Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform*

Amanda E. Kowalski

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 extrapolating it to Massachusetts. Even though Oregon compliers increased their ER utilization, they were adversely selected relative to 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.

JEL Codes: C00, H00, I10. **Keywords:** compliers, LATE, marginal treatment effect, program evaluation, untreated outcome test.

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 effectively created a randomized experiment that facilitated evaluation of the impact of expanding health insurance coverage.

*aekowals@umich.edu. I thank Saumya Chatrath, Neil Christy, Simon Essig Aberg, Ryan Fraser, Aigerim Kabdiyeva, Samuel Moy, Srajal Nayak, Ljubica Ristovska, Sukanya Sravasti, and Matthew Tauzer for excellent research assistance. I extend special thanks to Magne Mogstad and Ed Vytlacil for providing multiple rounds of comments. 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, Atul Gupta, John Ham, Guido Imbens, Dean Karlan, Larry Katz, Pat Kline, Michal Kolesar, Jonathan Levin, Rebecca McKibbin, Sarah Miller, Costas Meghir, Edward Norton, Mark Rosenzweig, Joseph Shapiro, Orie Shelef, Ashley Swanson, Eva Vivalt, 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 also 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.

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 from 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. In this paper, I demonstrate that it is possible to reconcile the Oregon and Massachusetts results as two different LATEs that arise from a common MTE function.

I begin by providing background on the Oregon and Massachusetts policies and the data used to

study them. In my view, the most important difference between the policies is that they expanded coverage to individuals who likely differed in their underlying health and the impact that coverage would have on their ER utilization. The Oregon policy expanded Medicaid coverage to uninsured low-income individuals who entered a lottery. In contrast, the Massachusetts policy expanded various types of coverage across all individuals in the state, many of whom might have enrolled mainly to avoid a penalty. Accordingly, individuals who gained coverage in Oregon might have been sicker and more eager to increase their ER utilization upon gaining coverage.

To allow the impact of health insurance on ER utilization to differ across individuals who differ in their probability of gaining coverage under the Oregon and Massachusetts policies, I specify a heterogeneous treatment effects model in which the "treatment" is health insurance coverage. I begin with an MTE model shown by Vytlacil (2002) to assume no more than the LATE assumptions of Imbens and Angrist (1994). I do so to make it easier to evaluate the underlying assumptions.

Combining the model with data, I find comparable adverse selection into health insurance in Oregon and Massachusetts. I identify adverse selection by comparing outcomes and covariates that measure health in the absence of insurance coverage across "always takers" who are treated regardless of the policy, "compliers" who are treated if and only if the policy is in place, and "never takers" who are untreated regardless of the policy, in the terminology of Angrist et al. (1996). In Oregon, I find adverse selection in terms of ER utilization, previous ER utilization, and self-reported health. In Massachusetts, I find adverse selection in terms of self-reported health. The Massachusetts reform expanded coverage from a higher level than did the Oregon experiment, and over half of lottery winners in Oregon did not take up coverage. The model thus implies that I should compare Massachusetts compliers to a subset of Oregon never takers. Patterns in self-reported health support this comparison and show quantitatively similar adverse selection in both contexts.

Next, I build on my adverse selection findings to identify and compare treatment effect heterogeneity in both contexts. To do so, I make an ancillary assumption beyond the LATE assumptions that allows selection and the treatment effect to vary linearly with the fraction treated. I motivate

linearity in selection with my findings on adverse selection, and I build on those findings by allowing the treatment effect to vary as selection varies. Under this ancillary assumption, I estimate an MTE function in Oregon that shows that the treatment effect decreases from always takers to compliers to never takers, such that even though the average treatment effect on compliers is positive, the treatment effect on never takers is negative. Under an additional shape restriction, I show that previous ER utilization, a proxy for health, explains this treatment effect heterogeneity. In Massachusetts, I recast results from my previous work (Hackmann et al., 2015) in terms of the model with the ancillary assumption to show that they also imply an MTE function in which the treatment effect on health care utilization decreases from always takers to compliers to never takers.

Given comparable adverse selection and treatment effect heterogeneity in both contexts, I reconcile the results from both contexts by extrapolating the Oregon MTE function to Massachusetts. Reweighting the Oregon MTE reconciles the positive LATE in Oregon with the negative LATE in Massachusetts by comparing Massachusetts compliers to a subset of Oregon never takers, who would have decreased their ER utilization. Accordingly, the reweighting predicts a decrease in ER utilization for Massachusetts compliers. This prediction is arguably economically significant because it is of the same order of magnitude as the decrease found by Miller (2012), but it is not statistically significant, which is unsurprising given imprecision in the underlying Oregon LATE. While the prediction provides a summary measure, I would not want to rely on it alone for the reconciliation, especially without evidence that supports the strong assumptions on which it relies. Rather, I see the findings that motivate the assumptions, specifically the findings of comparable adverse selection and treatment effect heterogeneity in Oregon and Massachusetts, as important components of the reconciliation that are not reflected in a single point estimate.

I do not claim that my approach is the only way to reconcile the Oregon and Massachusetts findings. It is possible to posit a myriad of theories for why the Oregon and Massachusetts LATEs could differ, and it is difficult to reject them on the basis of two data points that represent the two LATEs. However, my approach incorporates several data points in a way that is disciplined by the model, and it can reconcile the LATEs without additional theories. Moreover, my approach

is feasible given the available data. 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 to differences in other available observables. Accordingly, I demonstrate that an alternative approach that uses available observables to reweight the LATE from Oregon cannot reconcile the results.

In the process of reconciling the Oregon and Massachusetts results, I contribute to the empirical and theoretical literature on adverse selection. Previous work on the Oregon experiment has not emphasized adverse selection. My own previous work on the Massachusetts reform (Hackmann et al., 2012, 2015) has reported adverse selection in Massachusetts using an extension of the Einav et al. (2010) cost curve test. By recasting these results in terms of the MTE model that builds on the LATE assumptions, I ground them in terms of a familiar framework from the treatment effects literature and formalize the distinction between adverse selection and treatment effect heterogeneity. This distinction is less clear in my previous work and in other work that relies on the Einav et al. (2010) cost curve test, which can find selection heterogeneity, treatment effect heterogeneity, or a combination of the two, depending on the particular cost curve.

I also contribute to the literature on treatment effects. The MTE literature generally focuses on a single context, but I use treatment effect heterogeneity within Oregon to reconcile results across the Oregon and Massachusetts contexts. As the foundation for reconciliation, I introduce the concept of selection heterogeneity, which generalizes the concept of "selection bias" from the MTE and LATE literatures (see Heckman et al., 1998; Angrist, 1998). Selection bias is not identified under the LATE assumptions, but I demonstrate that it is possible to identify a special case of selection heterogeneity under the LATE assumptions with outcomes as well as covariates. Using the concept of selection heterogeneity, I find adverse selection. I build on this finding to interpret and motivate an ancillary assumption beyond the LATE assumptions that allows me to identify treatment effect heterogeneity. This assumption has been made in the MTE literature without such motivation (Brinch et al., 2017). Throughout, I present my results using simple figures I developed for an earlier version of this paper (Kowalski, 2016). All of the content from

that version has been subsumed here, except for content on bounds that are uninformative in this context. I use that content to provide informative bounds in the context of a mammography trial (Kowalski, 2020b). In my only other closely related work (Kowalski, 2020a), I explicitly do not break any new ground, but I show with stylized examples how recent advances from the treatment effects literature, inclusive of contributions made here, can inform external validity. I co-developed the Stata command mtebinary (Kowalski et al., 2018) to make the calculations performed here accessible for reconciliation of results from other experiments.

2 Background and Data

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). However, I provide some background on the Oregon experiment and the Massachusetts reform that motivates the approach that I take to reconcile the results.

The Oregon Health Insurance Experiment expanded Medicaid coverage to uninsured individuals who entered a lottery that took place from March 10, 2008 to September 30, 2009. Entrants were only required to provide eligibility documentation if they won. Therefore, many individuals who were already eligible for Medicaid entered the lottery, perhaps unaware of their eligibility. Many such individuals enrolled in Medicaid even if they lost the lottery, perhaps because an emergency room facilitated submission of their eligibility documentation after they received uncompensated care. On the other end of the spectrum, many lottery winners did *not* enroll in Medicaid, either because they did not submit eligibility documentation, or because they did not meet the eligibility requirements, which included income below the federal poverty level, or both.

The Massachusetts reform of 2006 expanded health insurance coverage to uninsured individuals statewide through a variety of mechanisms: an expansion of Medicaid eligibility and other subsidized coverage to individuals with incomes below three times the federal poverty level, reforms on the individual health insurance market, a mandate that employers had to offer coverage or pay a penalty, and a mandate that individuals had to enroll in coverage or pay a penalty. Mas-

sachusetts had a very high rate of coverage before the reform, and coverage expanded to a higher level after the reform. However, coverage was still not universal.

I analyze the Oregon experiment using publicly available individual-level data. 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 who 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 and statistically insignificant, but still positive, LATE of 0.27.

I analyze the Massachusetts reform using individual-level data that I previously used to analyze the Massachusetts reform in Kolstad and Kowalski (2012) and Hackmann et al. (2015). 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. Therefore, it is important that I do not attempt to reconcile estimates using Massachusetts data that are restricted to a subset of the population, such as individuals eligible for Medicaid.

3 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 any ancillary assumptions before doing so explicitly, I follow the exposition from Heckman and Vytlacil (2005) closely. I extend that exposition by illustrating the model with simple figures and by presenting it in terms of variables relevant to the Oregon and Massachusetts contexts.

3.1 First Stage: Enrollment

I specify a binary treatment D that represents enrollment in Medicaid in Oregon and enrollment in any health insurance coverage in Massachusetts. I have chosen these definitions of D because my goal is to reconcile estimates that have been discussed in the literature and the media: namely, estimates that show that enrollment in any health insurance reduces ER utilization in Massachusetts and that enrollment in Medicaid increases ER utilization in Oregon. Let V_T represent potential utility in the treated state, and let V_U represent potential utility in the untreated state. 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. Vytlacil (2002) shows that the additive separability in (2) is equivalent to the LATE monotonicity assumption of Imbens and Angrist (1994). In the Oregon context, Z indicates winning the lottery. In the Massachusetts context, Z indicates that the reform has occurred, which I also interpret as "winning the lottery" for parallelism. 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." I need different terminology for the intervention group $(Z = 1)$ and the treated group $(D = 1)$ because, in both contexts, not all lottery winners are treated. 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. (First Stage Independence) The random variable ν_D is independent of Z conditional on X.

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 completeness in Online

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:

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

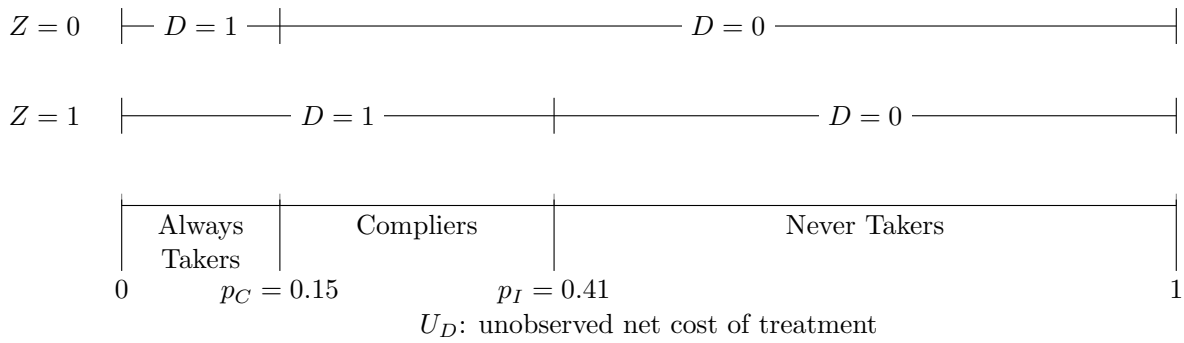
I show the derivation in Online Appendix B for completeness. Under A.3, the threshold is different for lottery winners and losers with the same vector of covariates X, which yields the following two cases:

$$D = \begin{cases} 1\{U_D \leq p_{CX}\}, & \text{if } Z = 0 \\ 1\{U_D \leq p_{IX}\}, & \text{if } Z = 1 \end{cases} \tag{4}$$

where $p_{CX} = P\big(D = 1 \mid Z = 0, X\big)$ is the probability of treatment in the control group conditional on X, and $p_{IX} = P\big(D = 1 \mid Z = 1, X\big)$ is the probability of treatment in the intervention group conditional on X.

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. I show these ranges in Figure 1 using data from the Oregon experiment. Within my analysis sample, 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 unobserved net cost of treatment U_D in $[0, 0.15]$. Treated lottery losers are always takers. In the middle line of Figure 1, I depict the lottery winners. By (4), the untreated lottery winners have unobserved net cost of treatment U_D in $(0.41, 1]$. 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 unobserved net cost of treatment U_D in the middle range, $(0.15, 0.41]$, 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



Note. 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.

The ordering from always takers to compliers to never takers along U_D is an ordering across an important margin: the margin of treatment takeup. In Oregon and Massachusetts, as 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. The model does not require me to specify what is included in U_D. Instead, it gives me a framework that I can use to examine which factors separate always takers from compliers and never takers.

3.2 Second Stage: ER Utilization

I relate treatment D to realized ER utilization Y as follows:

$$Y = Y_U + (Y_T - Y_U)D, \tag{5}$$

where I specify potential ER utilization if treated Y_T and if untreated Y_U as follows:

$$Y_T = g_T(X, U_D, \gamma_T) \tag{6}$$

$$Y_U = g_U(X, U_D, \gamma_U), \tag{7}$$

where $g_U(\cdot)$ and $g_T(\cdot)$ are unspecified functions, X is the 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 ensure that average treated and untreated potential outcomes are defined for each X, I assume:

A.4. (First and Second Stage Independence) The random vectors (ν_D, γ_T) and (ν_D, γ_U) are independent of Z conditional on X.

A.5. (Treated and Untreated) $0 < \mathrm{P}(D = 1 \mid X) < 1$.

A.6. (Second Stage Technical Assumption) The values of $E[Y_T]$ and $E[Y_U]$ are finite.

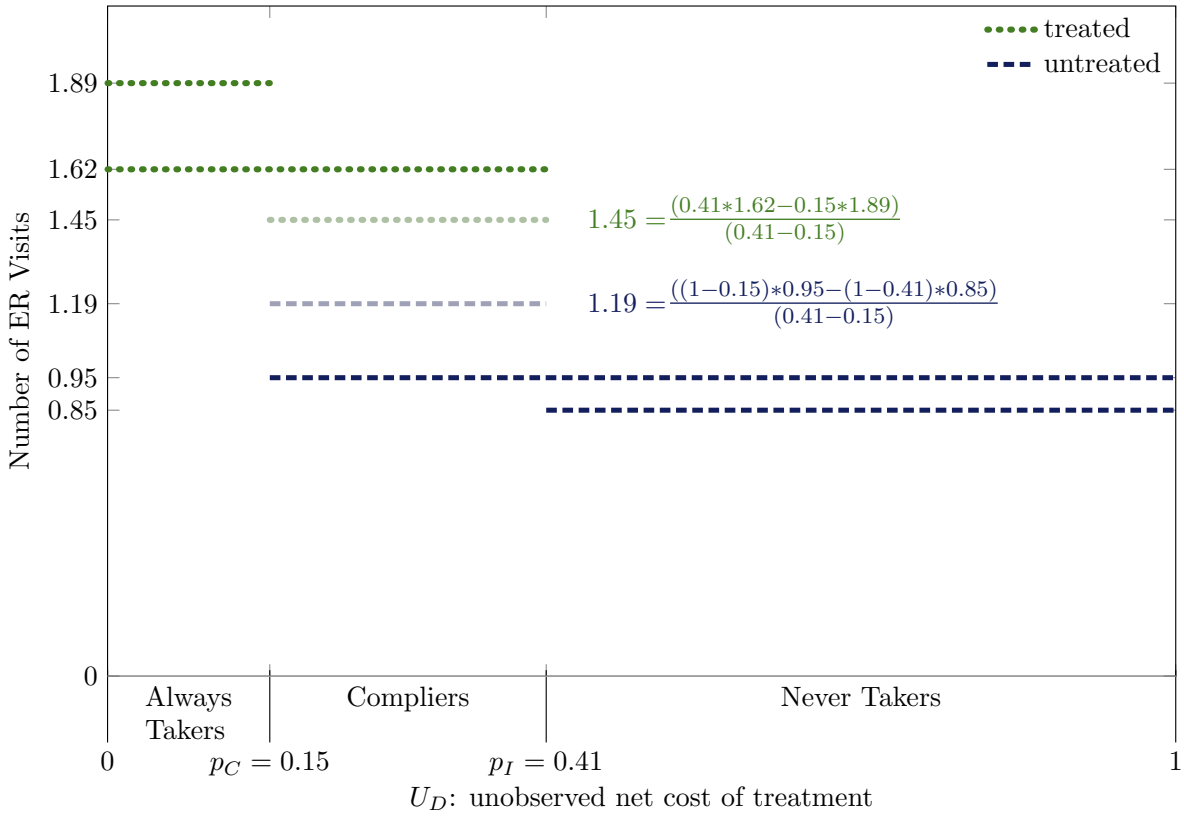
While A.4 implies A.2, I introduce them both to separate the first and second stage assumptions. As a whole, because I have only made stylistic changes to the model presented by Heckman and Vytlacil (2005), by the proof of Vytlacil (2002), the model, given by the utility equations (1)–(2), treatment equations (3)–(4), potential outcome equations (5)–(7), and assumptions A.1–A.6, assumes no more than the LATE assumptions. In particular, assumptions A.1–A.6 imply the LATE monotonicity assumption that no individuals would take up treatment if and only if they lose the lottery, which is defensible in this setting as winning the lottery should only increase coverage.

I illustrate implications of the model graphically in Figure 2 using data from Oregon. Along the horizontal axis, I depict the unobserved net cost of treatment U_D as I do in Figure 1. Along the vertical axis, I depict the number of ER visits after the lottery took place as the outcome Y. Using this figure, I illustrate how the model makes it possible to derive average outcomes for always takers, compliers, and never takers, consistent with the algebraic derivations of Imbens and Rubin (1997), Katz et al. (2001), Abadie (2002), and Abadie (2003) under the LATE assumptions. I provide an algebraic derivation under the assumptions of the model in Online Appendix C.

To derive average outcomes for always takers, compliers, and never takers, I begin by considering average outcomes in groups formed by the interaction of the treatment and the instrument. Treated individuals in control must be always takers, so the average treated outcome of always takers is $E[Y_T|0 \leq U_D < p_C] = E[Y|D = 1, Z = 0]$. In Figure 2, I plot this value of average ER utilization from the Oregon experiment, 1.89 visits, over the support of the unobserved net cost of treatment U_D for always takers, using a dotted line to indicate that it represents an average *treated* outcome. Untreated individuals in intervention must be never takers, so the average untreated outcome of never takers is $E[Y_U|p_I \leq U_D \leq 1] = E[Y|D = 0, Z = 1]$. I plot this value, 0.85 visits, over the support for never takers, using a dashed line to indicate that it represents an average *untreated* outcome. Treated individuals in intervention are always takers and compliers, so the average treated outcome among always takers and compliers is $E[Y_T|0 \leq U_D < p_I] = E[Y|D = 1, Z = 1]$. I plot this value, 1.62 visits, over the relevant support. The average treated outcome of compliers is a weighted average of the average treated outcome among always takers and compliers and the average treated outcome of always takers. Because I can determine the weights and the required average outcomes from the figure, I can back out the average treated outcome of compliers as $E[Y_T|p_C \leq U_D < p_I] = \frac{p_I E[Y_T|D=1,Z=1] - p_C E[Y_T|0 \leq U_D < p_C]}{p_I - p_C}$. I plot this value, 1.45 visits, using a lighter dotted line over the support for compliers. The derivation of average untreated outcomes of compliers is similar such that $E[Y_U|p_C \leq U_D < p_I] = \frac{(1-p_C)E[Y_U|D=0,Z=0] - (1-p_I)E[Y_U|p_I \leq U_D \leq 1]}{p_I - p_C}$. I plot this value, 1.19 visits, using a lighter dashed line over the support for compliers.

Figure 3 removes the intermediate steps of the derivation from Figure 2 to isolate the average

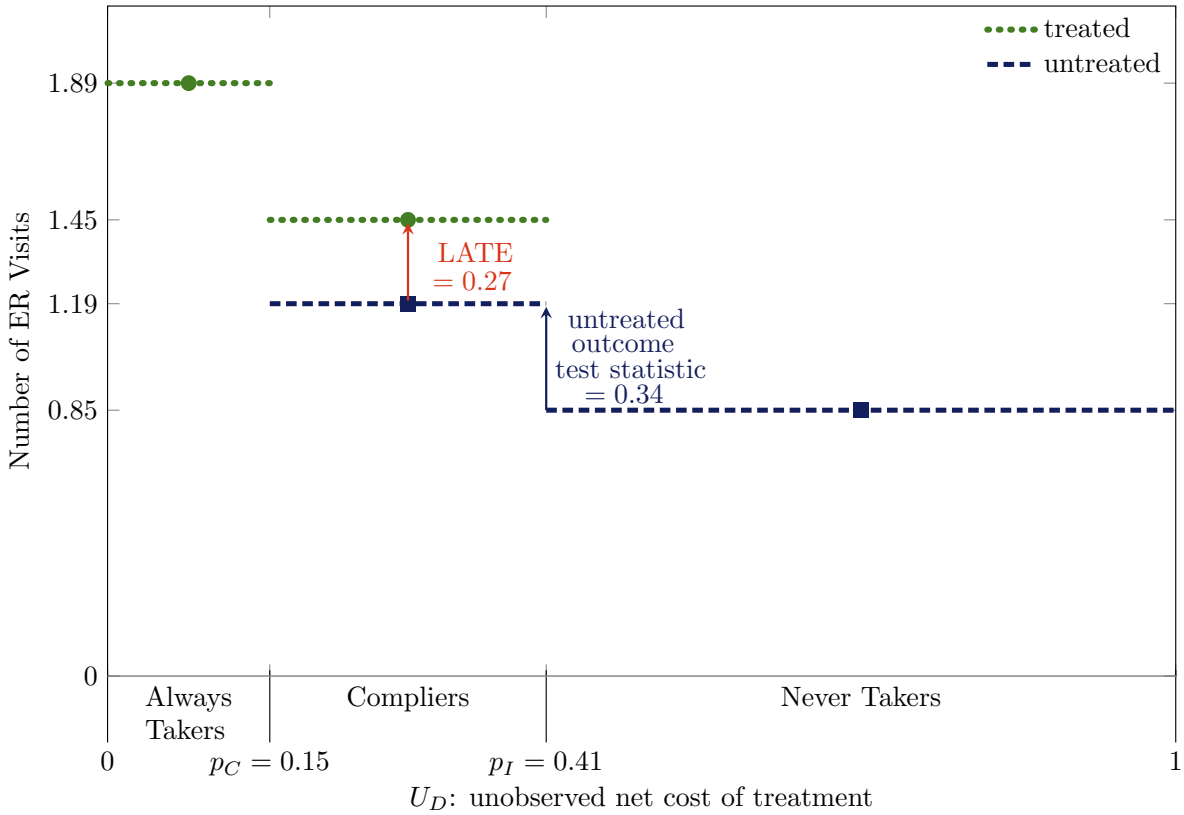
Figure 2: Model Implies a Derivation of Average ER Utilization for Always Takers, Compliers, and Never Takers (Average ER Utilization for Compliers in Lighter Shading)



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.

Figure 3: 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. Treatment represents enrollment in Medicaid. I discuss the untreated outcome test statistic in Section 4.1.1. Some differences between statistics might not appear internally consistent because of rounding.

outcomes of always takers, compliers, and never takers. 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. Figure 3 provides more information than the LATE alone. As shown by Angrist (1990) and Angrist and Krueger (1992), it is possible to calculate the LATE even if it is not possible to calculate the average outcomes of always takers, compliers, and never takers depicted in Figure 3.[1] This additional information can call into question whether the LATE applies to always and never takers, who represent substantial and distinct groups. 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. These differences in the average outcomes across always takers, compliers, and never takers could reflect selection or treatment effect heterogeneity.

3.3 Definitions of Selection and Treatment Effect Heterogeneity

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

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

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

I refer to the first function as the "marginal untreated outcome (MUO)" function. 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 MUO function traces out selection heterogeneity because, by definition, untreated outcomes do not include a treat-

[1] Calculation of the average outcomes depicted in Figure 3 requires the ability to calculate average outcomes formed by the interaction of the treatment and the instrument. In contrast, calculation of the LATE only requires the ability to calculate average outcomes formed by the treatment and the instrument separately. Using the Wald (1940) approach, the reduced form $\text{E}[Y|Z=1] - \text{E}[Y|Z=0]$ is equal to 0.07, and the first stage $\text{E}[D|Z=1] - \text{E}[D|Z=0]$ 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 3.

ment effect. The MTE literature uses the MUO function as an intermediate input in the derivation of the second function, the "marginal treatment effect (MTE)" function of Heckman and Vytlacil (1999, 2001, 2005). However, it does not use the MUO function to define selection heterogeneity (see Carneiro and Lee, 2009; Brinch et al., 2017). Instead, the MTE and LATE literatures focus on the following definition of "selection bias" (see Heckman et al., 1998; Angrist, 1998):

$$\text{Selection Bias: } \mathrm{E}[Y_U \mid D=1] - \mathrm{E}[Y_U \mid D=0]. \tag{8}$$

I demonstrate that selection heterogeneity generalizes selection bias by expressing (8) as the following weighted integral of the MUO function, where I omit dependence on X for simplicity:

$$\int_0^1 \Big[\frac{p_C}{p_C + \mathrm{P}(Z=1)(p_I - p_C)} 1\{0 \le p < p_C\} + \frac{\mathrm{P}(Z=1)(p_I - p_C)}{p_C + \mathrm{P}(Z=1)(p_I - p_C)} 1\{p_C \le p < p_I\}$$
$$- \frac{\mathrm{P}(Z=0)(p_I - p_C)}{\mathrm{P}(Z=0)(p_I - p_C) + 1 - p_I} 1\{p_C \le p < p_I\} - \frac{1 - p_I}{\mathrm{P}(Z=0)(p_I - p_C) + 1 - p_I} 1\{p_I \le p < 1\} \Big] \mathrm{MUO}(p)\, \mathrm{d}p.$$

The first and second terms in brackets yield the average untreated outcome conditional on receiving treatment. The first term places weight on always takers and the second term places weight on treated compliers. The third and fourth terms in brackets yield the average untreated outcome conditional on not receiving treatment. The third term places weight on untreated compliers and the fourth term places weight on never takers. This weighted integral 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 an ancillary assumption in a model with a binary instrument 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 an ancillary assumption.

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, doing so would have material implications because it would reverse the sign of the treatment effect.[2]

4 Findings

I have three main findings. First, I find comparable adverse selection within Oregon and Massachusetts along the unobservable that separates always takers, compliers, and never takers. Massachusetts compliers have similar self-reported health to a subset of Oregon never takers, and are thus healthier than Oregon compliers given the adverse selection I find. Second, I use this evidence of adverse selection to motivate an ancillary assumption to estimate treatment effect heterogeneity. I find that treatment effect heterogeneity in Oregon is decreasing such that even though compliers increase their ER utilization upon gaining coverage, never takers would decrease their ER utilization upon gaining coverage. I present comparable evidence of decreasing treatment effect heterogeneity in Massachusetts. Third, using the Oregon experiment as the "gold standard," I extrapolate treatment effect heterogeneity from within the Oregon experiment to Massachusetts, and I reconcile the positive LATE in Oregon with the negative LATE in Massachusetts. I show that the observables available in both settings cannot reconcile the results via LATE-reweighting.

4.1 Comparable Adverse Selection in Oregon and Massachusetts

4.1.1 Adverse Selection in Oregon

I identify a special case of selection heterogeneity using a test that I refer to as the "untreated outcome test." This test rejects selection heterogeneity if the test statistic, the average untreated outcome of compliers minus the average untreated outcome of never takers, is different from zero. I derive the average untreated outcomes required for the test statistic in the discussion of Figure 2. The untreated outcome test is similar or equivalent to tests proposed by Guo et al. (2014), Black

[2]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 is material to the definition of the treatment effect. Therefore, the distinction between treated and untreated is also material to the definition of treatment effect heterogeneity. Reversing the definition of the treatment would imply that treatment effect heterogeneity is heterogeneity in the untreated outcome minus the treated outcome, so there would still be a meaningful distinction between the MUO and MTO functions. Heterogeneity in untreated outcomes would represent the sum of selection and treatment effect heterogeneity, and heterogeneity in treated outcomes would represent selection heterogeneity.

et al. (2017), and Bertanha and Imbens (2020), which are generalized by Mogstad et al. (2018).[3] I emphasize that the untreated outcome test is also equivalent to the Einav et al. (2010) cost curve test for selection heterogeneity if applied to *uninsured ER* utilization. Whereas the Einav et al. (2010) cost curve test interprets both the insured and uninsured cost curves in terms of selection heterogeneity, I emphasize that the interpretation of the curves depends on whether they represent cost curves among the insured or uninsured. Heterogeneity in costs among the insured reflects selection plus treatment effect heterogeneity, but heterogeneity in costs among the uninsured reflects selection heterogeneity only.

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. 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 ($p_I < U_D \leq 1$). Randomization drives identification by creating a distinction between never takers and untreated compliers.

Applying the untreated outcome test to the Oregon experiment, I reject the null hypothesis of selection homogeneity. As shown in Figure 3 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. I

[3] I refer to the Bertanha and Imbens (2020) 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	
	(0.07)	(0.11)	(0.42)	
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)	(0.42)	

Note. Standard errors from a nonparametric bootstrap with 1,000 replications are in parentheses. The shaded cells report extrapolated values from MTE-reweighting via (9)–(11) 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. I discuss the shaded cells in Section 4.3.1. Some differences between statistics might not appear internally consistent because of rounding.

calculate the standard error of the test statistic as the standard deviation of the estimates from 1,000 nonparametric bootstrap 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.

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

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

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 vectors of always takers, compliers, and never takers. The calculation of the average covariate vectors for always and never takers is analogous to the calculation of their average outcomes. However, for compliers, even though their average *outcomes* should depend on whether they win or lose the lottery, their average *covariates* should not. Therefore, I weight the average covariates of compliers who win and lose the lottery by their respective probabilities:

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

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 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.

As shown in Table 2, within Oregon, I find that never takers are less likely to be in fair or

Table 2: Adverse Selection on Self-Reported Health and Previous ER Utilization Within Oregon and Massachusetts

	Means				Difference in Means	
	All	(1) Always Takers	(2) Compliers	(3) 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 Visits	0.17	0.26	0.19	0.15	0.07	0.04
	(0.003)	(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)	(0.003)	(0.01)	(0.01)
Other Observables						
Age	40.69	39.45	42.41	40.25	-2.96	2.16
	(0.09)	(0.27)	(0.42)	(0.20)	(0.51)	(0.58)
Female	0.56	0.72	0.53	0.53	0.19	0.003
	(0.004)	(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
	(0.01)		(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)	(1.48)	(0.49)	(1.55)	(1.64)
Female	0.51	0.52	0.43	0.38	0.10	0.04
	(0.003)	(0.005)	(0.05)	(0.02)	(0.06)	(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)	(0.03)
N	62,456	55,966	3,175	3,314		

Note. Standard errors from a nonparametric bootstrap with 1,000 replications 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. For the Oregon experiment, 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 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" is an indicator that equals one for the two worst values of self-reported 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.

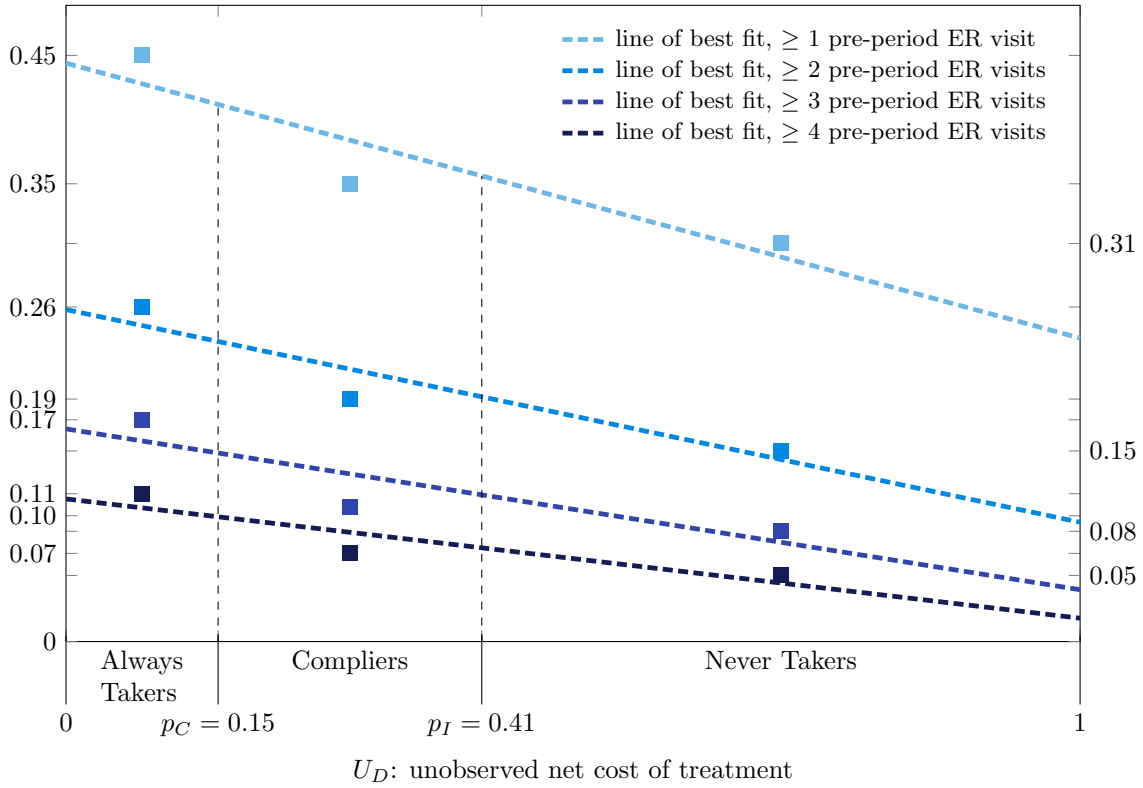
poor health than untreated compliers, and the difference is statistically significant. Therefore, I find adverse selection on self-reported health that is consistent with the adverse selection on ER utilization that I find using the untreated outcome test. This adverse selection indicates that never takers are healthier than compliers, which may be one reason why 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 because it was elicited after randomization.

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 4 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 than compliers. Similar relationships hold between compliers and never takers. Furthermore, the relationships appear 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 necessarily monotonic from always takers to compliers to never takers. The non-monotonicity in selection on age indicates that age may not explain adverse selection on ER utilization. Therefore, to the extent that health explains adverse selection on ER utilization, age is not a good proxy for health in this setting. In contrast, self-reported health and previous ER utilization, both of which are proxies for health, may explain adverse selection on ER utilization.

Figure 4: 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.

4.1.2 Adverse Selection in Massachusetts

As in Oregon, I find adverse selection on self-reported health in Massachusetts. Table 2 shows that in Massachusetts, 21% of untreated compliers are in fair or poor health, compared to 18% of never takers. As in Oregon, I do not examine self-reported health of always takers because I only observe them when treated, so their self-reported health could include a treatment effect. Unlike in Oregon, I do not examine previous ER utilization as an alternative proxy for health because I define the Massachusetts intervention as after reform and the Massachusetts control as before the reform, so it is unclear what "previous" ER utilization should represent. In Table 2, I provide evidence of selection heterogeneity on other observables available in Oregon and Massachusetts. It shows that always takers are more likely to speak English than compliers.

Not only is there evidence of adverse selection in terms of self-reported health in both contexts, but also these selection patterns are quantitatively similar across contexts. Before I can compare selection patterns in Oregon to selection patterns 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%.

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. Whereas the literature takes difference-in-differences approaches to account for trends in coverage, I only consider a single difference (before versus after the reform) for simplicity because I show in Kolstad and Kowalski (2012) that an additional difference that accounts for national trends has very little impact on the coverage results. Before the Massachusetts reform, the coverage rate in Massachusetts was among the highest in the nation. According to the Behavioral Risk Factor Surveillance System data used

to examine coverage in Kolstad and Kowalski (2012), the fraction treated increased from 89% before the reform to 94% after the reform. I label the probability of health insurance coverage before and after the reform in Massachusetts as $p_C^{MA} = 0.90$ and $p_I^{MA} = 0.95$ along the top axis of Figure 5. 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 0.90 and 0.95.

Institutional details support the comparison of Massachusetts compliers to a subset of Oregon never takers. The Massachusetts reform included a mandate that required uninsured individuals to pay a penalty. In contrast, the Oregon reform did not include a penalty for uninsurance. 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.

Patterns in adverse selection on self-reported health provide quantitative evidence that Massachusetts compliers are comparable to a subset of the Oregon never takers. Figure 5 plots the fraction of untreated compliers and never takers who report fair or poor health at the midpoints of the relevant supports of the unobserved net cost of treatment U_D in Massachusetts and Oregon. 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 values observed in Massachusetts, demonstrating that the health of Massachusetts compliers is like that of Oregon never takers with U_D between 0.90 and 0.95. These predictions also demonstrate that there is merit in linear extrapolation of Oregon heterogeneity.

4.2 Comparable Treatment Effect Heterogeneity in Oregon and Massachusetts

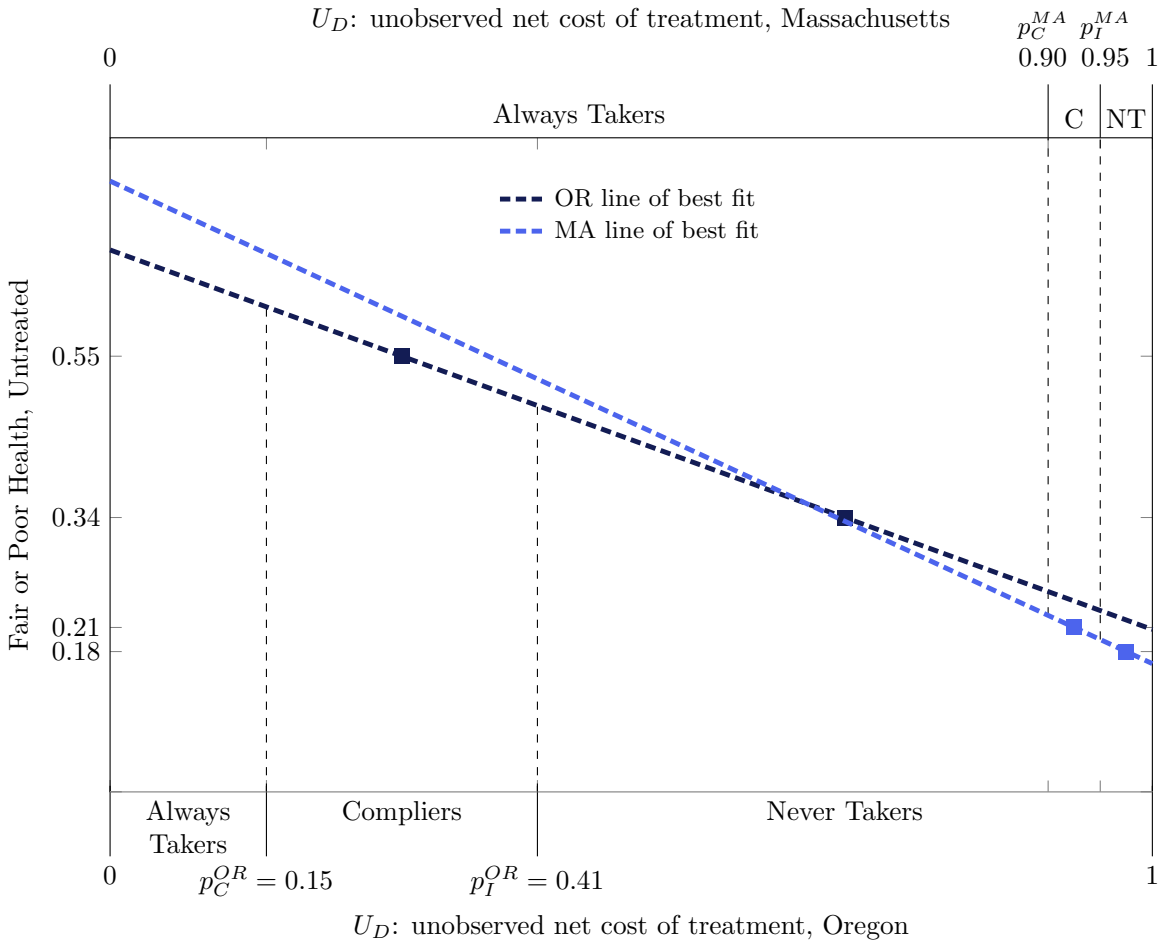
4.2.1 Treatment Effect Heterogeneity in Oregon

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

AA.1. (Linear Selection Heterogeneity and Linear Treatment Effect Heterogeneity) In (6) and (7), 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,

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

Figure 5: Self-Reported Health Similar for Massachusetts Compliers and Subset of Oregon Never Takers



Note. "Fair or Poor Health" is an indicator that equals one for the two worst values of self-reported health on a 5-point scale. "C" stands for "Compliers" and "NT" stands for "Never Takers." Filled markers indicate averages among Oregon and Massachusetts untreated compliers and never takers with nonmissing self-reported health, 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. The treatment and instrument in Massachusetts and Oregon are defined as in Table 2. p_C^{OR} is the probability of treatment in the control group, and p_I^{OR} is the probability of treatment in the intervention group in my full analysis sample for Oregon. p_C^{MA} the probability of treatment in the control group, and p_I^{MA} the probability of treatment in the intervention group in my full analysis sample for Massachusetts.

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

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

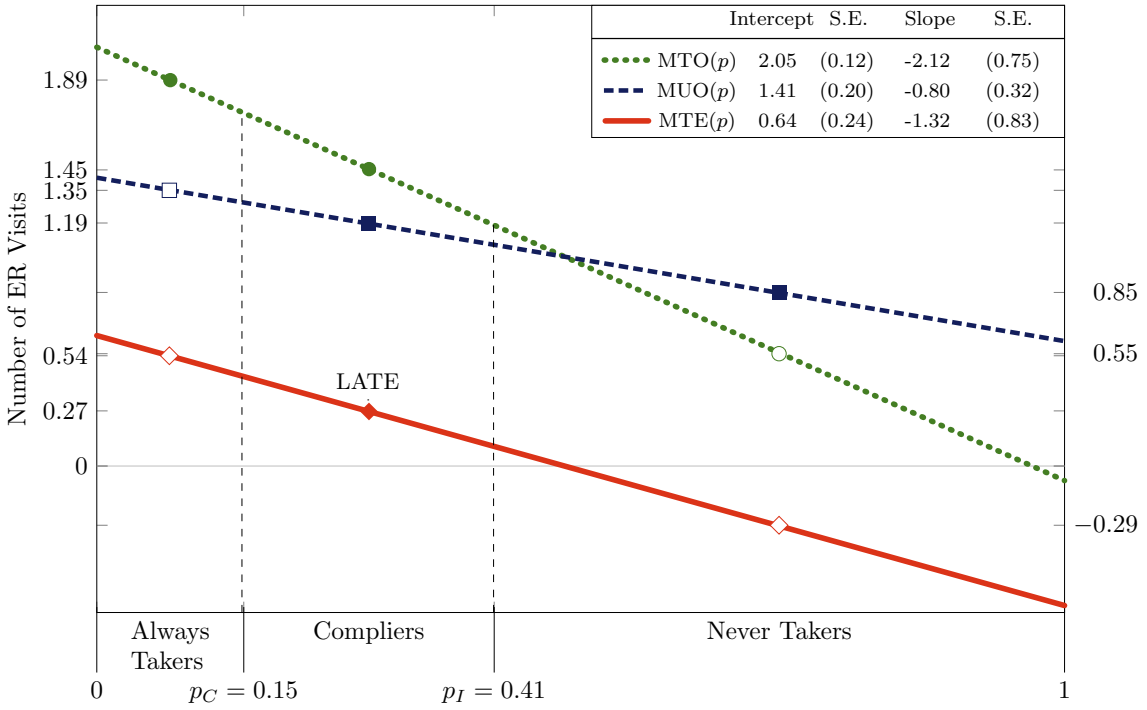
Brinch et al. (2017) impose an equivalent assumption to examine the impact of family size on child outcomes; Olsen (1980) imposes linearity of the MTO function to examine the impact of family size on maternal outcomes; and several other papers impose linearity of the MTE function in other applications (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. I impose AA.1 instead of the weak monotonicity assumption from Brinch et al. (2017) that I make in the context of a clinical trial on mammography in Kowalski (2020b) because the resulting bound on the treatment effect for always takers is uninformative about treatment effect heterogeneity here.

In the Oregon context, the empirical evidence of adverse selection provides motivation for AA.1. The evidence of approximately linear adverse selection with respect to previous ER utilization motivates the assumption that selection heterogeneity is linear. Furthermore, the evidence of adverse selection with respect to self-reported health indicates that the unobserved net cost of treatment U_D captures health. Therefore, the assumption that treatment effect heterogeneity is linear allows the treatment effect to vary with health as selection varies with health. I note 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 6 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,

Figure 6: Treatment Effect on ER Utilization

Decreases from Always Takers to Compliers to Never Takers



Note. Standard errors from a nonparametric bootstrap with 1,000 replications 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.

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.

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 (6) and (7), for $k \in \{T,U\}$, specify $g_k(X,U_D,\gamma_k) = \delta_k' X + \lambda_k U_D + \gamma_k$, where $\mathrm{E}[\gamma_k \mid U_D = p, X = x] = 0$. Therefore,

$$\mathrm{MUO}(p,x) = \mathrm{E}[Y_U \mid U_D = p, X = x] = \delta_U' x + \lambda_U p$$

$$\mathrm{MTE}(p,x) = \mathrm{E}[Y_T - Y_U \mid U_D = p, X = x] = (\delta_T - \delta_U)' x + (\lambda_T - \lambda_U) p$$

$$\mathrm{MTO}(p,x) = \mathrm{E}[Y_T \mid U_D = p, X = x] = \delta_T' x + \lambda_T p.$$

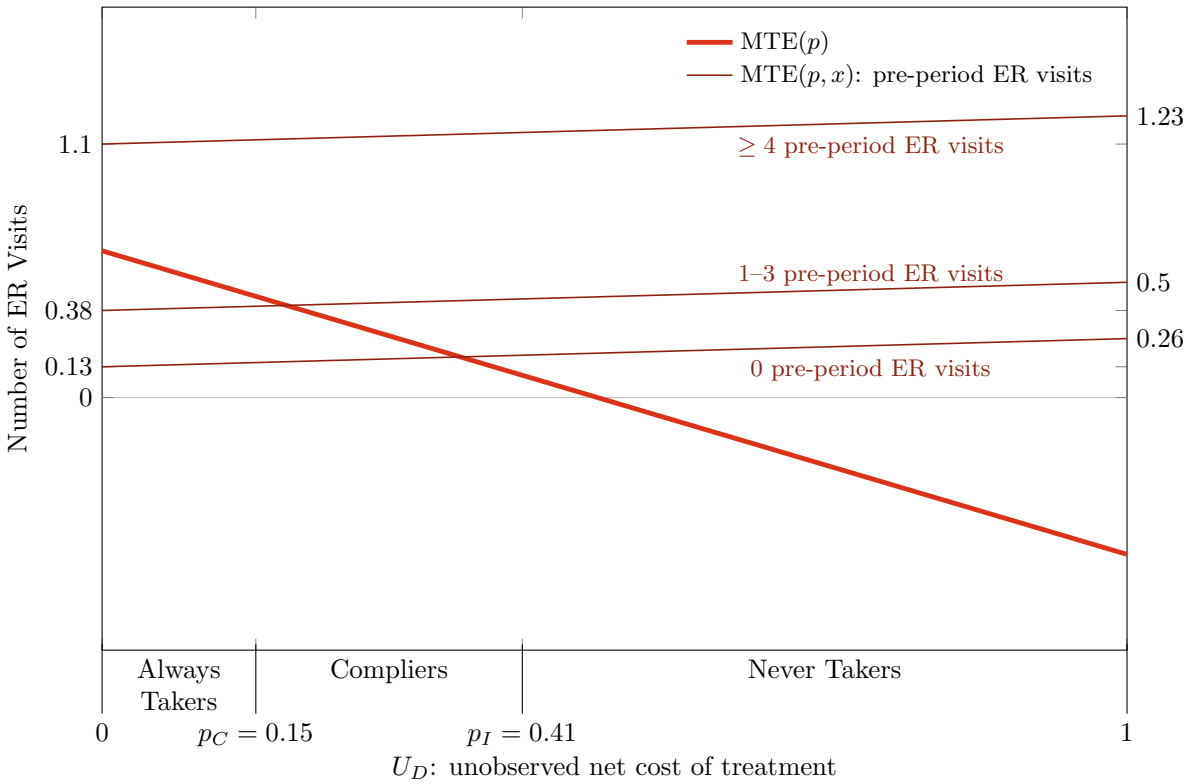
I present an algorithm for estimation of these functions that simplifies the Heckman et al. (2006) algorithm in Online Appendix D.

To evaluate whether health explains treatment effect heterogeneity, I incorporate proxies for health into the MTE function. Since self-reported health was elicited after randomization, it represents an outcome as opposed to an observable for treated individuals, and 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 function via AA.2, I find that previous ER utilization can explain the entire decrease in the 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 function, I obtain a separate MTE(p,x) for each visit 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(p,x) 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. The slope of the MTE function conditional on a particular covariate vector informs the importance of those covariates in explaining treatment effect heterogeneity because the unobserved net cost of treatment U_D is defined conditional on the included covariate vector. 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 7, the MTE(p,x) for individuals with 4 or more pre-period visits is much greater than the MTE(p,x) for individuals with zero pre-period visits. The institutional details of the Oregon experiment provide a plausible mechanism for why individuals with the highest number of pre-period ER visits have the largest treatment effects. Given that the ER can help to facilitate Medicaid enrollment, it could be possible that some individuals enrolled in Medicaid precisely because they showed up at the ER to receive care. Therefore, it is possible that individuals who were the most likely to enroll in Medicaid were also the most eager to increase their ER utilization, leading them to have the largest treatment effect of enrollment on ER utilization. The evidence shown in Figure 7 combined with institutional details from the Oregon experiment suggest that observable variation in previous ER utilization, a proxy for 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 as health improves.

Figure 7: 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, 1–3 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.

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 8. To do so, I present the expectation of the MTE function with respect to covariate x $\mathrm{E}[\mathrm{MTE}(p,x)]$, which averages included observed heterogeneity across all individuals. Consistent with the depiction in Figure 7, the inclusion of pre-period ER visits in $\mathrm{MTE}(p,x)$ results in a function that is flatter than $\mathrm{MTE}(p)$. Therefore, the inclusion of pre-period ER visits decreases unexplained heterogeneity in the treatment effect. In contrast, the inclusion of the other observables in $\mathrm{MTE}(p,x)$ results in a function that is steeper than $\mathrm{MTE}(p)$. 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.

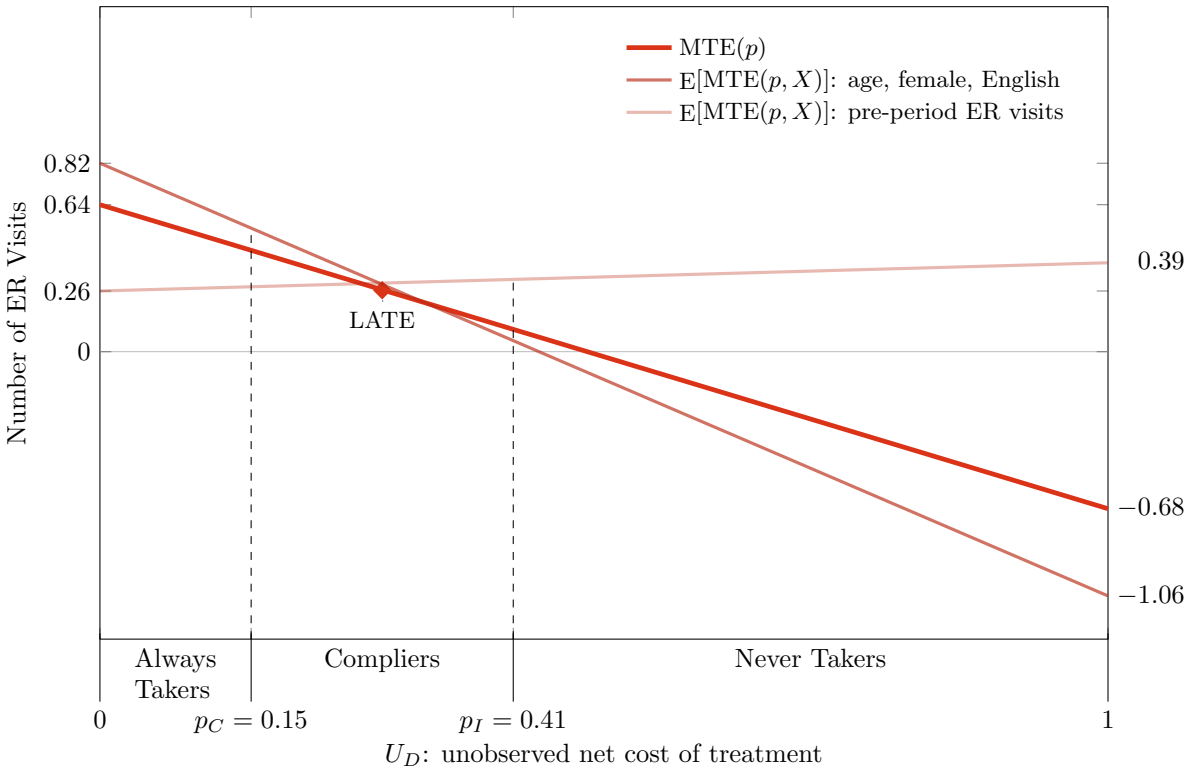
4.2.2 Treatment Effect Heterogeneity in Massachusetts

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 comparable individual-level data,[4] and I cannot identify average outcomes of always takers, compliers, and never takers with the available aggregate data.

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 a decreasing treatment effect on total health care spending implies a decreasing treatment effect on ER utilization along the unobserved

[4] 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 on ER visits aggregated to the county-quarter level, matched to county-level data on insurance before the reform because individual-level data on ER utilization and insurance coverage before and after the reform are not available. In Kolstad and Kowalski (2012), we use data from the Behavioral Risk Factor Surveillance System (BRFSS), which contains all the necessary elements except ER utilization. We 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.

Figure 8: 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, 1–3 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 common observables (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.

net cost of treatment U_D.

To examine treatment effect heterogeneity on total health care utilization within Massachusetts, 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. As a measure of total health care utilization, I consider the log premium, which captures total insurer costs because insurers must collect premiums to recover their costs. In Figure 9, 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 Hackmann et al. (2015), without the benefit of the separate definitions of selection and treatment effect heterogeneity that depend on whether costs are on behalf of the insured or uninsured, I interpret the cost curve in Figure 9 as a function that captures only adverse selection. Here, I interpret it as a function that captures treatment effect heterogeneity because it represents the difference between insured and uninsured costs. 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.

4.3 MTE-Reweighting Can Reconcile Oregon and Massachusetts LATEs

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

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 support $p_L < U_D \leq p_H$ as follows:

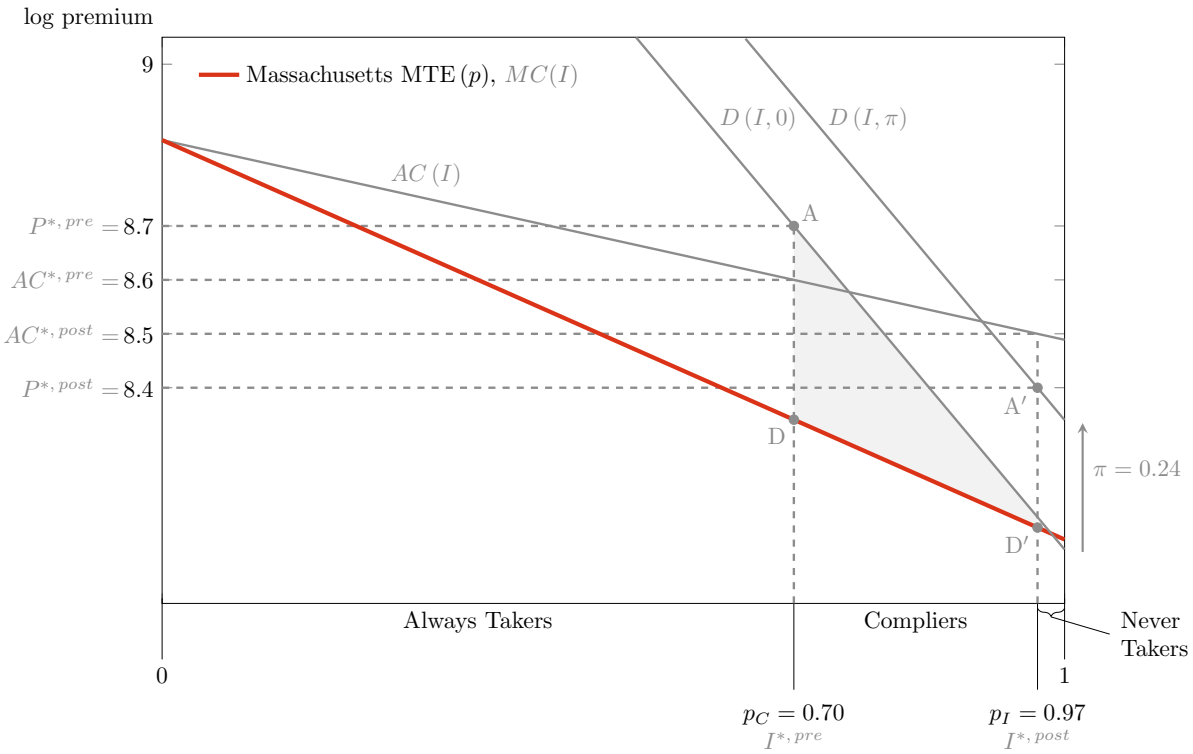
$$\mathrm{E}\left[Y_T \mid p_L < U_D \leq p_H\right] = \int_0^1 \omega(p, p_L, p_H)\mathrm{MTO}(p)\,\mathrm{d}p \tag{9}$$

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

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

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

Figure 9: Figure from Hackmann et al. (2015) Recast as Massachusetts MTE(p) Shows Treatment Effect Heterogeneity on Total Health Care Spending



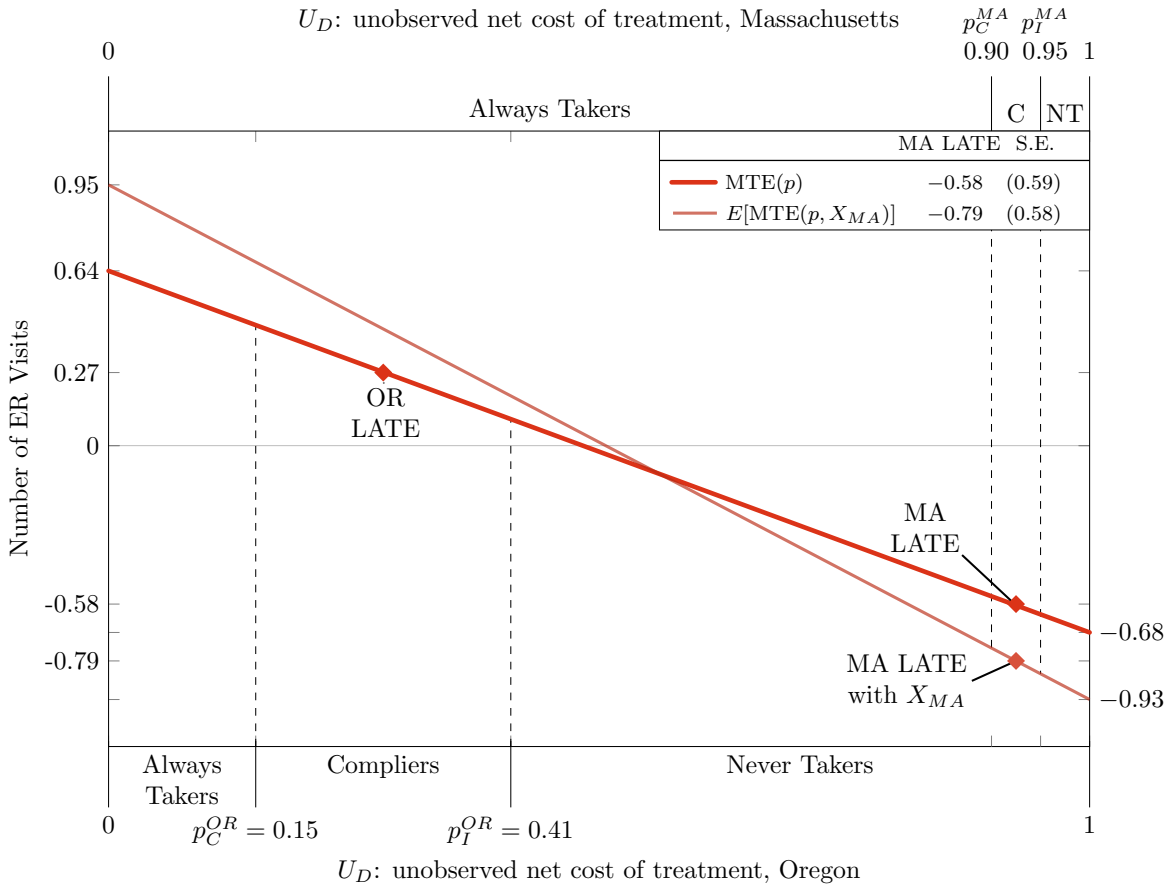
the average outcomes for Oregon always takers ($0 \leq U_D \leq p_C$), compliers ($p_C < U_D \leq p_I$), and never takers ($p_I < U_D \leq 1$) reported in Table 1. These weights also allow me to predict an untreated average outcome for Oregon always takers, a treated average outcome for Oregon never 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 6. These predictions indicate that always takers would visit the ER 1.35 times if they were not enrolled in Medicaid, implying a positive always taker average treatment effect of 0.54 visits. In contrast, never takers would visit the ER 0.55 times if they were enrolled in Medicaid, implying a negative never taker average treatment effect of 0.29 visits.

To extrapolate treatment effect heterogeneity from Oregon to Massachusetts, I reweight the Oregon MTE(p) over the implied support of the unobserved net cost of treatment U_D for Massachusetts compliers ($p_C^{MA} < U_D \leq p_I^{MA}$) via (11). To demonstrate the approach graphically, Figure 10 reproduces MTE(p) from Oregon and labels the support for Massachusetts compliers. As shown, the extrapolated Massachusetts LATE predicts that the Massachusetts reform decreased ER utilization by an average of 0.58 ER visits among Massachusetts compliers. To put the magnitude 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.[5] The decrease that I find over the 19 months from March 10, 2008 to September 30, 2009 translates into a decrease of 0.37 visits per person per year (=(0.58/19)*12), which is smaller but still negative. Therefore, reweighting the Oregon MTE(p) function can reconcile the increase in ER utilization in Oregon with the decrease 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

[5]Other estimates are not directly comparable. Chen et al. (2011) do not provide an estimate but report 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 10: Reweighting of MTE(p) or MTE(p,x) with Massachusetts Observables Reconciles Positive Oregon LATE and Negative Massachusetts 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. The treatment, instrument, "Age," "Female," and "English" in Massachusetts and Oregon are defined as in Table 2. 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, and p_I^{OR} is the probability of treatment in the intervention group in my full analysis sample for Oregon. p_C^{MA} the probability of treatment in the control group, and p_I^{MA} the probability of treatment in the intervention group in my full analysis sample for Massachusetts.

observables in the extrapolation, I evaluate the Oregon MTE(p,x) at the Massachusetts covariate vector X_{MA}, and I reweight the resulting function over the support for Massachusetts compliers. I depict the extrapolation in Figure 10. The resulting LATE 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.

Although my estimates can reconcile the results in terms of magnitudes, the extrapolated Massachusetts LATE is imprecise in both specifications. While the extrapolation provides a summary measure of the reconciliation exercise, it should not and does not constitute the *entire* reconciliation exercise. Rather, the extrapolation builds upon the evidence of comparable selection and treatment effect heterogeneity in the two settings that helps to justify the underlying assumptions. All of these components together comprise the reconciliation exercise.

Beyond imprecision, another concern with the extrapolation is that it averages the Oregon MTE function over a support of the unobserved net cost of treatment U_D that is far from the support for Oregon compliers. In this support, the linearity assumptions may not be appropriate. However, the estimated MTE function crosses the horizontal axis relatively close to the support for Oregon compliers, at an unobserved net cost of treatment U_D equal to 0.48, such that it is negative in more than half of its support. Therefore, even if the true Oregon MTE function is not linear at the extremes of the support, the extrapolation still qualitatively shows that coverage decreases ER utilization for many never takers, some who are relatively similar to compliers in terms of their unobserved net cost of treatment U_D.

4.3.2 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 LATE-reweighting (Hotz et al., 2005; Angrist and Fernandez-Val, 2013). One limitation of LATE-reweighting is that it only incorporates observables that are available in both contexts. The only observables available for all individuals in the Oregon and Massachusetts data are age, sex, and English-speaking status. Another limitation of LATE-reweighting is that it requires discretization of the observables, and it is unclear which discretization to use. To

illustrate a plausible approach, 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, with weights determined by the joint frequency of the three variables among Massachusetts compliers. This approach yields an increase of 0.23 visits among Massachusetts compliers, which is positive and therefore cannot reconcile the results.

It is not surprising that this approach to LATE-reweighting cannot reconcile the results because the observables on which it is based cannot explain treatment effect heterogeneity within Oregon. However, it is not clear that I could reconcile the results with LATE-reweighting even if a wider array of observables were available in both contexts. Taubman et al. (2014) report LATEs within many subgroups in Oregon, and almost all are positive. Any reweighting of positive LATEs from Oregon with positive weights from Massachusetts would not yield a negative estimate required to reconcile the results. Furthermore, even LATE-reweighting with observables that can potentially explain treatment effect heterogeneity within Oregon, such as previous ER utilization, would not necessarily reconcile the results. LATEs only give treatment effects on compliers, even within subgroups determined by observables. 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 observables. MTE-reweighting allows me to extrapolate from Oregon to Massachusetts using an unobservable that captures health.

5 Conclusion

I aim to shed light on why ER utilization increased following the Oregon Health Insurance Experiment but decreased following the Massachusetts reform. In both Oregon and Massachusetts, I find comparable evidence of adverse selection along the unobservable that separates always takers, compliers, and never takers. These patterns indicate that health improves from always takers to compliers to never takers. I also find comparable treatment effect heterogeneity, indicating that the effect of insurance on ER utilization decreases always takers to compliers to never takers. In Oregon, 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.

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.

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. *Journal of Business & Economic Statistics*, 38(3):593–612, 2020.

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.

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.

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.

Liran Einav, Amy Finkelstein, and Mark R Cullen. Estimating welfare in insurance markets using variation in prices. *The Quarterly Journal of Economics*, 125(3):877, 2010.

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. Health reform, health insurance, and selection: Estimating selection into health insurance using the massachusetts health reform. *American Economic Review Papers And Proceedings*, 102(3):498–501, May 2012.

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. *Health affairs (Project Hope)*, 29(3):389–432, 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.

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, 2018.

Amanda E Kowalski. How to examine external validity within an experiment. Working Paper

24834, National Bureau of Economic Research, October 2020a.

Amanda E Kowalski. Behavior within a clinical trial and implications for mammography guidelines. Working Paper 25049, National Bureau of Economic Research, November 2020b.

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.

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. *The annals of mathematical statistics*, 11(3):284–300, 1940.

Online Appendix

Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform

Amanda E. Kowalski

Online 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:

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

Online 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

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

Online Appendix C Derivation of Average Outcomes

Under the assumptions of the model, I identify the expected value of Y_T for always takers as follows, suppressing X for simplicity:

$$\mathrm{E}[Y \mid D = 1, Z = 0] = \mathrm{E}[Y_U + D(Y_T - Y_U) \mid D = 1, Z = 0] \qquad (\text{by (5)})$$

$$
\begin{aligned}
&= \mathrm{E}[Y_T \mid D=1, Z=0] \\
&= \mathrm{E}[Y_T \mid 0 \le U_D \le p_C, Z=0] && \text{(by (4), where } p_C = P(D=1|Z=0)) \\
&= \mathrm{E}[g_T(U_D, \gamma_T) \mid 0 \le U_D \le p_C, Z=0] && \text{(by (6))} \\
&= \mathrm{E}[g_T(U_D, \gamma_T) \mid 0 \le U_D \le p_C] && (Z \perp (U_D, \gamma_T) \text{ by A.4}) \\
&= \mathrm{E}[Y_T \mid 0 \le U_D \le p_C].
\end{aligned}
$$

Derivations for compliers and never takers are similar.

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

First, estimate a propensity score, $\widehat{p}$, for each individual 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. Second, estimate the average treated outcome (ATO) function, defined as $\mathrm{ATO}(p,x) = \mathrm{E}[Y_T \mid 0 \le U_D \le p, X = x] = \widetilde{\delta}_T' x + \widetilde{\lambda}_T p$, 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 $\mathrm{MTO}(p,x) = \frac{\mathrm{d}[p\mathrm{ATO}(p,x)]}{\mathrm{d}p}$. Therefore, $\mathrm{MTO}(p,x) = \widetilde{\delta}_T' x + 2\widetilde{\lambda}_T p = \delta_T' x + \lambda_T p$. So, estimates of the MTO parameters can be constructed as follows: $\delta_T = \widetilde{\delta}_T$ and $\lambda_T = 2\widetilde{\lambda}_T$. Third, estimate the average untreated outcome (AUO) function, defined as $\mathrm{AUO}(p,x) = \mathrm{E}[Y_U \mid p < U_D \le 1, X = x] = \widetilde{\delta}_U' x + \widetilde{\lambda}_U p$, 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 $\mathrm{MUO}(p,x) = \frac{\mathrm{d}[(1-p)\mathrm{AUO}(p,x)]}{\mathrm{d}(1-p)}$. Therefore, $\mathrm{MUO}(p,x) = \widetilde{\delta}_U' x - \widetilde{\lambda}_U + 2\widetilde{\lambda}_U p = \delta_U' x + \lambda_U p$. 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 $\widetilde{\lambda}_U$. Fourth, construct the estimate for MTE(p,x) using the estimated parameters of MTO(p,x) and MUO(p,x).

References

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