Behavior within a Clinical Trial and Implications for Mammography Guidelines*

Amanda E. Kowalski

December 12, 2018

Abstract

I examine behavior within a clinical trial to inform treatment guidelines. I use data from the Canadian National Breast Screening Study, an influential clinical trial on mammography. During the active study period of the trial, a substantial fraction of women in the control group received mammograms, and some women in the intervention group did not. Using this mammography behavior, random assignment within the trial, and a standard model from the economics literature, I divide participants into three groups that differ in how likely they are to receive mammograms. Making comparisons across these groups, I find two important relationships. First, I find heterogeneous selection into mammography: women more likely to receive mammograms are healthier. I find this relationship using a marginal treatment effect model that assumes no more than the local average treatment effect assumptions. Second, I find treatment effect heterogeneity along the margin of selection into mammography: women more likely to receive mammograms are more likely to experience harm from them. I find this relationship using an ancillary assumption that builds on the first empirical relationship. I find additional empirical support for the ancillary assumption using baseline covariates. My findings contribute to the literature concerned about harms from mammography by demonstrating variation across the margin of selection into mammography. This variation is problematic for current mammography guidelines for women in their 40s because it implies that they unintentionally encourage mammography for healthier women who are more likely to experience harm from them.

*aekowals@umich.edu. Saumya Chatrath, Neil Christy, Tory Do, Bailey Flanigan, Pauline Mourot, Dominik Piehlmaier, Ljubica Ristovska, Sukanya Sravasti, and Matthew Tauzer provided excellent research assistance. I thank Anthony Miller, Teresa To, Cornelia Baines, and Claus Wall, investigators of the Canadian National Breast Screening Study, for sharing data and for answering questions. I thank Magne Mogstad, Pat Kline, Atheendar Venkataramani, and seminar participants at the American Economic Association Annual Meeting, the Canadian Health Economics Study Group, the London-Paris Public Economics Workshop, the North American Summer Meetings of the Econometric Society, Princeton, and the University of Michigan for helpful comments. NSF CAREER Award 1350132 and NIA Grant P30-AG12810 provided support. I dedicate my research on breast cancer to Elisa Long.

1 Introduction

The U.S. Preventive Service Task Force (USPSTF) revived the debate on mammography when they updated their mammography guidelines in 2009 (U.S. Preventive Services Task Force, 2009). Although their previous guidelines recommended regular mammography for women aged 40 and older (U.S. Preventive Services Task Force, 2002), their updated guidelines left the mammography decision for women in their 40s to individual women and their doctors. The precise USPSTF guidelines for women in their 40s, as confirmed in 2016, state: "The decision to start screening mammography in women prior to age 50 years should be an individual one. Women who place a higher value on the potential benefit than the potential harms may choose to begin biennial screening between the ages of 40 and 49 years" (Siu, 2016).

These guidelines raise two important empirical questions. The first concerns heterogeneous selection into mammography: are women who are more likely to receive mammograms under these guidelines different? The second concerns heterogeneous treatment effects from mammography: are women who are more likely to receive mammograms under these guidelines more likely to experience benefits or harms from them?

To answer these questions, I examine behavior within a clinical trial. The medical literature on mammography, including the meta-analysis that informs the 2016 USPSTF mammography guidelines (Nelson et al., 2016), focuses on evidence from clinical trials (Moss et al., 2015; Miller et al., 2014; Bjurstam et al., 2003; Nyström et al., 2002; Tabar et al., 1995; Habbema et al., 1986). This evidence is based on the comparison of health outcomes between randomly assigned intervention and control groups. However, it rarely incorporates analysis of mammogram takeup, which reflects behavior as well as random assignment. In contrast, the economics literature on mammography focuses on impacts of policy interventions on mammography behavior (see Bitler and Carpenter (2016); Buchmueller and Goldzahl (2018); Kim and Lee (2017); Zanella and Banerjee (2016); Kadiyala and Strumpf (2016)). Uniting the medical literature with the economics literature, I focus on relationships between mammography behavior and health outcomes within a clinical trial. These relationships inform heterogeneous selection into mammography and accompanying heterogeneous treatment effects on health outcomes, which in turn inform the impact of mammography guidelines.

To conduct my analysis, I use individual-level data from the Canadian National Breast Screening Study (CNBSS), an influential clinical trial cited by the USPSTF in its mammography guidelines. The CNBSS enrolled almost 90,000 participants aged 40-59 between 1980 and 1985. All participants were randomly assigned to one of two groups: an intervention group and a control group. Intervention group participants received access to annual mam-

mograms during an active study period, consisting of the enrollment year and 3 to 4 years after enrollment. Control group women in their 40s at enrollment received usual care in the community, and control group women in their 50s at enrollment received access to annual clinical breast examinations in each year of the active study period. Given the change in mammography guidelines for women in their 40s, I focus on women in their 40s at enrollment, and I examine the robustness of my findings on women in their 50s at enrollment.

The CNBSS data allow me to examine mammography behavior and mortality for all participants. To the best of my knowledge, the CNBSS is the only trial considered by the meta-analsyis that informs the USPSTF guidelines that tracked the actual takeup of mammograms for all participants. These data show that a substantial fraction of women in the control group received mammograms, and some women in the intervention group did not. This variation in mammography behavior is crucial for my analysis. Furthermore, the CNBSS data allow me to observe mortality for all participants because they are linked to the Canadian Mortality Database. Among the trials on mammography considered by the meta-analsyis that informs the USPSTF guidelines, the CNBSS is the only trial with data that allows for examination of mortality for at least 20 years after enrollment for all participants (Nelson et al., 2016). The ability to examine mortality over a long time horizon proves important to my results.

To examine relationships between mammography and mortality, I begin with a standard treatment effect model in which the "treatment" is mammography. I present the model as a generalized Roy (1951) model of the marginal treatment effect (MTE) as introduced by Björklund and Moffitt (1987), in the tradition of Heckman and Vytlacil (1999, 2001, 2005), Carneiro et al. (2011), and Brinch et al. (2017). The specific MTE model that I use can also be characterized as a model that assumes no more than the local average treatment effect (LATE) assumptions of independence and monotonicity proposed by Angrist and Imbens (1994), given the proof by Vytlacil (2002).

Given the equivalence of the model to the LATE assumptions, I emphasize the link between the MTE model and the Angrist et al. (1996) terminology used in the LATE literature. In this terminology, "always takers" receive treatment regardless of random assignment, "compliers" receive treatment if and only if they are assigned to the intervention group, and "never takers" do *not* receive the treatment regardless of random assignment. The model excludes "defiers," who receive treatment if and only if they are assigned to the control group. It is possible to identify some individuals as always takers because they receive treatment despite assignment to the control group, and it is possible to identify other individuals as never takers because they do not receive treatment despite assignment to the intervention group. It is not possible to identify the remaining individuals as members of any one group.

However, the assumptions of the model, which rely on the randomization and the exclusion of defiers, make it possible to calculate aggregate statistics on always takers, compliers, and never takers. These statistics are useful because they allow for comparisons across three groups, not just the two groups generated by the intervention and control.

In my presentation of the model, I emphasize that it implies an ordering from always takers to compliers to never takers. This ordering was originally shown by Vytlacil (2002). I make the ordering clear using a simple figure that illustrates implications of the first stage of the model. Given the equivalence of the model to the LATE assumptions, I emphasize that this ordering is also true under the LATE assumptions. Within the experiment, always takers are most likely to receive treatment (they receive treatment with probability 1), followed by compliers (they receive treatment with probability equal to the probability of assignment to the intervention group), followed by never takers (they receive treatment with probability 0). The second stage of the model allows for differences in mortality across always takers, compliers, and never takers. I use the ordering across these groups to identify selection heterogeneity. I also use the ordering to motivate the ancillary assumption that I impose to identify treatment effect heterogeneity.

I define selection and treatment effect heterogeneity using functions from the MTE literature. The definition of selection heterogeneity generalizes the definition of "selection bias" used by Angrist (1998) and Heckman et al. (1998) among others. Under that definition, there is selection bias if the average untreated outcome of the treated participants is not equal to the average untreated outcome of the untreated participants. I show that within a trial, selection bias depends on the probability of assignment to the intervention group, a parameter explicitly chosen as part of the trial design, but selection heterogeneity does not. More importantly, selection bias is not identified under the model alone because it is not possible to calculate the average untreated outcome of all treated participants. However, an alternative special case of selection heterogeneity is identified because the trial generates exogenous variation in which participants receive treatment, making it possible to calculate the average untreated outcome of some treated participants.

I identify a special case of selection heterogeneity in the CNBSS without any ancillary assumptions. I do so using a test that I refer to as the "untreated outcome test" because it compares the average untreated outcomes of compliers and never takers. The untreated outcome test is equivalent or similar to tests proposed in the econometric literature by Bertanha and Imbens (2014), Guo et al. (2014), and Black et al. (2015) and generalized by Mogstad et al. (2018).[1] Unlike previous literature, I show that the untreated outcome

[1]The test proposed by Bertanha and Imbens (2014) is similar because they develop their test for a regression discontinuity context, but it is effectively an equivalent test. Bertanha and Imbens (2014) propose this test as one component of a test for external validity, but they do not propose it as a test of selection

test identifies a special case of selection heterogeneity. I find selection heterogeneity in the CNBSS: women more likely to receive mammograms are healthier.

I identify a special case of treatment effect heterogeneity in the CNBSS using an ancillary assumption that builds upon empirical selection heterogeneity. It requires weak monotonicity of marginal untreated outcomes along the margin of selection into treatment, where empirical selection heterogeneity determines the direction of the weak monotonicity. Brinch et al. (2017) impose this assumption in conjunction with a corresponding assumption on treated outcomes to test for treatment effect homogeneity. I demonstrate that only one assumption is necessary to test for treatment effect heterogeneity.[2] Furthermore, I demonstrate that this assumption yields a one-sided bound on the average treatment effect for always takers that is of interest in its own right. The ancillary assumption that I impose is weaker than related assumptions made by Olsen (1980), Heckman (1979), and Brinch et al. (2017), discussed by Kline and Walters (2018). In the context of the CNBSS, given empirical selection heterogeneity, the ancillary assumption implies that women more likely to receive mammograms have weakly better health, measured by long-term mortality in the absence of mammograms. Measures of socioeconomic status and health behavior collected at baseline provide support for the ancillary assumption. Under the ancillary assumption, I find treatment effect heterogeneity: women more likely to receive mammograms are more likely to be harmed by them.

The possibility that harms of mammograms can outweigh benefits is surprising, but an extensive literature considers the possibility (Lannin and Wang, 2017; Baines et al., 2016; Nelson et al., 2016; Miller et al., 2014; Bleyer and Welch, 2012; Baum, 2013). The article that conveys the 2016 USPSTF guidelines notes, "The most important harm is the diagnosis and treatment of noninvasive breast cancer that would otherwise not have become a threat to a woman's health, or even apparent, during her lifetime (that is, overdiagnosis and overtreatment)" (Siu, 2016). To illustrate how this mechanism could lead the harms of mammograms to outweigh the benefits, suppose that two women receive mammograms. Both are diagnosed with breast cancer, and both indeed have breast cancer. Unbeknownst to the women and their doctors, one woman would die within 20 years in the absence of breast

heterogeneity. Similarly, Guo et al. (2014) propose this test as one component of a test for unmeasured confounding, but they do not discuss it as a test for selection heterogeneity. Black et al. (2015) propose this test as one of two tests for selection bias.

[2]This paper includes material from NBER Working Paper 22363, "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments" (Kowalski, 2016). I include content related to weak monotonicity assumptions in this paper, and I include content related to linearity assumptions, as well as content related to the Oregon Health Insurance Experiment, in Kowalski (2018a). In Kowalski (2018b), I discuss external validity for a nontechnical audience, and I do not break new ground. I co-developed the Stata commands mtemore to accompany (Kowalski, 2016) and mtebinary to accompany Kowalski (2018a).

cancer treatment, but the other woman would not because her tumor would grow more slowly. Unable to separate the two, both women receive breast cancer treatment, which has its own mortality risks. Both women die within 20 years. The first woman would have died in the absence of breast cancer treatment, so she is neither harmed nor helped. However, the second women would have survived in the absence of breast cancer treatment, so she is harmed. In this example, the harms of mammograms outweigh the benefits.

My finding that the harms of mammograms outweigh the benefits for the women most likely to receive them is problematic for the current USPSTF mammography guidelines for women in their 40s, which leave the mammography decision to individual women and their doctors. Beyond the mammography context, my analysis demonstrates the importance of examining relationships between behavior and health outcomes in a world that encourages personalized health care. Fortunately, some relationships can be identified in existing clinical trial data.

In the next section, I begin by replicating previous results from the CNBSS. In Section 3, I model selection and treatment effect heterogeneity within the CNBSS. In Section 4, I identify selection and treatment effect heterogeneity to arrive at my two main findings, and I present evidence in support of the ancillary assumption required for my second finding. I show that my results are robust to a wide variety of alternative specifications in Section 5. I conclude by discussing implications for mammography guidelines and future research in Section 6.

2 Replication of CNBSS Results

A great deal has been written on the CNBSS in the medical literature. Viewing the CNBSS as an influential trial, my focus is not to evaluate the CNBSS itself or previous work on it. Rather, my focus is to extend analysis of the CNBSS to examine selection and treatment effect heterogeneity. Using CNBSS data, I am able to produce an exact replication of the latest result published by CNBSS investigators in Miller et al. (2014). This result shows that access to mammography does not have a statistically significant impact on breast cancer mortality, which is consistent with results published by CNBSS investigators at earlier follow-up lengths (Miller et al., 1992a,b, 1997, 2000, 2002, 2014). This result is also consistent with other clinical trial results considered by the USPSTF (Nelson et al., 2016).

In the replication that serves as the foundation for my subsequent analysis, to increase the relevance of my findings to the USPSTF guidelines for women in their 40s, I depart from the exact replication of Miller et al. (2014) in four ways. First, because the USPSTF guidelines changed specifically for women in their 40s, I only include women aged 40-49 at enrollment in my main analysis sample, and I examine the robustness of my results among women aged 50-59 at enrollment. Second, because the USPSTF guidelines are intended for

asymptomatic women without a genetic predisposition for breast cancer, I exclude women from my main analysis sample if they have any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer. My main analysis sample includes 19,505 women. I examine robustness in the full sample of 50,430 women aged 40-49 at enrollment and in the subsample of excluded women. Third, I focus on all-cause mortality, which I refer to as "mortality" for simplicity, because mortality is ultimately the most important outcome for guidelines to consider, but I examine robustness to breast cancer mortality. Fourth, to make the timing of my findings easier to interpret, I focus on results at a fixed follow-up length of 20 years after enrollment, as opposed to a fixed cutoff date that reflects various follow-up lengths. I also examine robustness at earlier follow-up lengths to understand whether more limited follow-up data would still yield the same implications for mammography guidelines.

3 Model

I use an MTE model to allow for selection and treatment effect heterogeneity within the CNBSS. I make only stylistic changes to the model used by Heckman and Vytlacil (2005)[3] to ensure that the model assumes no more than the LATE assumptions of Angrist and Imbens (1994), as proven by Vytlacil (2002). I emphasize that the model implies an ordering from always takers to compliers to never takers. I present implications of this ordering using simple figures.

3.1 First Stage: Mammography

In the context of the CNBSS, I use "treatment" to refer to mammography, which I represent with D. I define mammography such that $D = 1$ if a participant receives a mammogram in at least one year during the active study period after the enrollment year, and I set $D = 0$ otherwise. If data is missing in any year, I construct D such that the participant did not receive a mammogram in that year.

Let V_T represent potential utility in the treated state, and let V_U represent potential utility in the untreated state. I relate the potential utilities to realized utility V such that:

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

[3]One stylistic change that I make is that I do not condition on an optional covariate vector. Because randomization in the CNBSS was not stratified, randomization was not conditional on a covariate vector. The absence of conditioning on a covariate vector simplifies the exposition and emphasizes the role of the unobserved net cost of treatment.

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

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

where $\mu_D(\cdot)$ is an unspecified function, Z is an observed binary instrument such that $Z = 1$ represents random assignment to the intervention group and $Z = 0$ represents random assignment to the control group, and ν_D is an unobserved term with an unspecified distribution. I assume:

A.1. (Instrument Relevance) $\mu_D(Z)$ is a nondegenerate random variable,

A.2. (Continuity) The cumulative distribution function of ν_D, which I denote with $F(\cdot)$, is absolutely continuous with respect to the Lebesgue measure.

A.3. (Independence) The random vectors (U_D,γ_T) and (U_D,γ_U) are independent of Z, where $U_D = F(\nu_D)$, and γ_T and γ_U are unobserved terms introduced in the second stage.

A.1 is verifiable. Under A.2, the transformation of ν_D by $F(\cdot)$ is a normalization that implies that $U_D = F(\nu_D)$ is uniformly distributed between 0 and 1. For completeness, I show the proof in Appendix A. The term ν_D enters negatively into the net benefit of treatment in (2), so I interpret it as a net cost of treatment. I therefore interpret U_D as the normalized "unobserved net cost of treatment."

The current USPSTF guidelines recommend mammography for women in their 40s "who place a higher value on the potential benefit than the potential harms" (Siu, 2016). In terms of the model, the guideline recommends mammography for women with a value of (2) that is greater than zero. These women are precisely the women who receive mammograms within the CNBSS. As I show for completeness in Appendix B, given (2) greater than 0, A.3 implies the following treatment equation:

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

This equation shows that women receive mammograms if and only if their unobserved net cost of treatment U_D is weakly less than an observed threshold. If A.1 holds, then the observed threshold is different for the control and intervention groups, resulting in two special cases of the treatment equation:

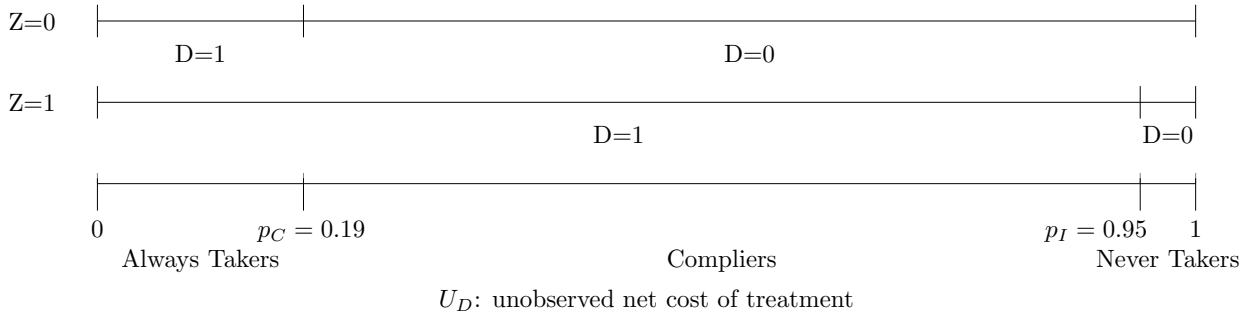
$$D = 1\{U_D \leq p_C\} \qquad \text{where } p_C = P\big(D = 1 \mid Z = 0\big), \tag{4}$$

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

where the treatment probabilities p_C and p_I can be estimated in the control group ($Z = 0$) and the intervention group ($Z = 1$), respectively.

I present the implications of the first stage of the model graphically in Figure 1 using values from my main analysis sample from the CNBSS. The net unobserved cost of treatment U_D ranges from 0 to 1. Using (4) and (5), I partition the unobserved net cost of treatment U_D into three distinct ranges. The top line depicts the ranges for control group participants. In my main analysis sample from the CNBSS, 19% of control group participants receive mammograms, so $p_C = 0.19$. By (4), control group participants that receive mammograms have $0 \leq U_D \leq 0.19$. In the Angrist et al. (1996) terminology from the LATE literature, these participants must be "always takers," participants who receive the treatment regardless of random assignment. The middle line of Figure 1 depicts ranges of U_D for intervention group participants. In my main analysis sample, 95% of intervention group participants receive mammograms, so $p_I = 0.95$. By (5), intervention group participants that do not receive mammograms have $0.95 < U_D \leq 1$. These participants must be "never takers," participants who do *not* receive the treatment regardless of random assignment (Angrist et al., 1996). I depict U_D for participants in the control and intervention groups on the same axis in the bottom line of Figure 1. Participants in the middle range ($0.19 < U_D \leq 0.95$) receive mammograms if and only if they are assigned to the intervention group. They must be "compliers," participants who receive the treatment if and only if they are assigned to the intervention group (Angrist et al., 1996). The depiction in the bottom line of Figure 1 emphasizes the ordering from always takers to compliers to never takers. In the CNBSS, this ordering reflects behavior within the trial: always takers are the most likely to receive mammograms, followed by compliers, followed by never takers.

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



Note. The treatment is mammography, which is equal to one if a participant receives a mammogram in at least one year during the active study period after the enrollment year. Missing mammogram data in any year is set to no mammogram in that year. The sample includes women aged 40-49 at enrollment, excluding women with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer.

3.2 Second Stage: Mortality

I relate mortality Y to mammography D as follows:

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

Y_T represents potential treated mortality and Y_U represents potential untreated mortality, which I specify as follows:

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

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

where $g_T(\cdot)$ and $g_U(\cdot)$ are unspecified functions, U_D is the unobserved net cost of treatment from the first stage, and γ_T and γ_U are unobserved terms with unspecified distributions. I assume:

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

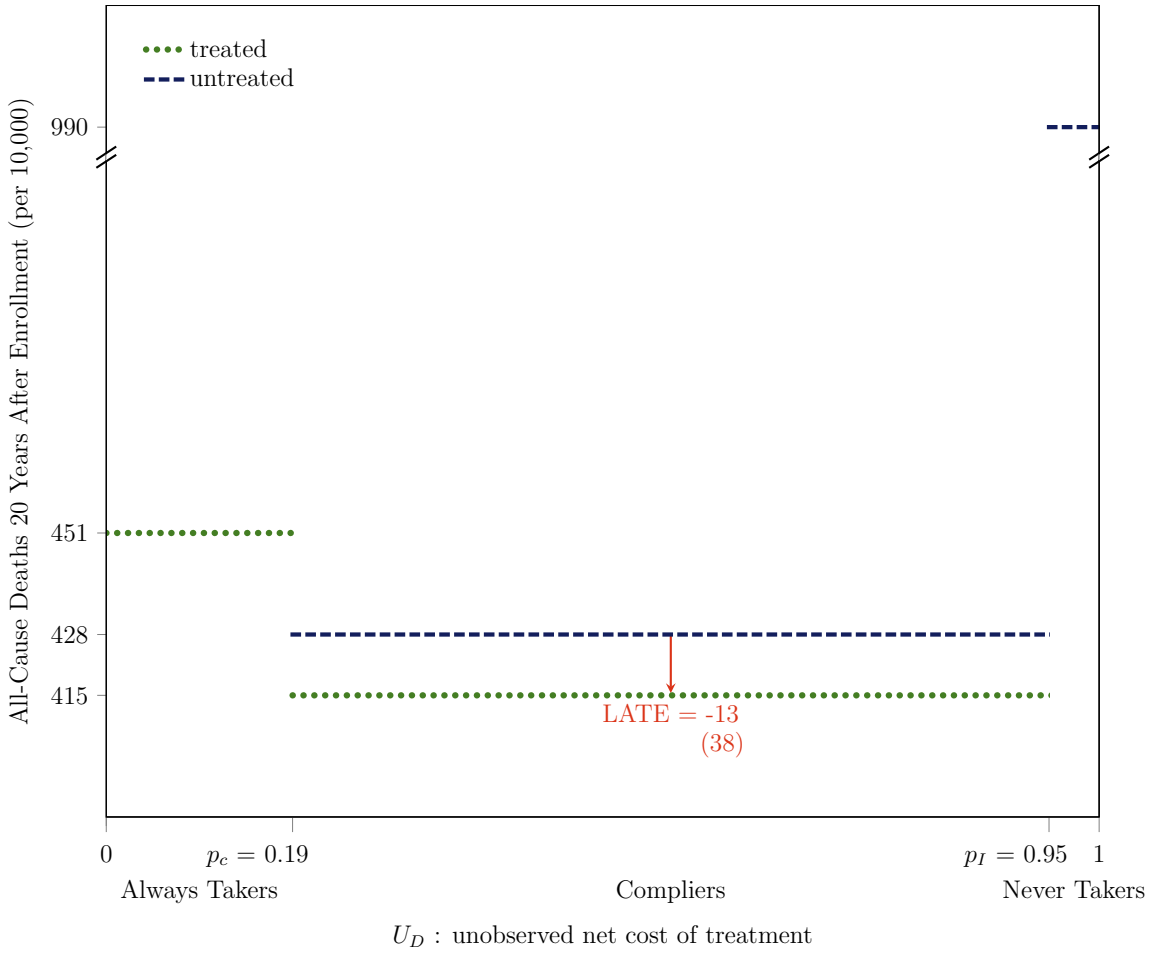
A.5. (Finite Average Outcomes) The values of $E[Y_T]$ and $E[Y_U]$ are finite.

A.4 is verifiable. A.5 ensures that average treated and untreated potential outcomes are defined.

The model, given by the utility equations (1) and (2), the treatment equations (3)–(5), the potential outcome equations (6)–(8), and assumptions A.1–A.5, assumes no more than the LATE assumptions. This claim follows because my algebraic presentation of the model differs only stylistically from the model presented in Heckman and Vytlacil (2005). Heckman and Vytlacil (2005) invokes the proof in Vytlacil (2002) to claim that the model assumes no more than the LATE assumptions.

I illustrate implications of the model using statistics from my main analysis sample from the CNBSS in Figure 2. The horizontal axis depicts implications of the first stage as in Figure 1. The vertical axis depicts implications of the second stage in terms of average mortality with and without mammograms. Always takers always receive mammograms within the experiment, so it is not possible to derive their average mortality without mammograms without ancillary assumptions. Similarly, never takers never receive mammograms within the experiment, so it is not possible to derive their average mortality with mammograms without ancillary assumptions. However, it is possible to derive the other average mortality statistics that I report in Figure 2 using the model, as I show in Appendix C. These derivations yield the same values that I would obtain using the derivations by Imbens and Rubin (1997), Katz et al. (2001), Abadie (2002), and Abadie (2003), which rely on the LATE assumptions.

Figure 2: Average Mortality for Always Takers, Compliers, and Never Takers



Note. Bootstrapped standard errors in parentheses. All-cause deaths are measured 20 years after enrollment for all participants, based on the exact calendar date of enrollment. The treatment is mammography, which is equal to one if a participant receives a mammogram in at least one year during the active study period after the enrollment year. Missing mammogram data in any year is set to no mammogram in that year. The sample includes women aged 40-49 at enrollment, excluding women with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer.

The depiction in Figure 2 makes clear that the LATE represents the average treatment effect on compliers. However, always and never takers make up sizeable fractions of the sample. Furthermore, their average mortality rates appear very different from the average mortality rates of compliers. By 20 years after enrollment, the cumulative mortality rate is 4.51% for always takers, but it is only 4.28% for untreated compliers and 4.15% for treated compliers. In stark contrast, the cumulative mortality rate for never takers is 9.90%, more than double any of the other reported rates, necessitating a break in the axis to avoid being "off the chart." These differences in mortality rates provide a starting point for identification of selection and treatment effect heterogeneity.

3.3 Definitions of Selection and Treatment Effect Heterogeneity in the Model

I define selection and treatment effect heterogeneity using the following functions from the MTE literature (see Carneiro and Lee, 2009; Brinch et al., 2017):

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

where p is a realization of the unobserved net cost of treatment U_D.

The first function, which I refer to as the "marginal untreated outcome (MUO)" function, defines selection heterogeneity along the entire margin of U_D. Although this function has been used in the literature, to the best of my knowledge, it has not been used as the definition of selection heterogeneity. In Appendix D, I show that "selection bias," as defined by Angrist (1998) and Heckman et al. (1998) among others, which is equal to the difference in average untreated outcomes between treated and untreated participants, is a special case of selection heterogeneity defined by the MUO function. Furthermore, I show that selection bias depends on the fraction of individuals assigned to the intervention group, a parameter explicitly chosen as part of the trial design. I also provide intuition for why selection bias is not identified without ancillary assumptions by showing that it depends on the average untreated outcome of always takers, which cannot be calculated within a trial without ancillary assumptions. I show in Section 4 that a different special case of selection heterogeneity is identified without ancillary assumptions because it compares the average untreated outcomes of compliers and never takers, which can be calculated within a trial without ancillary assumptions.

The second function, which I refer to as the "marginal treated outcome (MTO)" function, characterizes the sum of selection and treatment effect heterogeneity along the entire margin of U_D. Differences in untreated outcomes can only reflect selection heterogeneity, but differences in treated outcomes can reflect selection heterogeneity, treatment effect heterogeneity, or both. It is tempting to think that there should be no material distinction between treated outcomes and untreated outcomes in the definitions of selection and treatment effect heterogeneity. However, the treatment effect is defined as the treated outcome minus the untreated outcome, not the untreated outcome minus the treated outcome. Therefore, the treatment effect has magnitude *and* direction, which is why I represent the LATE with an arrow in Figure 2. Renaming the untreated outcome as the treated outcome and vice versa would change the direction of the treatment effect, illustrating why there is a material distinction between treated and untreated outcomes in the definitions.

The third function is the "marginal treatment effect (MTE)" function of Heckman and Vytlacil (1999, 2001, 2005). It defines treatment effect heterogeneity along the entire margin of U_D. In the CNBSS, the MTE function characterizes how the impact of mammography on mortality changes along the margin of selection into mammography.

4 Findings

Applying the model to the CNBSS, I identify and estimate two main relationships between mammography behavior and mortality. First, under the model that assumes no more than the LATE assumptions, I find selection heterogeneity: women who are more likely to receive mammograms are healthier. Second, under an ancillary assumption, I find treatment effect heterogeneity along the margin of mammography: women more likely to receive mammograms are more likely to be harmed by them. I find support for the ancillary assumption using covariates collected at baseline.

4.1 Selection Heterogeneity: Women More Likely to Receive Mammograms are Healthier

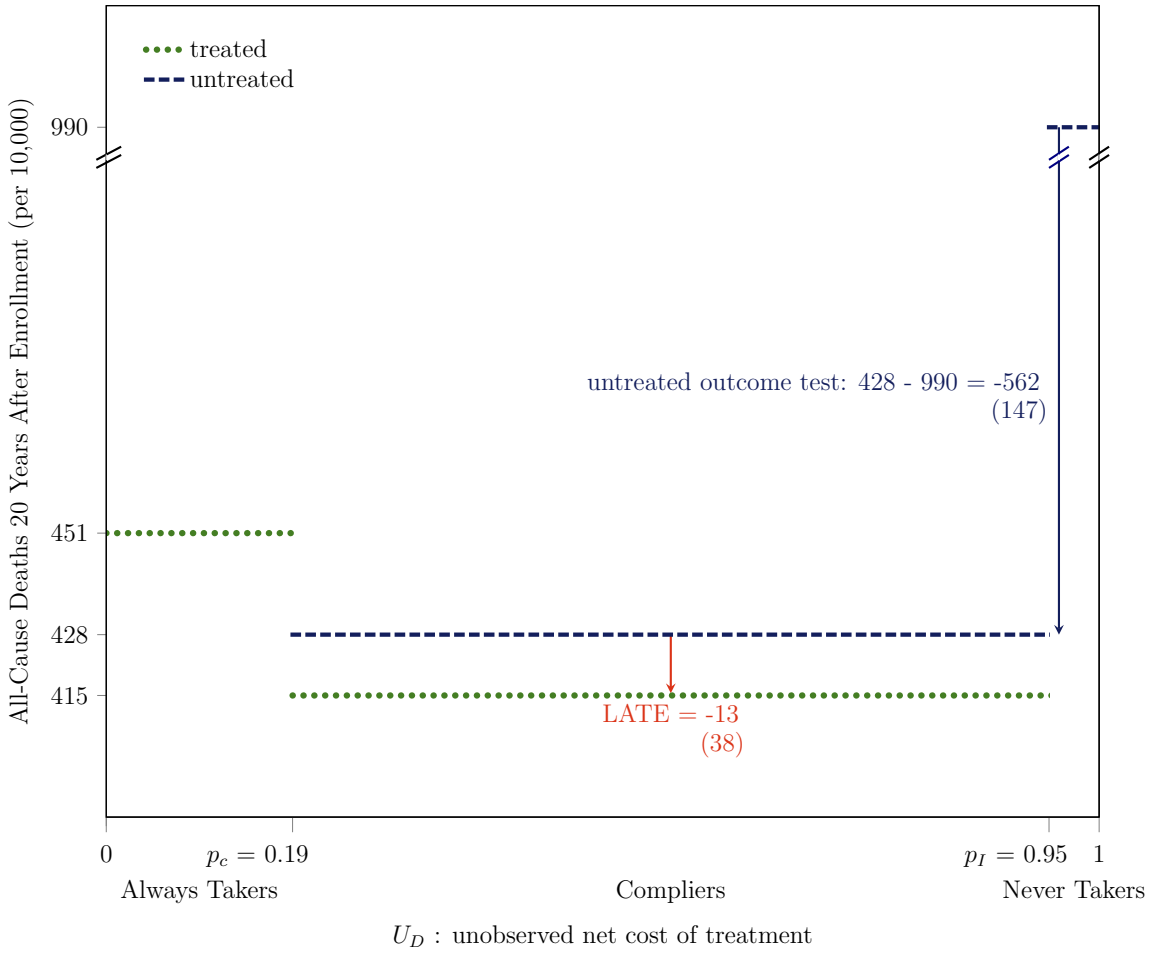
I identify selection heterogeneity using a test that I refer to as the "untreated outcome test" because it compares average untreated outcomes of compliers ($p_C < U_D \leq p_I$) and never takers ($p_I < U_D \leq 1$) with the following test statistic:

$$\mathrm{E}[Y_U \mid p_C < U_D \leq p_I] - \mathrm{E}[Y_U \mid p_I < U_D \leq 1] = \int_0^1 (\omega(p, p_C, p_I) - \omega(p, p_I, 1)) \,\mathrm{MUO}(p) \,\mathrm{d}p, \tag{9}$$

where $\omega(p, p_L, p_H) = 1\{p_L \leq p < p_H\}/(p_H - p_L)$. A negative test statistic indicates negative selection heterogeneity, and a positive test statistic indicates positive selection heterogeneity. The test of the null hypothesis that this test statistic is equal to zero is equivalent to or similar to tests proposed by Bertanha and Imbens (2014), Guo et al. (2014), and Black et al. (2015), which are generalized by Mogstad et al. (2018). Unlike previous literature, I define selection heterogeneity with the MUO function. I demonstrate that the untreated outcome test identifies a special case of selection heterogeneity by expressing the untreated outcome test statistic as a weighted integral of the MUO function in (9).

Applying the untreated outcome test to my main analysis sample from the CNBSS, I find that by 20 years after enrollment, never takers had died at a rate that was 5.62 percentage points higher than the untreated complier mortality rate of 4.28 percent. This difference is statistically different from zero, so I reject selection homogeneity. Mortality is a measure of health, and compliers are more likely to receive mammograms than never takers. Therefore, my finding indicates selection heterogeneity: women more likely to receive mammograms are

Figure 3: Untreated Outcome Test Rejects Selection Homogeneity:
Women More Likely to Receive Mammograms are Healthier



Note. Bootstrapped standard errors in parentheses. All-cause deaths are measured 20 years after enrollment for all participants, based on the exact calendar date of enrollment. The treatment is mammography, which is equal to one if a participant receives a mammogram in at least one year during the active study period after the enrollment year. Missing mammogram data in any year is set to no mammogram in that year. The sample includes women aged 40-49 at enrollment, excluding women with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer.

healthier.

4.2 Treatment Effect Heterogeneity: Women More Likely to Receive Mammograms are More Likely to be Harmed by Them

To identify treatment effect homogeneity, I impose the following ancillary assumption:

M.1. (Weak Monotonicity of the MUO Function) For all p_1, $p_2 \in [0,1]$ such that $p_1 < p_2$: $\mathrm{E}[Y_U \mid U_D = p_1] \leq \mathrm{E}[Y_U \mid U_D = p_2]$ or $\mathrm{E}[Y_U \mid U_D = p_1] \geq \mathrm{E}[Y_U \mid U_D = p_2]$.

While the model imposes LATE monotonicity in the first stage, as shown by Vytlacil (2002), this assumption imposes a corresponding weak monotonicity in the second stage. Brinch

et al. (2017) impose this assumption in conjunction with an analogous weak monotonicity assumption on the MTO function to test for treatment effect homogeneity. I emphasize that either of the Brinch et al. (2017) assumptions is sufficient. In Appendix E, I present the weak monotonicity assumption on the MTO function and an alternative weak montonicity assumption on the MTE function, and I discuss why I do not impose them in the CNBSS. I impose weak monotonicity on the MUO function via M.1 in the CNBSS because I can find empirical support for it using covariates.

In the CNBSS, the ancillary assumption M.1 implies that average health, measured by mortality in the absence of mammograms, varies monotonically from always takers to compliers to never takers. It is not possible to observe the average health of always takers in the absence of mammograms. However, the untreated outcome test shows that the average complier is healthier than the average never taker. Therefore, M.1 implies that the average always taker is weakly healthier than the average complier.

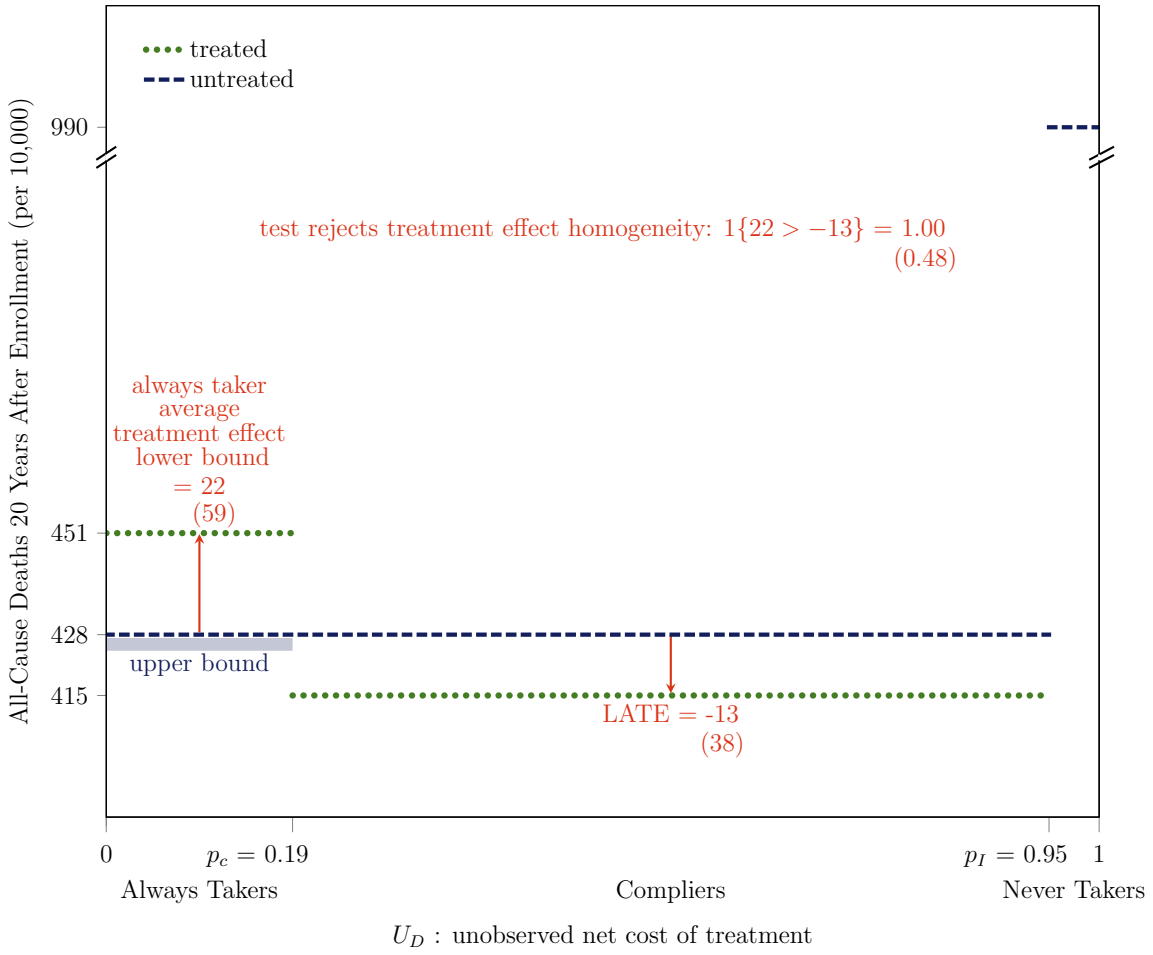
Imposing M.1, I test the null hypothesis of treatment effect homogeneity using the following decision rule, which has an outcome that is equal to 1 if the test rejects treatment effect homogeneity and 0 otherwise:

$$1\left\{\begin{array}{l} \mathrm{E}[Y_T \mid 0 \leq U_D \leq p_C] - \mathrm{E}[Y_U \mid p_C < U_D \leq p_I] > \mathrm{E}[Y_T - Y_U \mid p_C < U_D \leq p_I] \\ \qquad\qquad \text{if } \mathrm{E}[Y_U \mid p_C < U_D \leq p_I] - \mathrm{E}[Y_U \mid p_I < U_D \leq 1] \leq 0, \\ \mathrm{E}[Y_T \mid 0 \leq U_D \leq p_C] - \mathrm{E}[Y_U \mid p_C < U_D \leq p_I] < \mathrm{E}[Y_T - Y_U \mid p_C < U_D \leq p_I] \\ \qquad\qquad \text{if } \mathrm{E}[Y_U \mid p_C < U_D \leq p_I] - \mathrm{E}[Y_U \mid p_I < U_D \leq 1] > 0. \end{array}\right\} \tag{10}$$

As shown, this decision rule has two cases. The first case is the case in which the untreated outcome test statistic is negative, as it is in the CNBSS, and the second case is the case in which the untreated outcome test statistic is positive. In the first case, under M.1, average mortality for untreated compliers is an *upper* bound on the average untreated mortality of always takers, which is not observed. The average treatment effect for always takers is equal to the average treated mortality of always takers minus the average untreated mortality of always takers. Therefore, as illustrated in Figure 4, the upper bound on the average untreated mortality of always takers implies a *lower* bound on the average treatment effect for always takers. The first line of (10) compares this bound to the LATE, the average treatment effect for compliers. If the lower bound on the average treatment effect for always takers is strictly greater than the LATE, then average treatment effect for always takers cannot be equal to the average treatment effect for compliers, so the test rejects treatment effect homogeneity. The logic of the second case follows similarly.

Applying the test to the CNBSS, I reject treatment effect homogeneity. As depicted in

Figure 4: Under Ancillary Assumption, Test Rejects Treatment Effect Homogeneity: Women More Likely to Receive Mammograms are More Likely to be Harmed by Them



Note. Bootstrapped standard errors in parentheses. All-cause deaths are measured 20 years after enrollment for all participants, based on the exact calendar date of enrollment. The treatment is mammography, which is equal to one if a participant receives a mammogram in at least one year during the active study period after the enrollment year. Missing mammogram data in any year is set to no mammogram in that year. The sample includes women aged 40-49 at enrollment, excluding women with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer. Some differences between statistics might not appear internally consistent because of rounding.

Figure 4, I derive a lower bound on the average treatment effect of always takers that indicates that always takers experience at least an additional 22 deaths per 10,000 participants when they receive mammograms. The lower bound on the always taker average treatment effect is strictly greater than the LATE, so always takers face a strictly greater average treatment effect than compliers. Therefore, the decision rule in (10) yields a value of one. The standard error indicates that it is statistically different from zero. Therefore, my finding indicates treatment effect heterogeneity: women more likely to receive mammograms are more likely to be harmed by them.

4.2.1 Support for the Ancillary Assumption Using Baseline Covariates

The ancillary assumption that allows me to find treatment effect heterogeneity in the CNBSS extends my finding of selection heterogeneity, which shows that women more likely to receive mammograms are healthier. My finding of selection heterogeneity is based on the comparison of long-term mortality in the absence of mammograms for compliers and never takers. I do not observe long-term mortality in the absence of mammograms for always takers because, by definition, all always takers received mammograms during the active study period. However, I do observe baseline covariates for always takers, as well as compliers and never takers. I use these baseline covariates as proxies for health in the absence of mammograms, allowing me to investigate support for my ancillary assumption by comparing average baseline covariates across always takers, compliers, and never takers. I obtain average covariates for each group in the same way that I obtain average outcomes for each group, but I combine the averages for treated and untreated compliers, as I discuss at the end of Appendix C.

Table 1: Baseline Summary Statistics for Always Takers, Compliers, and Never Takers

	Means			Difference in Means	
	(1) Always Takers	(2) Compliers	(3) Never Takers	(1)-(2)	(2)-(3)
Baseline Socioeconomic Status					
University, trade or business school	0.50	0.46	0.39	0.04	0.08
	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)
In work force	0.65	0.64	0.65	0.02	-0.02
	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)
Age at first birth	24.28	23.98	23.57	0.30	0.41
	(0.12)	(0.05)	(0.21)	(0.14)	(0.22)
No live birth	0.16	0.15	0.13	0.01	0.01
	(0.01)	(0.00)	(0.01)	(0.01)	(0.02)
Married	0.80	0.81	0.75	-0.01	0.06
	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)
Husband in work force if alive	0.81	0.81	0.76	-0.00	0.05
	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)
Baseline Health Behavior					
Non-Smoker	0.78	0.75	0.63	0.03	0.12
	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)
Body Mass Index	23.87	24.42	24.48	-0.56	-0.06
	(0.10)	(0.05)	(0.21)	(0.12)	(0.22)
Used oral contraception	0.74	0.71	0.67	0.03	0.04
	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)
Used estrogen	0.13	0.13	0.15	-0.00	-0.02
	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)
Mammograms prior to enrollment	0.23	0.13	0.13	0.10	-0.00
	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)
Practiced breast self-examination	0.47	0.44	0.38	0.03	0.06
	(0.01)	(0.00)	(0.02)	(0.01)	(0.02)

Note. Bootstrapped standard errors in parentheses. Missing values correspond to redacted numbers in accordance with Data Use Agreement. The treatment is mammography, which is equal to one if a participant receives a mammogram in at least one year during the active study period after the enrollment year. Missing mammogram data in any year is set to no mammogram in that year. The sample includes women aged 40-49 at enrollment, excluding women with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer. Baseline breast-related covariates are not reported here because they are all zero based on the sample restriction. Some differences between statistics might not appear internally consistent because of rounding.

As shown in Table 1, baseline measures of socioeconomic status tend to vary monotonically from always takers to compliers to never takers, with always takers having the highest socioeconomic status. Since literature has shown a negative correlation between socioeconomic status and health (see Pappas et al. (1993); Cutler and Lleras-Muney (2010); National Center for Health Statistics (2012)), this finding supports my ancillary assumption that women more likely to receive mammograms are healthier. Other covariates that measure health behavior suggest a potential mechanism, demonstrating that women more likely to receive mammograms exhibit better health behaviors. As shown, smoking status, body mass index, and breast self-examination vary monotonically from always takers to compliers to never takers, with always takers exhibiting the best health behaviors.

5 Robustness

I examine the robustness of my two main findings by estimating my main specification with alternative sample restrictions, alternative definitions of mammography, and alternative outcomes. To facilitate comparisons with my main specification, I summarize important statistics from Figures 3 and 4 in tables, starting with Table 2. A specification shows my two main findings — selection and treatment effect heterogeneity such that women more likely to receive mammograms are healthier and more likely to be harmed by them — when the untreated outcome test statistic in column (1) is negative and the decision rule indicates that the test rejects treatment effect homogeneity in column (4).

Table 2: Alternative Sample Restrictions and Alternative Outcome

	N	(1) Untreated Outcome Test	(2) Always Taker Average Treatment Effect Lower Bound	(3) Local Average Treatment Effect LATE	(4) Test Rejects Treatment Effect Homogeneity
Main Specification					
Main specification	19,505	-562	22	-13	1.00
		(147)	(59)	(38)	(0.48)
Alternative Sample Restrictions					
All excluded participants aged 40-49 at enrollment	30,925	-759	60	27	1.00
		(135)	(39)	(40)	(0.47)
All participants aged 40-49 at enrollment	50,430	-672	53	9	1.00
		(103)	(31)	(27)	(0.34)
All participants aged 50-59 at enrollment	39,405	-1,216	-83	15	0.00
		(154)	(51)	(46)	(0.26)
Alternative Outcome					
Breast cancer mortality	19,505	-43	30	-12	1.00
		(47)	(25)	(13)	(0.43)

Note. Bootstrapped standard errors in parentheses. The treatment is mammography, which is equal to one if a participant receives a mammogram in at least one year during the active study period after the enrollment year. Missing mammogram data in any year is set to no mammogram in that year. The sample for the main specification includes women aged 40-49 at enrollment, excluding women with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer. Some differences between statistics might not appear internally consistent because of rounding.

5.1 Alternative Sample Restrictions

5.1.1 Participants Aged 40-49 at Enrollment

In Table 2, I summarize results from specifications that include all *excluded* participants aged 40-49 at enrollment and *all* participants aged 40-49 at enrollment. (The excluded participants are those with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer.) In both samples, the untreated outcome test statistic is negative in column (1) and the test rejects treatment effect homogeneity in column (4). Therefore, the main results are robust.

5.1.2 Participants Aged 50-59 at Enrollment

I focus on women aged 40-49 at enrollment because the change in the 2009 USPSTF recommendations affected them, but I examine the robustness of my results among women aged 50-59 at enrollment. As shown in Table 2, selection heterogeneity goes in the same direction regardless of age group at enrollment: women more likely to receive mammograms are significantly healthier. The lower bound on the average treatment effect for always takers, reported in column (2), does not rule out the LATE, reported in column (3), so I cannot reject treatment effect homogeneity for women aged 50-59 at enrollment. However, the lower bound still allows for treatment effect heterogeneity such that women more likely to receive mammograms are more likely to be harmed by them. To rule out such treatment effect heterogeneity, the specification would have to show that women more likely to receive mammograms are *less* likely to be harmed by them, which would require a positive untreated outcome test statistic in column (1) and a rejection of treatment effect heterogeneity in column (4).

5.2 Alternative Definitions of Mammography

In the CNBSS, I define mammography D such that $D = 1$ if a participant receives a mammogram in at least one year during the active study period after the enrollment year, and I set $D = 0$ otherwise. If mammogram data is missing for a given participant in a given year, I construct D such that the participant did not receive a mammogram in that year. In Table 3, I consider narrower and broader definitions of mammography. Under the narrowest definition, participants must receive a mammogram in *all* active study period years after enrollment to be considered "treated." Under the broadest definition, participants must receive a mammogram *or* be missing mammogram data in any active study period year after enrollment to be considered "treated." The narrowest and broadest definitions are arguably too extreme, so it is notable that all reported specifications yield negative untreated outcome test statistics, demonstrating selection heterogeneity consistent with the main specification:

women more likely to receive mammograms are healthier. Because the untreated outcome test statistic is negative, the two specifications in which the test does not reject treatment effect homogeneity still allow for the possibility that there could be treatment effect heterogeneity such that women more likely to receive mammograms are more likely to be harmed by them.

Table 3: Alternative Definitions of Mammography

	N	(1) Untreated Outcome Test	(2) Always Taker Average Treatment Effect Lower Bound	(3) Local Average Treatment Effect LATE	(4) Test Rejects Treatment Effect Homogeneity
Main Specification					
Mammogram in at least one year after enrollment during the active study period, missing in year = no mammogram in year					
Main specification	19,505	-562	22	-13	1.00
		(147)	(59)	(38)	(0.48)
Narrower Definitions of Mammography					
Mammogram in more than one year after enrollment during the active study period, missing in year = no mammogram in year					
At least two active study period years	19,505	-465	-27	-12	0.00
		(106)	(77)	(35)	(0.49)
At least three active study period years	19,505	-420	56	-12	1.00
		(94)	(145)	(36)	(0.48)
All active study period years	19,505	-225	-135	-15	0.00
		(75)	(138)	(42)	(0.37)
Broader Definition of Mammography					
Mammogram in at least one year after enrollment during the active study period					
Missing in year = mammogram in year	19,505	-776	103	-24	1.00
		(835)	(43)	(69)	(0.43)

Note. Bootstrapped standard errors in parentheses. All-cause deaths are measured 20 years after enrollment for all participants, based on the exact calendar date of enrollment. The treatment is mammography. In the main specification, mammography is equal to one if a participant receives a mammogram in at least one year during the active study period after the enrollment year. Missing mammogram data in any year is set to no mammogram in that year. The sample includes women aged 40-49 at enrollment, excluding women with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer. Some differences between statistics might not appear internally consistent because of rounding.

5.3 Alternative Outcomes

5.3.1 Breast Cancer Mortality

For comparison with the literature, I examine breast cancer mortality as an alternative outcome in lieu of all-cause mortality. Breast cancer mortality could overstate collateral harms from mammography if women who receive mammograms are endogenously more likely to have their cause of death reported as breast cancer. Alternatively, breast cancer mortality could understate collateral harms from mammography if collateral harms from mammograms manifest themselves through causes of death that are not reported as breast cancer. As reported in Table 2, the results in terms of breast cancer mortality corroborate the main results.

5.3.2 Mortality at All Earlier Follow-up Lengths

Table 4: Alternative Outcomes: Mortality at All Earlier Follow-Up Lengths

Years Since Enrollment	N	(1) Untreated Outcome Test	(2) Always Taker Average Treatment Effect Lower Bound	(3) Local Average Treatment Effect LATE	(4) Test Rejects Treatment Effect Homogeneity
Main specification: 20	19,505	-562	22	-13	1.00
		(147)	(59)	(38)	(0.48)
19	19,505	-485	50	-13	1.00
		(142)	(58)	(37)	(0.40)
18	19,505	-492	54	-8	1.00
		(139)	(56)	(35)	(0.41)
17	19,505	-456	18	-8	1.00
		(135)	(50)	(33)	(0.48)
16	19,505	-471	15	-16	1.00
		(134)	(46)	(31)	(0.47)
15	19,505	-480	-11	-15	1.00
		(131)	(42)	(31)	(0.50)
14	19,505	-396	-38	-21	0.00
		(121)	(38)	(30)	(0.45)
13	19,505	-365	-30	-24	0.00
		(115)	(36)	(28)	(0.49)
12	19,505	-334	-23	-27	1.00
		(106)	(32)	(27)	(0.50)
11	19,505	-351	-30	-10	0.00
		(105)	(28)	(25)	(0.42)
10	19,505	-306	-41	-15	0.00
		(97)	(25)	(23)	(0.37)
9	19,505	-314	-35	-12	0.00
		(97)	(21)	(20)	(0.36)
8	19,505	-340	-14	-2	0.00
		(97)	(21)	(18)	(0.44)
7	19,505	-351	-15	-6	0.00
		(97)	(18)	(17)	(0.46)
6	19,505	-317	-24	-5	0.00
		(93)	(16)	(15)	(0.33)
5	19,505	-269	-12	-5	0.00
		(86)	(15)	(13)	(0.45)
4	19,505	-218	-3	-9	1.00
		(77)	(14)	(11)	(0.49)
3	19,505	-209	-3	-6	1.00
		(76)	(11)	(9)	(0.50)
2	19,505	-194	-3	-3	1.00
		(67)	(9)	(9)	(0.50)
1	19,505	-55	-5	-5	0.00
		(40)	(5)	(5)	(0.00)

Note. Bootstrapped standard errors in parentheses. The treatment is mammography, which is equal to one if a participant receives a mammogram in at least one year during the active study period after the enrollment year. Missing mammogram data in any year is set to no mammogram in that year. The sample includes women aged 40-49 at enrollment, excluding women with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer. Some differences between statistics might not appear internally consistent because of rounding.

I investigate the robustness of the main specification, which measures mortality 20 years after enrollment, using specifications that measure mortality at all earlier annual follow-up lengths in Table 4. At all earlier follow-up lengths, the untreated outcome test statistic is negative, consistent with the selection heterogeneity that I find in the main specification: women more likely to receive mammograms are healthier. Furthermore, the test rejects treatment effect homogeneity at some early follow-up lengths and at all follow-up lengths starting 15 years after enrollment, consistent with the treatment effect heterogeneity that I find in the main specification: women more likely to receive mammograms are more likely to be harmed by them. This time pattern suggests that collateral harms from mammograms emerge over time, which makes access to long-term outcomes in the CNBSS particularly valuable.

6 Implications for Guidelines and Future Research

Clinical guidelines are often based on analysis of health outcomes from clinical trials. The success of guidelines in improving health outcomes depends on how they affect behavior in practice. I demonstrate that behavior within a clinical trial can inform how guidelines will affect in behavior in practice, ultimately leading to effects on health outcomes. To do so, I examine relationships between behavior and health outcomes within existing clinical trial data.

Specifically, I examine relationships between mammography behavior and mortality in the CNBSS, an influential and extensive trial on mammography. To the best of my knowledge, the CNBSS trial is the only trial cited by the analysis that informs the USPSTF mammography guidelines that collected data on mammography behavior for all participants. Furthermore, to the best of my knowledge, the CNBSS is the only trial that has tracked mortality for all participants for at least 20 years after enrollment. The long follow-up period proves valuable in capturing collateral harms that manifest themselves over long time horizons.

Within the CNBSS, I identify two key relationships between mammography behavior and mortality. First, under an MTE model that assumes no more than the LATE assumptions, I find that women more likely to receive mammograms are healthier. This relationship reflects heterogeneous selection into mammography. Second, under an ancillary assumption that builds on the first relationship, I find that women more likely to receive mammograms are more likely to experience harm from them. This relationship reflects treatment effect heterogeneity from mammography. Putting both relationships together, women more likely to receive mammograms are healthier and are more likely to experience harm from them.

My first finding, that women more likely to receive mammograms are healthier, is not particularly surprising. The CNBSS allows me to examine a broad measure of health: mor-

tality 20 years after enrollment for women who did not receive mammograms during the active study period. Because I cannot observe this measure for women who received mammograms during the active study period, I impose an ancillary assumption that extends my first finding to them. I find support for the ancillary assumption using using other measures of health that were collected for all trial participants before the trial began. These covariates show that women more likely to receive mammograms during the trial had higher socioeconomic status. Socioeconomic status is often positively correlated with health behavior and health itself. Indeed, within the data collected before the trial began, women more likely to receive mammograms were more likely to be nonsmokers, and they had lower body mass index.

My second finding, that women more likely to receive mammograms are more likely to be harmed by them, is more surprising. However, a growing literature expresses concern about collateral harms from mammograms. Furthermore, there are plausible mechanisms through which women more likely to receive mammograms could be more likely to be harmed by them. For example, it is plausible that women more likely to receive mammograms detect less severe breast cancers but pursue similar aggressive treatments. It is also plausible that women more likely to receive mammograms detect breast cancers of the same severity but pursue more aggressive treatments.

Given my findings, the current USPSTF mammography guidelines for women in their 40s could have unintended consequences. The current guidelines leave the mammography decision for women in their 40s to individual women and the doctors. In doing so, my findings imply that they unintentionally encourage more mammograms for healthier women who are more likely to be harmed by them.

The active study period of the CNBSS took place in the 1980s, so it is unclear if my findings apply in the current environment. However, changes in environment are an inherent limitation of any long-term analysis. The current USPSTF guidelines are based on long-term health outcomes from the CNBSS and other trials. Given that findings based on health outcomes from the CNBSS are already taken into account in the determination of the current USPSTF guidelines, it is arguably worth considering my findings based on behavior. My first finding, which shows that women more likely to receive mammograms are healthier, still seems applicable in the current environment. My second finding, which shows that women more likely to receive mammograms are more likely to be harmed by them, could be attenuated or exacerbated in the current environment. Treatments for breast cancer are becoming less aggressive, which could attenuate my second finding in the current environment. On the other hand, mammograms are becoming more detailed, potentially leading to the diagnosis and treatment of more breast cancers that would never grow to be life-threatening, which

could exacerbate my second finding in the current environment.

My findings support the need for more clinical trials on mammograms. Mammograms are now so ubiquitous that it could be difficult to conduct such trials. Indeed, mammography was already so widespread in the U.S. in the 1980s that most of the trials considered by the USPSTF in its recommendations were conducted outside of the U.S. Even in the Canadian National Breast Screening Study, a substantial fraction of women in the control group received mammograms. In the present environment, many factors encourage mammography, including mandatory health insurance coverage for mammograms under the Affordable Care Act, public outreach efforts, and risk aversion on the part of doctors and patients. Very few factors discourage mammography or encourage more evidence to be collected on it, which is potentially a reason to take my findings even more seriously.

Beyond the context of mammograms, my findings support the need for clinical trials to collect data on behavior and for that data to be used in the development of guidelines. In many trials, individual-level data on takeup of treatment are not collected, especially for participants assigned to the control group. Furthermore, even when they are collected, to the best of my knowledge, they are not taken into account in the development of guidelines. Whenever the USPSTF determines that "there is at least moderate certainty that the net benefit is small," the USPSTF issues a "C recommendation," as it did in the case of mammography for women in their 40s, which means that "the USPSTF recommends selectively offering this service to individual patients based on professional judgment and patient preferences" U.S. Preventive Service Task Force (2017). Such a guideline presupposes that there is selection and treatment effect heterogeneity such that the individuals most likely to benefit from a treatment will be the most likely to receive it. However, such a guideline is not based on evidence of selection and treatment effect heterogeneity. By demonstrating that it is possible to examine selection and treatment effect heterogeneity within existing clinical trial data, I advance the ability of future guidelines to progress toward personalized health care.

Appendix

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

The uniform distribution of U_D between 0 and 1 is not a separate assumption of the model. Instead, it is due to the "probability integral transformation," which shows that the cumulative distribution function of any random variable applied to itself must be distributed uniformly between 0 and 1 (for example, see Casella and Berger (2002, page 54)). 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 proof that follows shows that $U_D = F(\nu_D)$ is distributed uniformly between

0 and 1.

$$
\begin{aligned}
F_{U_D}(u) &= P(U_D \leq u) \\
&= P(F(\nu_D) \leq u) \\
&= P(\nu_D \leq F^{-1}(u)) && (F(\cdot) \text{ absolutely continuous by A.2}) \\
&= F(F^{-1}(u)) = u.
\end{aligned}
$$

■

Appendix B Derivation of the treatment equation

Treatment D is given by

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

where the last equality follows from

$$
\begin{aligned}
F(\mu_D(Z)) &= \mathrm{P}(\nu_D \leq \mu_D(Z)) \\
&= \mathrm{P}(\nu_D \leq \mu_D(z) \mid Z = z) && (\nu_D \perp Z \text{ by A.3}) \\
&= \mathrm{P}(0 \leq \mu_D(Z) - \nu_D \mid Z = z) \\
&= \mathrm{P}(0 \leq V_T - V_U \mid Z = z) \\
&= \mathrm{P}(D = 1 \mid Z = z).
\end{aligned}
$$

■

Appendix C Derivation of average outcomes and covariates

I begin by deriving the average treated outcome of always takers. To do so, I begin with the average outcome of treated control group participants because those participants must be always takers. After some manipulation, I invoke the independence assumption A.3 to show that the average outcome of treated control group participants must be equal to the average treated outcome of all always takers, regardless of whether they are assigned to the control

or intervention group:

$$\begin{aligned}
\mathrm{E}[Y \mid D=1, Z=0] &= \mathrm{E}[Y_U + D(Y_T - Y_U) \mid D=1, Z=0] && \text{(by (6))} \\
&= \mathrm{E}[Y_T \mid D=1, Z=0] \\
&= \mathrm{E}[Y_T \mid 0 \leq U_D \leq p_C, Z=0] && \text{(by (4), where } p_C = P(D=1|Z=0)) \\
&= \mathrm{E}[g_T(U_D, \gamma_T) \mid 0 \leq U_D \leq p_C, Z=0] && \text{(by (7))} \\
&= \mathrm{E}[g_T(U_D, \gamma_T) \mid 0 \leq U_D \leq p_C] && (Z \perp (U_D, \gamma_T) \text{ by A.3}) \\
&= \mathrm{E}[Y_T \mid 0 \leq U_D \leq p_C].
\end{aligned}$$

In the CNBSS, I obtain an average treated outcome of always takers of 451 deaths per 10,000 women, which I plot over the relevant range ($0 \leq U_D \leq p_I$) in Figure C1, using a dotted line to indicate that it represents an average treated outcome.

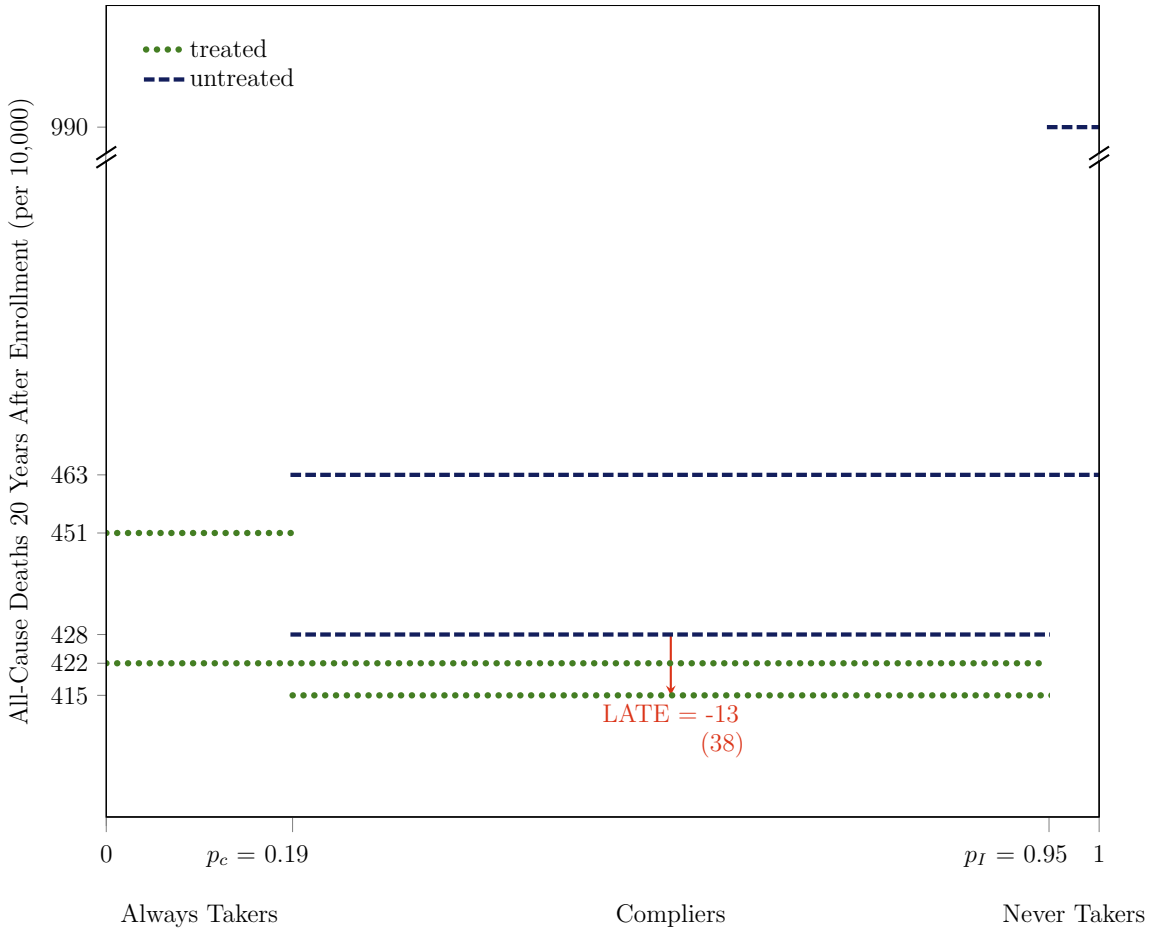
Next, I derive the average treated outcome of compliers. To do so, I begin with the average outcome of treated intervention group participants because those participants must be always takers and treated compliers. Following a similar derivation to the derivation for always takers yields $\mathrm{E}[Y \mid D=1, Z=1] = \mathrm{E}[Y_T \mid 0 \leq U_D \leq p_I]$. Therefore, I plot the average outcome of treated intervention group participants, which is 422 deaths per 10,000 women in the CNBSS, over the relevant range ($0 \leq U_D \leq p_I$) in Figure C1, using a dotted line to indicate that it represents an average treated outcome. As the figure makes clear, the fractions of always takers and compliers are known, so it is possible to back out the average treated outcome of compliers as follows:

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

In the CNBSS, the average treated outcome of compliers is 415 deaths per 10,000 women, which I plot over the relevant range ($p_C < U_D \leq p_I$) in Figure C1, using a dotted line to indicate that it represents an average treated outcome.

Turning to average untreated outcomes, I begin with the average untreated outcome of intervention group participants because those participants must be never takers. Following a similar derivation to the derivation for always takers yields $\mathrm{E}[Y \mid D=0, Z=1] = \mathrm{E}[Y_U \mid p_I < U_D \leq 1]$. In the CNBSS, I obtain an average untreated outcome of never takers of 990 deaths per 10,000 women, which I plot over the relevant range ($p_I < U_D \leq 1$) in Figure C1, using a dashed line to indicate that it represents an average untreated outcome. Similarly, I derive the average outcome of untreated control group participants $\mathrm{E}[Y \mid D=0, Z=0] = \mathrm{E}[Y_U \mid p_C < U_D \leq 1]$, which is equal to 463 deaths per 10,000 women, and I plot it over the

Figure C1: Derivation of Average Mortality for Always Takers, Compliers, and Never Takers



Note. Bootstrapped standard errors in parentheses. All-cause deaths are measured 20 years after enrollment for all participants, based on the exact calendar date of enrollment. The treatment is mammography, which is equal to one if a participant receives a mammogram in at least one year during the active study period after the enrollment year. Missing mammogram data in any year is set to no mammogram in that year. The sample includes women aged 40-49 at enrollment, excluding women with any nonzero values of the following breast-related covariates at baseline: breast cancer in family; any other breast disease; patient reported symptoms; referred for review by nurse; abnormality found by nurse; ever told has breast cancer.

relevant range ($p_C < U_D \leq 1$) in Figure C1, using a dashed line to indicate that it represents an average untreated outcome. Using these two values, I calculate the average untreated outcome of compliers as follows:

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

In the CNBSS, the average untreated outcome of compliers is 428 deaths per 10,000 women, which I plot over the relevant range ($p_C < U_D \leq p_I$) in Figure C1, using a dashed line to indicate that it represents an average untreated outcome.

To derive the average covariates for always takers, compliers, and never takers, I follow the same approach with a covariate X in lieu of an outcome Y. However, while average outcomes should be different for treated and untreated compliers, average covariates should be the same for treated and untreated compliers. I therefore obtain the average covariate vector for compliers weighting the average covariate vectors for treated and untreated compliers by the probabilities of being assigned to the intervention and control groups, respectively:

$$\mathrm{E}[X \mid p_C < U_D \leq p_I] = \mathrm{P}(Z=1)\mathrm{E}[X \mid p_C < U_D \leq p_I] + \mathrm{P}(Z=0)\mathrm{E}[X \mid p_C < U_D \leq p_I].$$

Appendix D Selection bias

The definition of "selection bias" used by Angrist (1998) and Heckman et al. (1998), among others, is as follows:

$$\mathrm{E}[Y_U \mid D=1] - \mathrm{E}[Y_U \mid D=0]$$

I can express selection bias as the following weighted integral of the MUO function, demonstrating that it is a special case of selection heterogeneity as defined by the MUO function with weights $\omega(p, p_L, p_H) = 1\{p_L \leq p < p_H\}/(p_H - p_L)$:

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

This weighted integral depends on the probability of assignment to the intervention group $\mathrm{P}(Z=1)$, which is a feature of the trial design.

I can also express selection bias as a weighted average of the untreated outcomes of always takers ($0 \leq U_D \leq p_C$), compliers ($p_C < U_D \leq p_I$), and never takers ($p_I < U_D \leq 1$):

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

While all other terms are identified without ancillary assumptions, the average untreated outcome of always takers $\mathrm{E}[Y_U \mid 0 \leq U_D \leq p_C]$ is not, so selection bias is not identified without ancillary assumptions.

Appendix E Alternative ancillary weak monotonicity assumptions

Following Brinch et al. (2017), I impose the following ancillary assumption on the MUO function:

M.1. (Weak Monotonicity of the MUO function) For all p_1, $p_2 \in [0,1]$ such that $p_1 < p_2$: $\mathrm{E}[Y_U \mid U_D = p_1] \leq \mathrm{E}[Y_U \mid U_D = p_2]$ or $[Y_U \mid U_D = p_1] \geq \mathrm{E}[Y_U \mid U_D = p_2]$.

Brinch et al. (2017) also impose the following analogous assumption on the MTO function:

M.2. (Weak Monotonicity of the MTO function) For all p_1, $p_2 \in [0,1]$ such that $p_1 < p_2$: $\mathrm{E}[Y_T \mid U_D = p_1] \leq \mathrm{E}[Y_T \mid U_D = p_2]$ or $[Y_T \mid U_D = p_1] \geq \mathrm{E}[Y_T \mid U_D = p_2]$.

I also consider imposing the following analogous assumption on the MTE function, which is similar to the Manski (1997) assumption of monotone treatment response:

M.3. (Weak Monotonicity of the MTE function) For all p_1, $p_2 \in [0,1]$ such that $p_1 < p_2$: $\mathrm{E}[Y_T - Y_U \mid U_D = p_1] \leq \mathrm{E}[Y_T - Y_U \mid U_D = p_2]$ or $[Y_T - Y_U \mid U_D = p_1] \geq \mathrm{E}[Y_T - Y_U \mid U_D = p_2]$.

I emphasize that just as M.1 implies an upper or lower bound on the average treatment effect for always takers, M.2 implies an upper or lower bound on the average treatment effect for never takers. Because either assumption yields a bound on the average treatment effect for a group of participants other than compliers, which can be compared to the LATE, either assumption facilitates a test of treatment effect homogeneity.

However, I do not impose M.2 in the CNBSS because it is harder to defend than M.1. As I discuss in Section 4.2.1, I can proxy for untreated outcomes using baseline covariates, and doing so provides support for M.1 in the CNBSS. Because variation in treated outcomes can reflect selection and treatment effect heterogeneity, I cannot think of a proxy for them, which makes M.2 harder to defend.

I also do not impose M.3 in the CNBSS because it it is harder to defend than M.1. I could imagine using measures of breast cancer treatment intensity as proxies for treatment effects, where greater treatment intensity should lead to greater treatment effects. However, such measures are not available in the CNBSS data because women diagnosed with breast cancer in CNBSS study centers received breast cancer treatment elsewhere.

References

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

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

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

Angrist, J. D. and G. W. Imbens (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

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

Baines, C. J., T. To, and A. B. Miller (2016). Revised estimates of overdiagnosis from the canadian national breast screening study. *Preventive medicine 90*, 66–71.

Baum, M. (2013). Harms from breast cancer screening outweigh benefits if death caused by treatment is included. *Bmj 346*(jan23 1).

Bertanha, M. and G. W. Imbens (2014, December). External validity in fuzzy regression discontinuity designs. Working Paper 20773, National Bureau of Economic Research.

Bitler, M. P. and C. S. Carpenter (2016). Health insurance mandates, mammography, and breast cancer diagnoses. *American Economic Journal: Economic Policy 8*(3), 39–68.

Björklund, A. and R. Moffitt (1987). The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, 42–49.

Bjurstam, N., L. Björneld, J. Warwick, E. Sala, S. W. Duffy, L. Nyström, N. Walker, E. Cahlin, O. Eriksson, L.-O. Hafström, et al. (2003). The gothenburg breast screening trial. *Cancer 97*(10), 2387–2396.

Black, D. A., J. Joo, R. LaLonde, J. A. Smith, and E. J. Taylor (2015). Simple tests for selection bias: Learning more from instrumental variables.

Bleyer, A. and H. G. Welch (2012). Effect of three decades of screening mammography on breast-cancer incidence. *New England Journal of Medicine 367*(21), 1998–2005.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017). Beyond late with a discrete instrument. *Journal of Political Economy 125*(4), 000–000.

Buchmueller, T. C. and L. Goldzahl (2018). The effect of organized breast cancer screening on mammography use: Evidence from france. Technical report, National Bureau of Economic Research.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating marginal returns to education. *The American economic review 101*(6), 2754–2781.

Carneiro, P. and S. Lee (2009). 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.

Casella, G. and R. L. Berger (2002). *Statistical inference*, Volume 2. Duxbury Pacific Grove, CA.

Cutler, D. M. and A. Lleras-Muney (2010). Understanding differences in health behaviors by education. *Journal of health economics 29*(1), 1–28.

Guo, Z., J. Cheng, S. A. Lorch, and D. S. Small (2014). Using an instrumental variable to test for unmeasured confounding. *Statistics in medicine 33*(20), 3528–3546.

Habbema, J., G. J. v. Oortmarssen, D. J. van Putten, J. T. Lubbe, and P. J. v. d. Maas (1986). Age-specific reduction in breast cancer mortality by screening: an analysis of the results of the health insurance plan of greater new york study. *Journal of the National Cancer Institute 77*(2), 317–320.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–162.

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998). Characterizing selection bias using experimental data. *Econometrica 66*(5), 1017–1098.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*(3), 669–738.

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

Heckman, J. J. and E. J. Vytlacil (2001). Local instrumental variables. In C. Hsiao, K. Morimune, and J. L. Powell (Eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, pp. 1–46. Cambridge University Press.

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

Kadiyala, S. and E. Strumpf (2016). How effective is population-based cancer screening? regression discontinuity estimates from the us guideline screening initiation ages. In *Forum for Health Economics and Policy*, Volume 19, pp. 87–139. De Gruyter.

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

Kim, H. B. and S. Lee (2017). When public health intervention is not successful: Cost sharing, crowd-out, and selection in korea's national cancer screening program. *Journal of Health Economics 53*, 100 – 116.

Kline, P. M. and C. R. Walters (2018, March). On heckits, late, and numerical equivalence. Working Paper XX, National Bureau of Economic Research.

Kowalski, A. (2016, June). 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.

Kowalski, A. (2018a, May). Extrapolation using selection and moral hazard heterogeneity from within the oregon health insurance experiment. Working Paper 24647, National Bureau of Economic Research.

Kowalski, A. (2018b, July). How to examine external validity within an experiment. Working Paper 24834, National Bureau of Economic Research.

Kowalski, A., Y. Tran, and L. Ristovska (2016, December). MTEBINARY: Stata module to compute Marginal Treatment Effects (MTE) With a Binary Instrument. Statistical Software Components, Boston College Department of Economics.

Kowalski, A., Y. Tran, and L. Ristovska (2018, July). MTEMORE: Stata module to compute Marginal Treatment Effects (MTE) With a Binary Instrument. Statistical Software Components, Boston College Department of Economics.

Lannin, D. R. and S. Wang (2017). Are small breast cancers good because they are small or small because they are good? *New England Journal of Medicine 376*(23), 2286–91.

Manski, C. F. (1997). Monotone treatment response. *Econometrica: Journal of the Econometric Society*, 1311–1334.

Miller, A. B., C. J. Baines, T. To, and C. Wall (1992a). Canadian national breast screening study: 1. breast cancer detection and death rates among women aged 40 to 49 years. *CMAJ: Canadian Medical Association Journal 147*(10), 1459–1476.

Miller, A. B., C. J. Baines, T. To, and C. Wall (1992b). Canadian national breast screening study: 2. breast cancer detection and death rates among women aged 50 to 59 years. *CMAJ: Canadian Medical Association Journal 147*(10), 1477–1488.

Miller, A. B., T. To, C. J. Baines, and C. Wall (1997). The canadian national breast screening study: update on breast cancer mortality. *JNCI Monographs 1997*(22), 37–41.

Miller, A. B., T. To, C. J. Baines, and C. Wall (2000). Canadian national breast screening study-2: 13-year results of a randomized trial in women aged 50–59 years. *Journal of the National Cancer Institute 92*(18), 1490–1499.

Miller, A. B., T. To, C. J. Baines, and C. Wall (2002). The canadian national breast screening study-1: breast cancer mortality after 11 to 16 years of follow-up: a randomized screening trial of mammography in women age 40 to 49 years. *Annals of internal medicine 137*(5_Part_1), 305–312.

Miller, A. B., C. Wall, C. J. Baines, P. Sun, T. To, and S. A. Narod (2014). Twenty five year follow-up for breast cancer incidence and mortality of the canadian national breast screening study: randomised screening trial. *Bmj 348*, g366.

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

Moss, S. M., C. Wale, R. Smith, A. Evans, H. Cuckle, and S. W. Duffy (2015). Effect of mammographic screening from age 40 years on breast cancer mortality in the uk age trial at 17 years' follow-up: a randomised controlled trial. *The Lancet Oncology 16*(9), 1123–1132.

National Center for Health Statistics (2012). Health, United States, 2011: With special feature on socioeconomic status and health.

Nelson, H. D., R. Fu, A. Cantor, M. Pappas, M. Daeges, and L. Humphrey (2016). Effectiveness of breast cancer screening: Systematic review and meta-analysis to update the 2009 us preventive services task force recommendation effectiveness of breast cancer screening. *Annals of internal medicine 164*(4), 244–255.

Nelson, H. D., M. Pappas, A. Cantor, J. Griffin, M. Daeges, and L. Humphrey (2016). Harms of breast cancer screening: systematic review to update the 2009 us preventive services task force recommendationharms of breast cancer screening. *Annals of internal medicine 164*(4), 256–267.

Nyström, L., I. Andersson, N. Bjurstam, J. Frisell, B. Nordenskjöld, and L. E. Rutqvist (2002). Long-term effects of mammography screening: updated overview of the swedish randomised trials. *The Lancet 359*(9310), 909–919.

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

Pappas, G., S. Queen, W. Hadden, and G. Fisher (1993). The increasing disparity in mortality between socioeconomic groups in the united states, 1960 and 1986. *New England journal of medicine 329*(2), 103–109.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford economic papers 3*(2), 135–146.

Siu, A. L. (2016). Screening for breast cancer: Us preventive services task force recommendation statementscreening for breast cancer. *Annals of internal medicine 164*(4), 279–296.

Tabar, L., G. Fagerberg, H.-H. Chen, S. W. Duffy, C. R. Smart, A. Gad, and R. A. Smith (1995). Efficacy of breast cancer screening by age. new results swedish two-county trial. *Cancer 75*(10), 2507–2517.

U.S. Preventive Service Task Force (2017). Grade definitions. u.s. preventive services task force. november 2017. https://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions. Online. Accessed June 8, 2018.

U.S. Preventive Services Task Force (2002). Screening for breast cancer: Recommendations and rationale. *Annals of Internal Medicine 137*(5), 344–346.

U.S. Preventive Services Task Force (2009). Screening for breast cancer: U.s. preventive services task force recommendation statement. *Annals of Internal Medicine 151*(10), 716–726.

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

Zanella, G. and R. Banerjee (2016). Experiencing breast cancer at the workplace. *Journal of Public Economics 134*, 53–66.