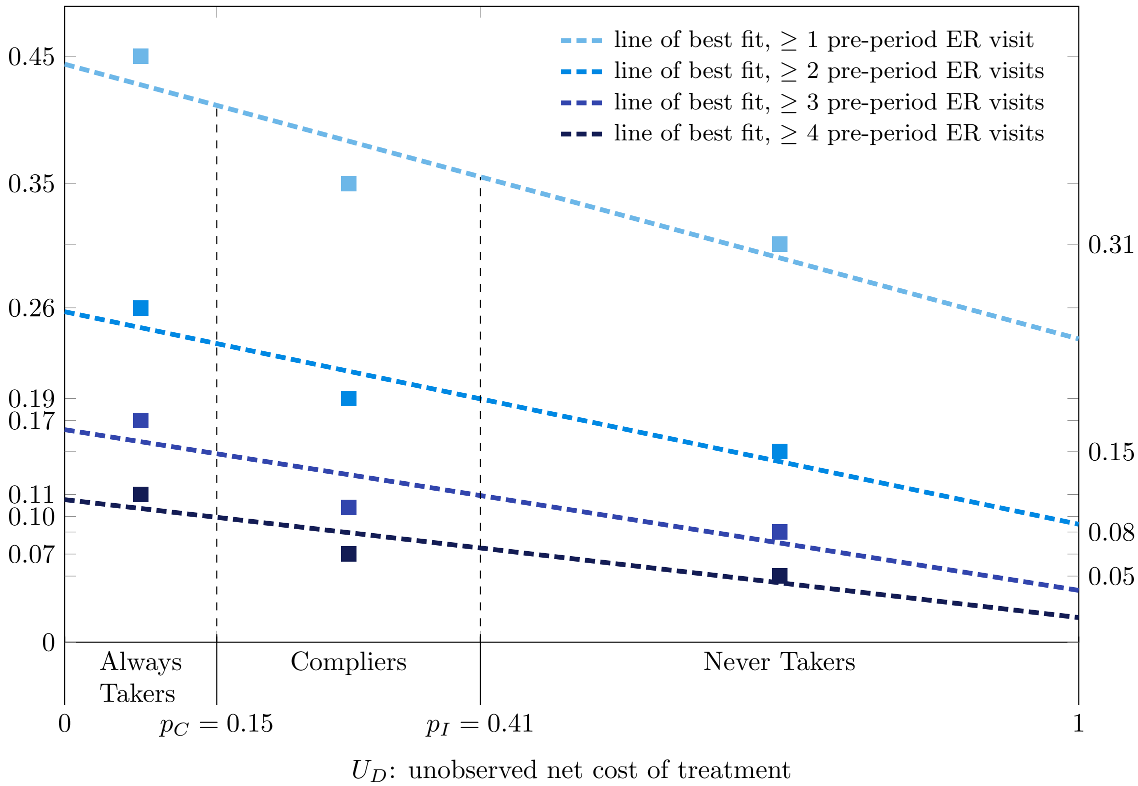
To further illustrate how previous ER utilization varies from always takers to compliers to never takers, Figure 3 plots statistics on previous ER utilization from Table 2. Always takers are more likely to have at least one, at least two, at least three, and at least four previous ER visits

Table 2: Adverse Selection on Self-Reported Health and Previous ER Utilization Within Oregon and 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)
≥ 1 Pre-period ER Visits	0.34	0.45	0.35	0.31	0.10	0.05
	(0.003)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)
≥ 2 Pre-period ER Visit	0.17	0.26	0.19	0.15	0.07	0.04
	(0.002)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
≥ 3 Pre-period ER Visits	0.10	0.17	0.10	0.08	0.07	0.02
	(0.002)	(0.01)	(0.01)	(0.004)	(0.01)	(0.01)
≥ 4 Pre-period ER Visits	0.06	0.11	0.07	0.05	0.05	0.02
	(0.002)	(0.01)	(0.01)	(0.004)	(0.01)	(0.01)
Other 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)
Other Observables						
Age	42.00	42.15	42.42	38.98	-0.26	3.43
	(0.09)	(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 sex 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. Means obtained as described in Appendix C. [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.

Figure 3: Adverse Selection on Previous ER Utilization Appears Approximately Linear: Average Previous ER Utilization for Always Takers, Compliers, and Never Takers by Number of Pre-Period ER Visits, Relative to Zero Visits



Note. The number of pre-period ER visits represents the number of ER visits before the experiment took place 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.

than compliers. Similar relationships hold between compliers and never takers. Furthermore, the relationships are approximately linear in the unobserved net cost of treatment U_D.

Not all observables exhibit such a clear selection pattern. As shown in Table 2, age, sex, and English-speaking status, exhibit patterns that are not even necessarily monotonic from always takers to compliers to never takers. Therefore, these observables cannot explain adverse selection on ER utilization. In contrast, self-reported health and previous ER utilization, which are both proxies for health, can explain adverse selection on ER utilization, implying that individuals with higher unobserved net cost of treatment U_D tend to be healthier.

3.1.3 Treatment Effect Heterogeneity on ER Utilization

To identify treatment effect heterogeneity on ER utilization along U_D, I make a transparent ancillary assumption beyond the model:

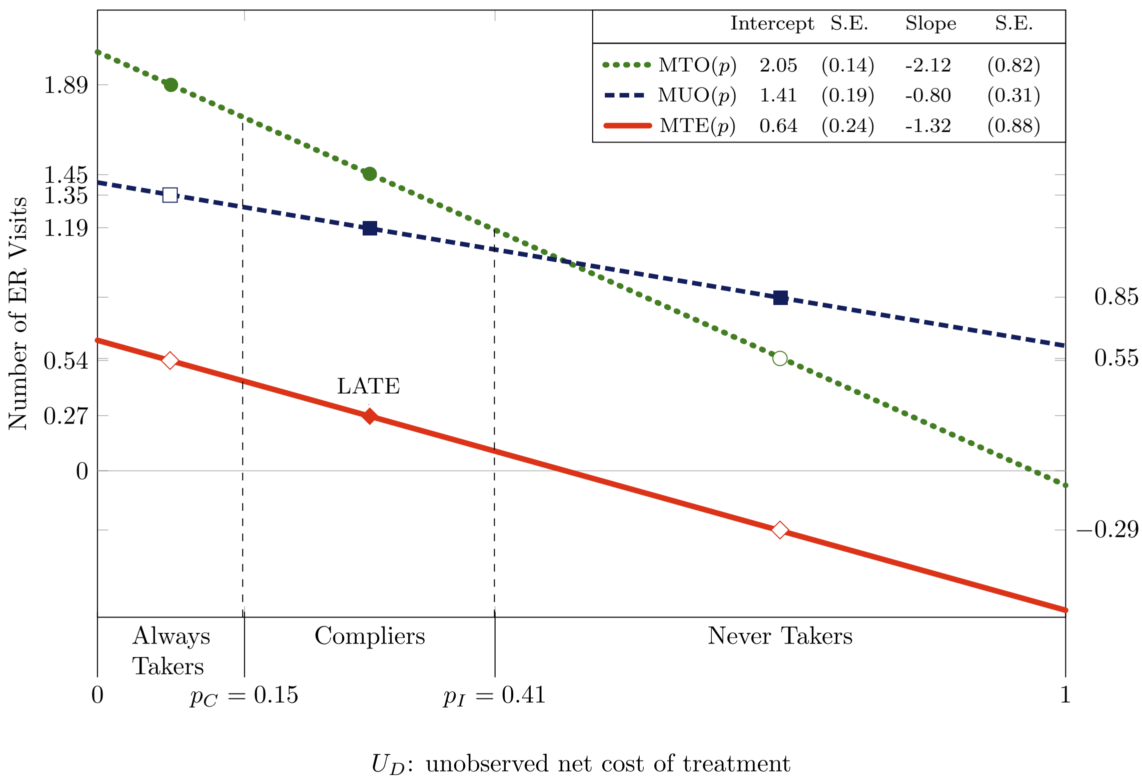
AA.1. (Linear Selection Heterogeneity and Linear Treatment Effect Heterogeneity) In (7) and (8), for $k \in \{T, U\}$, specify $g_k(X, U_D, \gamma_k) = \alpha_k + \beta_k U_D + \gamma_k$, where $\mathrm{E}[\gamma_k \mid U_D = p] = 0$. Therefore,

$$\mathrm{MUO}(p) = \mathrm{E}[Y_U \mid U_D = p] = \alpha_U + \beta_U p$$

$$\text{MTE}(p) = \text{E}\left[Y_T - Y_U \mid U_D = p\right] = (\alpha_T - \alpha_U) + (\beta_T - \beta_U)\,p$$
$$\text{MTO}(p) = \text{E}\left[Y_T \mid U_D = p\right] = \alpha_T + \beta_T p.$$

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 all other policies using the LATE also makes a stronger, implicit assumption that there is no treatment effect heterogeneity, which implies that the MTE function is linear and has zero slope.

Figure 4: Treatment Effect on ER Utilization
Decreases from Always Takers to Compliers to Never Takers



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. Filled markers indicate average values of MUO, MTO, and MTE identified without AA.1, whereas unfilled markers indicate average values of MUO, MTO, and MTE identified with AA.1. Some differences between statistics might not appear internally consistent because of rounding.

In the Oregon context, the empirical findings of adverse selection on ER utilization, self-reported health, and previous ER utilization provide motivation for AA.1. The adverse selection findings imply that healthier individuals have higher unobserved net cost of treatment U_D. Furthermore, selection heterogeneity on previous ER utilization is approximately linear in U_D. Therefore, the assumption that treatment effect heterogeneity is linear in U_D allows the treatment effect to vary with health as selection varies with health. I note, however, that AA.1 allows for selection and treatment effect heterogeneity to have slopes of different signs and magnitudes. I also note that AA.1 allows for the possibility that there is no selection or treatment effect heterogeneity, which occurs when the MUO and MTE slope coefficients are both equal to zero.

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 filled 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 filled 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, the 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 the unobserved net cost of treatment U_D increases.

3.1.4 Previous ER Utilization Explains Treatment Effect Heterogeneity

Given the evidence that health improves as the unobserved net cost of treatment increases, a natural question is whether differences in health explain treatment effect heterogeneity. To answer this question, I quantify how much treatment effect heterogeneity I can explain with observables, particularly those that proxy for health. To do so, I incorporate observables into the MTE function using a shape restriction commonly used in the MTE literature (see Cornelissen et al., 2018; 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 $g_k(X, U_D, \gamma_k) = \delta_k' X + \lambda_k U_D + \gamma_k$, where $\mathrm{E}[\gamma_k \mid X = x, U_D = p] = 0$. Therefore,

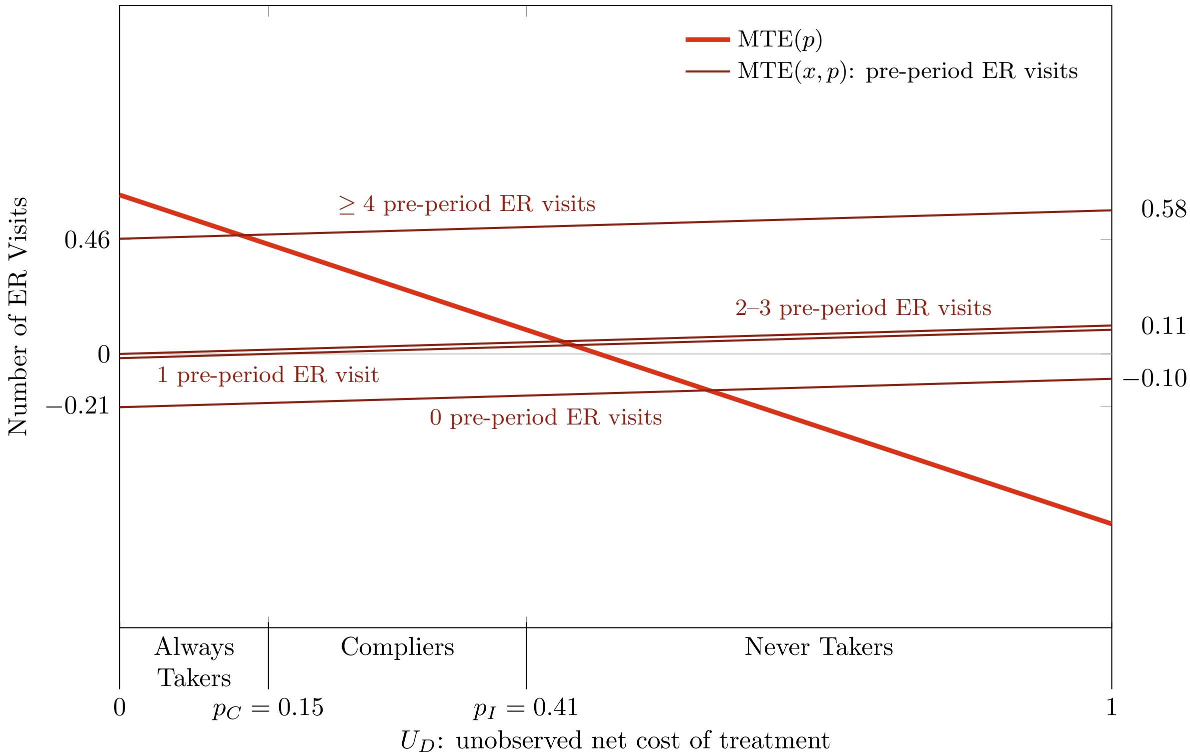
$$\begin{aligned}
\mathrm{MUO}(x,p) &= \mathrm{E}[Y_U \mid X = x, U_D = p] = \delta_U' x + \lambda_U p \\
\mathrm{MTE}(x,p) &= \mathrm{E}[Y_T - Y_U \mid X = x, U_D = p] = (\delta_T - \delta_U)' x + (\lambda_T - \lambda_U) p \\
\mathrm{MTO}(x,p) &= \mathrm{E}[Y_T \mid X = x, U_D = p] = \delta_T' x + \lambda_T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

I present an algorithm for estimation of these functions that simplifies the Heckman et al. (2006) algorithm in Appendix D.

To evaluate whether health explains treatment effect heterogeneity, I incorporate proxies for health into the MTE function. Because self-reported health represents an outcome as opposed to an observable for treated individuals, I do not incorporate it into the MTE function as an observable. However, I do incorporate previous ER utilization into the MTE function because it represents an observable for all individuals.

Incorporating previous ER utilization into the MTE via AA.2, I find that previous ER utilization

Figure 5: Previous ER Utilization Explains Treatment Effect Heterogeneity

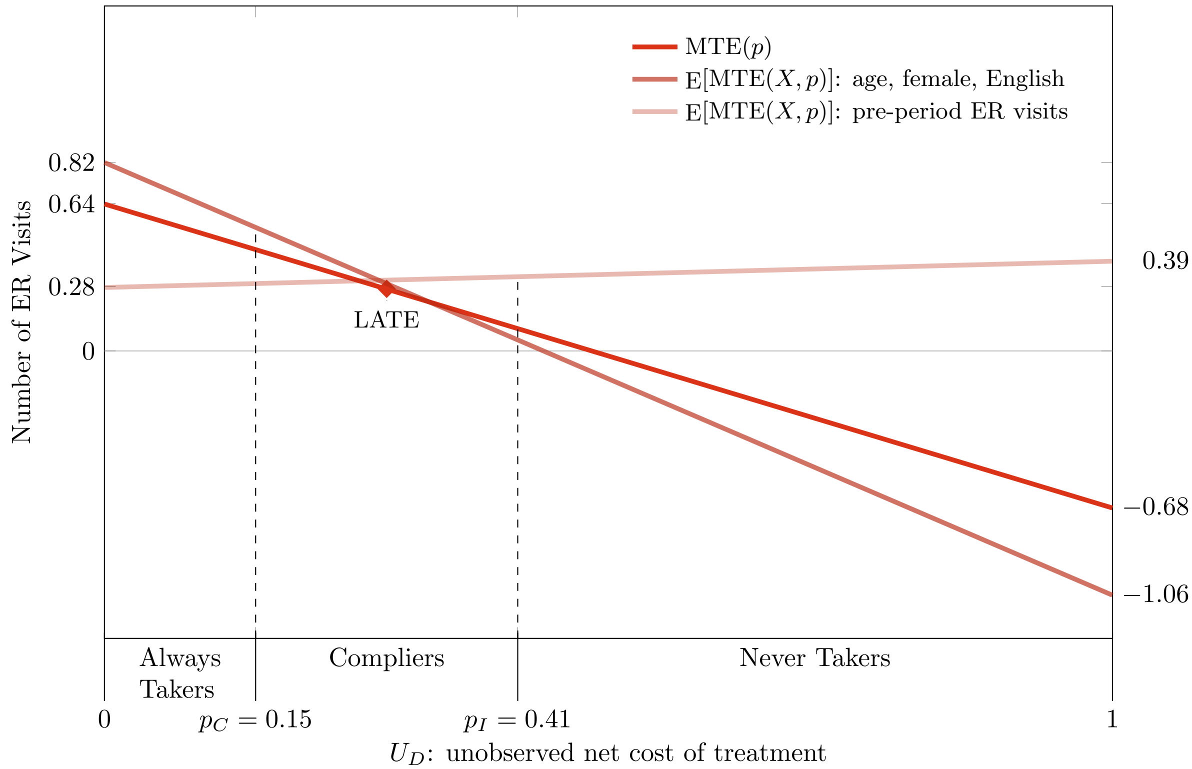


Note. The number of ER visits represents the total number of visits to the emergency department during the study period from March 10, 2008 to September 30, 2009. Pre-period ER visits refers to a group of indicators for visiting the ER 0 times, 1 time, 2–3 times, and 4 or more times during the pre-period 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.

can explain the entire decrease in treatment effect from always takers to compliers to never takers. Incorporating indicators for each of four visit ranges (zero pre-period visits, one pre-period visit, 2 to 3 pre-period visits, and 4 or more pre-period visits) into the MTE, I obtain a separate MTE(x, p) for each visit range. As depicted in Figure 5, 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 slightly positive. The remaining slope in the MTE with observables is not of a meaningful magnitude. Therefore, previous ER utilization can explain all of the treatment effect heterogeneity in MTE(p).

Looking beyond the slope of the MTE function to its level 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 5, 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

Figure 6: Other Observables Do Not Explain Treatment Effect Heterogeneity



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 sex of the respondent. "English" is a binary indicator that equals one for individuals who requested materials in English. The specification with other 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.

decrease their ER utilization upon gaining coverage. This finding suggests that observable variation in health can explain treatment effect heterogeneity. In particular, these results suggest that health improves from always takers to compliers to never takers, and treatment effects decrease from positive to negative as health improves.

Not all observables can explain treatment effect heterogeneity. When I include age, sex, and English-speaking status as well as their two-way interactions in the MTE, substantial heterogeneity remains unexplained. I compare unexplained heterogeneity across various MTE functions in Figure 6. To do so, I present E[MTE(x, p)] functions, which average included observed heterogeneity across all individuals. Consistent with the depiction in Figure 5, 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 other observables in MTE(x, p) results in a function that is steeper than MTE(p). Therefore, the inclusion of these other observables *increases* unexplained heterogeneity in the treatment effect. My analysis demonstrates that the choice of which observables to include in the MTE function matters.

3.2 Comparable Heterogeneity within Massachusetts

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 Massachusetts MTE function is the same as the Oregon MTE function, however, I assess whether such an assumption is plausible. Therefore, I examine evidence from Massachusetts to determine if selection and treatment effect heterogeneity are similar in both states.

3.2.1 Adverse Selection on Self-Reported Health

Just as I find adverse selection on self-reported health in Oregon, I also find adverse selection on self-reported health in Massachusetts. I do so using the Behavioral Risk Factor Surveillance System (BRFSS) data that I used to study the Massachusetts reform in Kolstad and Kowalski (2012). I define the treatment D as enrollment in any health insurance in Massachusetts, and I define Z as after the reform. These definitions are in line with the literature on the Massachusetts reform. Table 2 shows that 21% of untreated Massachusetts compliers are in fair or poor health, whereas 18% of Massachusetts never takers are in fair or poor health, consistent with adverse selection on self-reported health. As in Oregon, I do not examine self-reported health of always takers because it is elicited after the reform. I do not observe previous ER utilization as an alternative proxy for health.

Using the same BRFSS data, Table 2 also provides evidence of selection heterogeneity on other observables in Massachusetts. In particular, always takers are more likely to be female and speak English than compliers and never takers. These same selection patterns do not appear as clearly in Oregon. However, my analysis of the Oregon data does not provide reason to expect that selection on sex and English-speaking status translates into differential treatment effects; there is more unexplained treatment effect heterogeneity in Oregon when I include age, sex, and English-speaking status into the MTE function. In Oregon, selection heterogeneity on these other observables is second order to selection on health, which explains treatment effect heterogeneity.

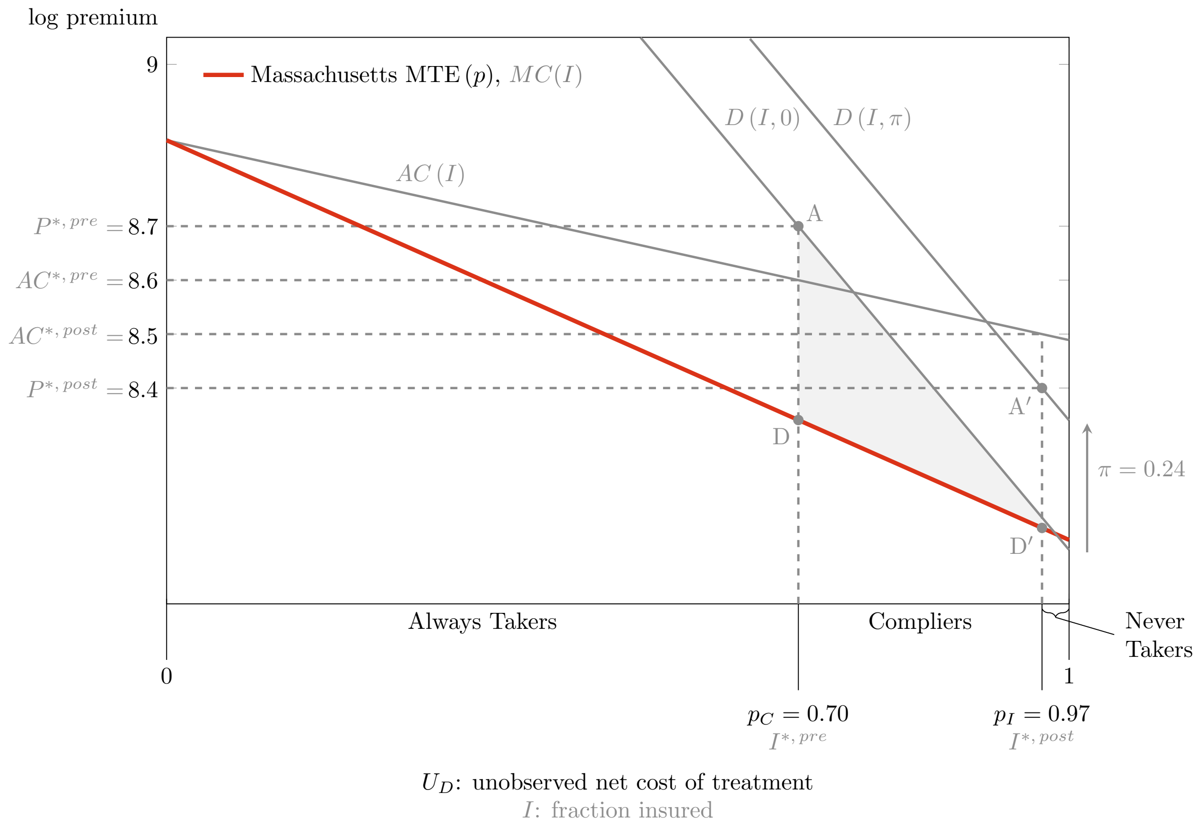
3.2.2 Treatment Effect Heterogeneity On Total Health Care Utilization

I have shown that treatment effects on ER utilization decrease from always takers to compliers to never takers in Oregon. I do not observe ER utilization in the BRFSS data, so I cannot examine treatment effect heterogeneity on ER utilization using those data. Furthermore, none of the studies that examine the impact of the Massachusetts reform on ER utilization use individual-level data,[3]

[3] 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 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.

and I cannot identify average outcomes of always takers, compliers, and never takers with the available aggregate data.

Figure 7: Figure from Hackmann et al. (2015) Recast as Massachusetts MTE(p) Shows Treatment Effect Heterogeneity on Total Utilization



However, I can examine total health care utilization as a proxy for ER utilization. Taubman et al. (2014) report evidence from the Oregon experiment that shows that ER spending and total health care spending are complements. If that is the case, then decreasing treatment effect on total health care spending implies a decreasing treatment effect on ER utilization along the unobserved net cost of treatment U_D.

To examine treatment effect heterogeneity on total health care utilization, 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. To do so, 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 7. I recast the marginal cost function estimated in Hackmann et al. (2015) as 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 the unobserved net cost of treatment U_D increases. Therefore, extrapolation from Oregon to Massachusetts has potential to be a meaningful exercise.

3.3 MTE-Reweighting Can Reconcile Oregon and Massachusetts LATEs

3.3.1 MTE-Reweighting with and without Common Observables Can Reconcile LATEs

Before I can extrapolate using the Oregon MTE function to estimate a LATE in Massachusetts, I must assess where the Massachusetts compliers should lie along the support of the Oregon unobserved net cost of treatment U_D. Recall that in Oregon, the fraction treated among the control group is $p_C = 0.15$, and the fraction treated among the intervention group is $p_I = 0.41$. These values partition the support of the unobserved net cost of treatment U_D. Therefore, an alternative interpretation of U_D is that it represents the fraction treated among the intervention or control groups. As U_D increases from 0 to 1, the fraction treated increases from 0% to 100%.

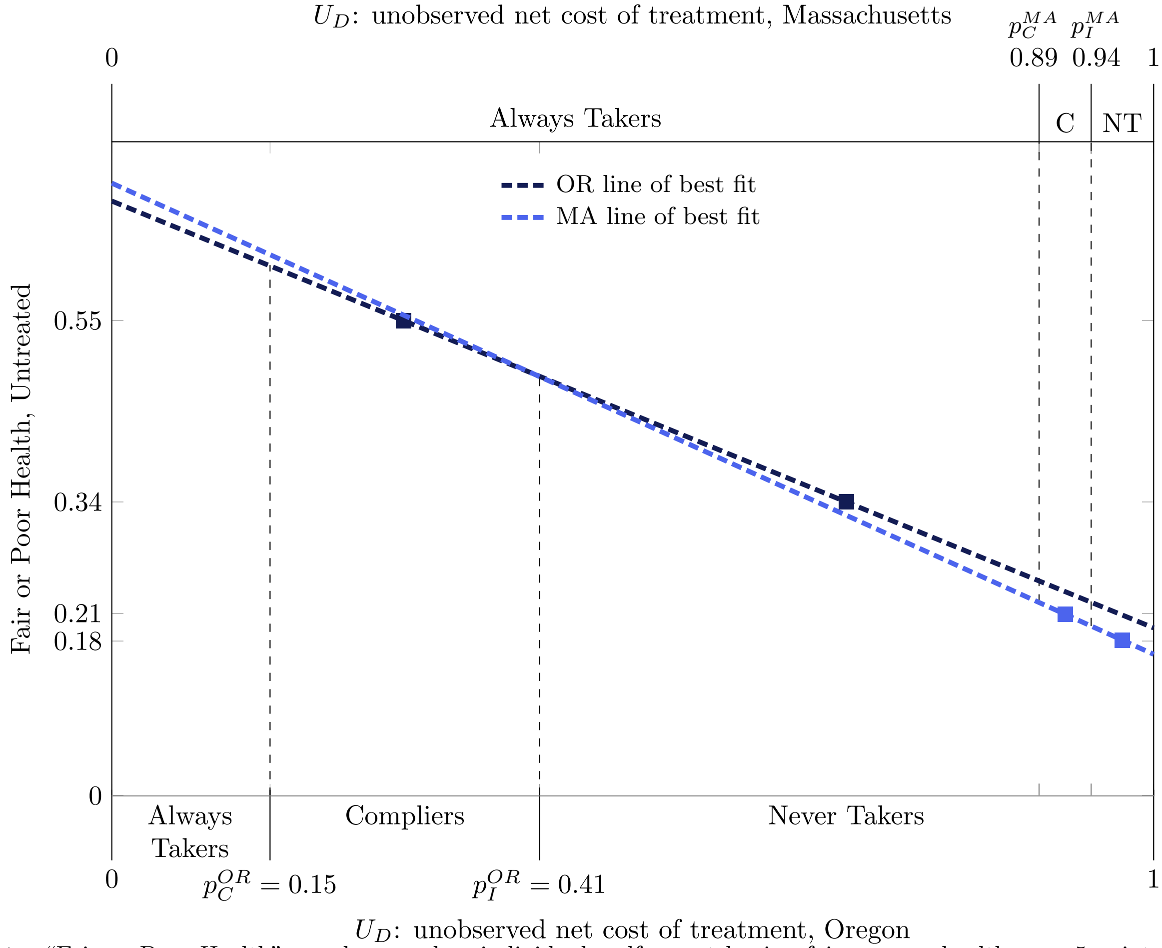
In the Massachusetts BRFSS data, the fraction of the state population with health insurance coverage increased from 89% before the reform to 94% after the reform. Unlike the Hackmann et al. (2015) data, which only capture enrollment in the individual health insurance market, the BRFSS data capture enrollment in the entire state. Because estimates from 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) reflect enrollment in the entire state, it is important that my reconciliation exercise employs data that does not restrict the sample to a segment of the market.

As a starting point for comparison, I interpret U_D as the fraction treated among the entire state population before and after the reform in Massachusetts, just as I can interpret it as the fraction treated among the intervention and control groups in Oregon. I label the probability of health insurance coverage before and after the reform in Massachusetts as $p_C^{MA} = 0.89$ and $p_I^{MA} = 0.94$ along the top axis of Figure 8. For comparison, I re-label the Oregon probabilities as $p_C^{OR} = 0.15$ and $p_I^{OR} = 0.41$ along the bottom axis. The comparison implies that Massachusetts compliers are comparable to the subset of Oregon never takers with an unobserved net cost of treatment between 0.89 and 0.94. I provide two rationales that support this comparison.

First, the Massachusetts reform included a mandate that required individuals to pay a penalty if they did not have health insurance. In contrast, the Oregon reform did not include a penalty. The Massachusetts penalty may have induced individuals who would have been never takers in the Oregon context to take up coverage and thus become compliers in the Massachusetts context.

Second, statistics on self-reported health provide quantitative evidence that Massachusetts compliers are comparable to Oregon never takers. Recall from Table 2 that untreated compliers report worse health than never takers *within* Oregon and Massachusetts. To compare these statistics *across* Oregon and Massachusetts, Figure 8 plots the fraction of untreated compliers and never takers who report fair or poor health in each state at the midpoint of the relevant range of unobserved net cost of treatment U_D. It also plots lines of best fit within Massachusetts and Oregon. These lines use the observed points within each state to predict self-reported health at other values of U_D. The Oregon prediction is quantitatively close to the observed Massachusetts points, demonstrating that the health of Massachusetts compliers is like that of Oregon never takers with U_D between 0.89

Figure 8: Self-Reported Health Similar for Massachusetts Compliers and Subset of Oregon Never Takers



Note. "Fair or Poor Health" equals one when individuals self-report having fair or poor health on a 5-point scale. "C" stands for "Compliers" and "NT" stands for "Never Takers." Filled markers represent the average value of this variable among Oregon and Massachusetts compliers and never takers, as reported in Table 2. 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.

and 0.94. These predictions also demonstrate that there is merit in linear extrapolation of Oregon heterogeneity.

Formally, I extrapolate Oregon selection and treatment effect heterogeneity by reweighting the Oregon MTE function and its component MTO and MUO functions over a general range $p_L < U_D \leq p_H$ as follows:

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

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

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

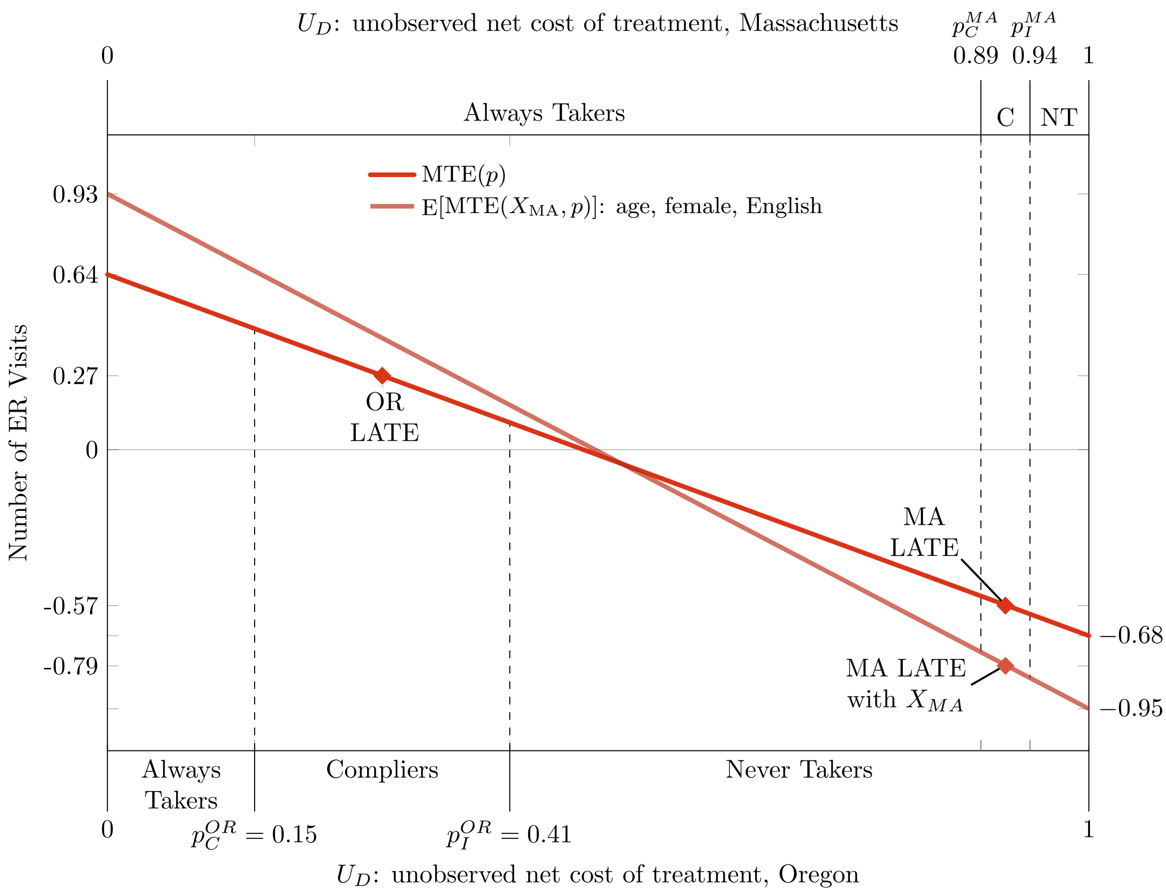
using weights $\omega(p, p_L, p_H) = 1\{p_L < p \leq p_H\}/(p_H - p_L)$. These weights are special cases of general weights for MTE-reweighting given by Heckman and Vytlacil (2007). Unlike the weights used by Brinch et al. (2017), these weights allow me to recover the exact values of the LATE and the average outcomes for Oregon always takers ($0 \leq U_D \leq p_C$), compliers ($p_C < U_D \leq p_I$), and never takers ($p_I < U_D \leq 1$) reported in Table 1. These weights also allow me to predict an untreated average outcome for Oregon always takers, a treated average outcome for Oregon never takers, and separate average treatment effects for Oregon always and never takers, as reported in the shaded cells of Table 1 and the unfilled points of Figure 4. These predictions indicate that if always takers were not enrolled in Medicaid, then they would visit the ER 1.35 times, implying a positive always taker average treatment effect of 0.54 visits. In contrast, if never takers were enrolled in Medicaid, then they would visit the ER 0.55 times, implying a negative never taker average treatment effect of 0.29 visits.

To extrapolate treatment effect heterogeneity from Oregon to Massachusetts, I begin by reweighting the Oregon MTE(p) over the implied range of the unobserved net cost of treatment U_D for Massachusetts compliers ($p_C^{MA} < U_D \leq p_I^{MA}$) via (12). I demonstrate the approach graphically in Figure 9 by reproducing MTE(p) from Oregon and labeling the support for Massachusetts compliers. As depicted, the extrapolated Massachusetts LATE predicts that the Massachusetts reform decreased ER utilization by an average of 0.57 ER visits among Massachusetts compliers. To put the magnitude of this result in context, Miller (2012) finds that Massachusetts compliers decreased their ER utilization by 0.67 to 1.28 visits per person per year, depending on the empirical strategy.[4] 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 but still negative. Therefore, reweighting of the Oregon MTE(p) function 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.

Next, I refine the extrapolation to account for differences in observables between Oregon and Massachusetts. The only observables that are available for all individuals in the Oregon and Massachusetts data are age, sex, and English-speaking status, reported in Table 2. To account for these observables in the extrapolation, I evaluate the Oregon MTE(x, p) at the Massachusetts covariate vector X_{MA}, and I reweight the resulting function over the support for Massachusetts compliers. In depict the extrapolation in Figure 9. The resulting LATE that accounts for Massachusetts observables implies an average decrease of 0.79 visits over an approximately 19-month period, which implies an annual decrease of 0.50 visits (=(0.79/19)*12). This prediction is even closer to the Miller (2012) estimates.

[4] 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 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.

Figure 9: Reweighting of MTE(p) or MTE(x,p) with Massachusetts Observables Reconciles Positive Oregon LATE and Negative Masschusetts LATE



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. "Female" is a binary indicator for the sex of the respondent. "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 observables (age, female, English) includes all two-way interactions. "C" stands for "Compliers" and "NT" stands for "Never Takers." 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.

3.3.2 LATE-Reweighting with Common Observables Cannot Reconcile LATEs

Finally, I consider whether I could have reconciled the positive LATE in Oregon with the negative LATE in Massachusetts using only the common observables available for all individuals in both contexts. I reweight the Oregon LATE in the tradition of Angrist and Fernandez-Val (2013) and Hotz et al. (2005). To examine variation in the observables available for LATE-reweighting, I use each observable to divide the sample into two subgroups, and I report LATEs within each subgroup in Table E1. The LATEs within all but one subgroup are positive. Taubman et al. (2014) also report LATEs within a wide variety of observable subgroups and find that almost all are positive. To reweight the Oregon LATE, I calculate a LATE within each joint realization of age (an indicator that age is at least the Oregon median), sex, and English-speaking status in Oregon. I then take a weighted average of these eight LATEs according to the joint frequency of the three variables among Massachusetts compliers. This LATE-reweighting protocol yields an increase of 0.23 visits among Massachusetts compliers. Therefore, LATE-reweighting using only the available observables cannot reconcile the positive treatment effect in Oregon with the negative treatment effect in Massachusetts.

It is not surprising that LATE-reweighting with age, sex, and English-speaking status cannot reconcile the results because these observables cannot explain treatment effect heterogeneity within Oregon. 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. LATE-reweighting is limited by which observables are available, but MTE-reweighting effectively allows me to extrapolate from Oregon to Massachusetts using an unobservable that captures health. This unobservable is important because treatment effects vary with health and Massachusetts compliers are healthier than Oregon compliers.

4 Conclusion

I aim to shed light on why 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 selection heterogeneity across the unobservable that separates always takers, compliers, and never takers: ER utilization decreases and health improves from always takers to compliers to never takers. I also find treatment effect heterogeneity: the effect of insurance on ER utilization decreases always takers to compliers to never takers. Differences in health explain this treatment effect heterogeneity. Although Oregon compliers increase their ER utilization upon gaining coverage, Oregon never takers, who are healthier, would decrease their ER utilization upon gaining coverage.

After demonstrating similar selection and treatment effect heterogeneity in the Massachusetts context, 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. 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. 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, suppressing X for simplicity:

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

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

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

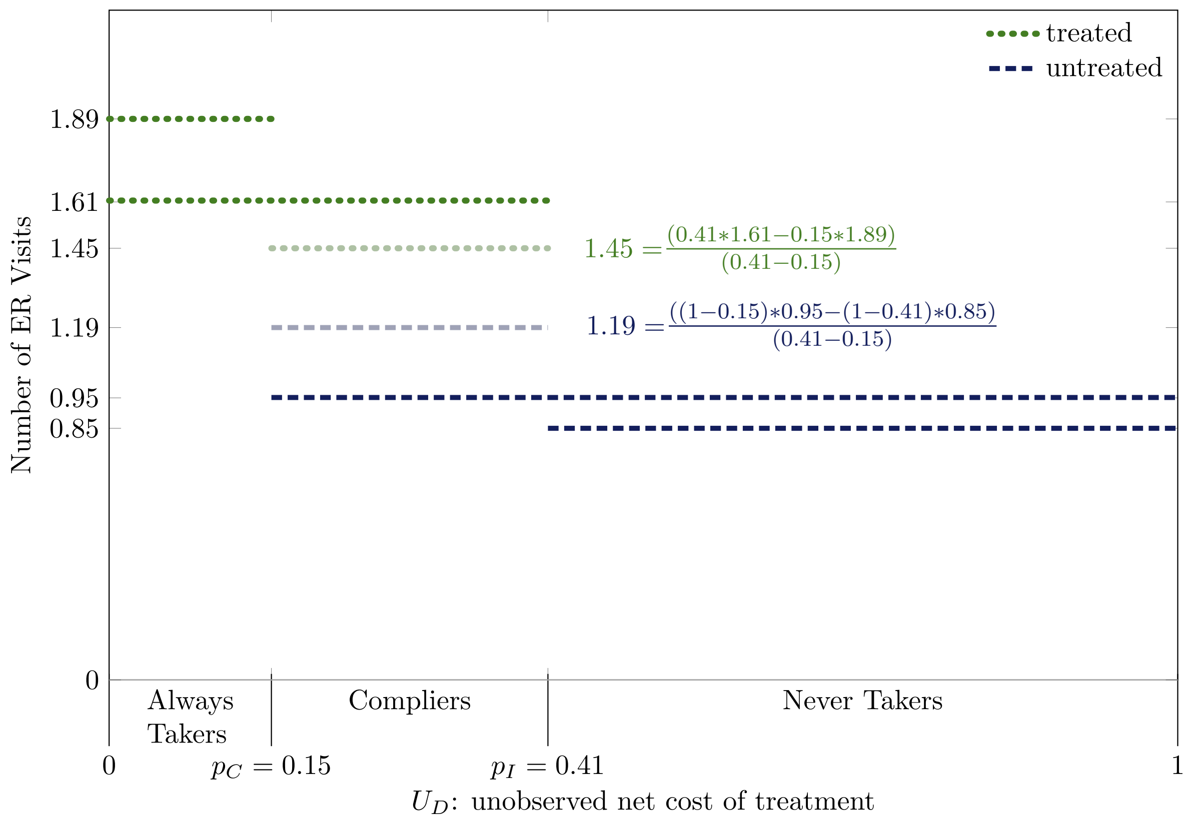
and the expected value of Y_U for compliers not enrolled in Medicaid:

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

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 ($0 \leq U_D \leq p_C$) and lottery winners enrolled in Medicaid ($0 \leq U_D \leq p_I$). 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 ($p_C < U_D \leq 1$) 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, av-

Figure C1: Derivation of Average Treated and Untreated ER Visits
for Compliers (Lighter 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.

erage *observables* of compliers should not. Therefore, I weight the average observables of compliers who win and lose the lottery by their respective probabilities:

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

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

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

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:

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

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:

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

Therefore,

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

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:

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

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

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

Therefore,

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

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 $\text{MTE}(x,p)$ using the estimated parameters of $\text{MTO}(x,p)$ and $\text{MUO}(x,p)$:

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

Appendix E Subgroup Analysis

Table E1: Oregon LATEs Are Positive in Most Subgroups

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	All	Age $\geq$ median[a]	Age $<$ median[a]	Female	Male	English	Non-English
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)
N	19,622	9,816	9,806	10,932	8,690	17,871	1,751

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. "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 sex of the respondent. "English" is an indicator for individuals in the Oregon Health Insurance Experiment who requested materials in English and 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.

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 overidentification in the LATE framework. 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.

Thomas Cornelissen, Christian Dustmann, Anna Raute, and Uta Schönberg. Who benefits from universal child care? estimating marginal returns to early child care attendance. *Journal of Political Economy*, 126(6):2356–2409, 2018.

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.

Amanda E Kowalski. Behavior within a clinical trial and implications for mammography guidelines. Working Paper 25049, National Bureau of Economic Research, September 2018. URL http://www.nber.org/papers/w25049.

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.