NBER WORKING PAPER SERIES

BEHAVIOR WITHIN A CLINICAL TRIAL AND IMPLICATIONS FOR MAMMOGRAPHY GUIDELINES

Amanda E. Kowalski

Working Paper 25049
http://www.nber.org/papers/w25049

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2018, Revised November 2019

Previously circulated as "Biology Meets Behavior in a Clinical Trial: Two Relationships Between Mortality and Mammogram Receipt." The first working paper version of this paper was circulated in September 2018 as Kowalski (2018b). This paper includes content from Kowalski (2016) related to weak monotonicity assumptions. Other content from Kowalski (2016) appears in Kowalski (2018a). I co-developed the Stata commands mtemore to accompany Kowalski (2016) and mtebinary to accompany Kowalski (2018a). Saumya Chatrath, Neil Christy, Tory Do, Simon Essig Aberg, 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 Zach Brown, Emily Horton, Pat Kline, Lee Lockwood, Magne Mogstad, Brock Rowberry, Atheendar Venkataramani, graduate public finance students at the University of Michigan, and seminar participants at the 2018 American Economic Association Annual Meeting, ASHEcon 2019, the 2018 Canadian Health Economics Study Group, Columbia University, the 2018 London-Paris Public Economics Workshop, the NBER Health Care 2018 Fall Meeting, the NBER Aging 2019 Spring Meeting, the 2018 North American Summer Meetings of the Econometric Society, Indiana University, 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. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Behavior within a Clinical Trial and Implications for Mammography Guidelines
Amanda E. Kowalski
NBER Working Paper No. 25049
September 2018, Revised November 2019
JEL No. C18,I1,I12

ABSTRACT

Current mammography guidelines reflect evidence that mammography could be harmful on average through the overdiagnosis of breast cancers that would not eventually cause symptoms in the long term. To inform targeting within these guidelines, I investigate whether some women are more likely to experience overdiagnosis than others on the basis of their mammography behavior. Using data on mammography behavior within an influential clinical trial, random assignment, and a model, I proceed in two steps. First, I find that women who are more likely to receive mammograms are healthier and have higher socioeconomic status. Second, building on the first finding, I find that the 20-year level of overdiagnosis is at least 3.5 times higher among women who are more likely to receive mammograms, such that at least 36% of their cancers are overdiagnosed. Current guidelines presuppose that the women most likely to receive mammograms are the women most likely to benefit from them. My findings imply that these guidelines could have unintended consequences by effectively encouraging mammograms among healthier women who are more likely to be overdiagnosed by them.

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

1 Introduction

The U.S. Preventive Services Task Force (USPSTF) updated their mammography guidelines in 2009 (U.S. Preventive Services Task Force, 2009) in response to evolving evidence from clinical trials. 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 the following questions, which motivate my analysis: Do current guidelines effectively target mammograms to women most likely to benefit from them, and can behavior within a clinical trial help to inform targeting within guidelines? I aim to answer these questions using data from a clinical trial. I proceed in two steps. First, I investigate heterogeneous selection: are women who are more likely to receive mammograms different from other women? Second, I investigate treatment effect heterogeneity: are women who are more likely to receive mammograms more likely to experience better or worse health outcomes because of them?

Mammography can lead to better health outcomes through the early detection and treatment of breast cancer that would eventually grow to be life-threatening, but mammography can also lead to worse health outcomes through the early detection and treatment of breast cancer that would *not* eventually grow to be life-threatening. 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). Overdiagnosis is distinct from a false-positive diagnosis. The latter refers to "a positive test in an individual who is subsequently recognized not to have cancer. By contrast, an overdiagnosed patient has a tumor that fulfills the pathological criteria for cancer"(Welch and Black, 2010). An extensive literature considers the possibility of overdiagnosis.[1]

Overdiagnosis can pose significant health risks. It can expose women to unnecessary chemotherapy, radiotherapy, and surgery, which can all be life-threatening.[2] Even absent subsequent medical care, breast cancer diagnosis itself can be harmful. Providing a perspective in the *New England Journal of Medicine*, Welch and Fisher (2017) argue that "the psycholocial effects of overutilization

[1] Etzioni et al. (2002); Pohl and Welch (2005); Zackrisson et al. (2006); Jørgensen and Gøtzsche (2009); Bleyer and Welch (2012); Marmot et al. (2012); Baum (2013); Duffy and Parmar (2013); Biller-Andorno and Jüni (2014); Helvie et al. (2014); Miller et al. (2014); Patz et al. (2014); Welch and Passow (2014); Harding et al. (2015); Baines et al. (2016); McCaffery et al. (2016); Nelson et al. (2016); Welch et al. (2016); Jørgensen et al. (2017); Lannin and Wang (2017); Raffle and Gray (2019)

[2] Evidence shows that chemotherapy for early-stage breast cancers increases the risk of second cancers, such as chemotherapy-induced acute myeloid leukaemia (Aidan et al., 2013; Martin et al., 2009; Praga et al., 2005). Radiotherapy for breast cancer has been shown to significantly increase mortality from lung cancer and heart disease (Early Breast Cancer Trialists' Collaborative Group, 2005). Surgeries such as mastectomy and lumpectomy also pose risks.

and overdiagnosis are also worrisome: turning people into patients may undermine their sense of resilience, which is fundamental to health."

It is possible that overdiagnosis is so harmful that the harms of mammograms outweigh the benefits on average. The 2016 USPSTF guidelines are based on a meta-analysis (Nelson et al., 2016) of average impacts from clinical trials.[3] Combining results across trials, there is no statistically significant reduction in all-cause mortality across all age groups or within any age group. Furthermore, some trials show statistically insignificant increases in all-cause mortality within some age groups, suggesting that the harms can outweigh the benefits on average (Nyström et al., 2002; Miller et al., 2014).

Instead of focusing on average health impacts, I advance the literature by examining whether health impacts vary with mammography behavior within a clinical trial. The meta-analysis that informs the 2016 USPSTF mammography guidelines (Nelson et al., 2016) examines health outcomes within clinical trials, but it says little about mammography behavior within those trials. Outside of the clinical trial literature, a large literature examines mammography behavior in response to policy interventions that yield natural experiments, but it says little about how health impacts vary with such behavior.[4] This literature has been limited because the methods that it employs do not allow it to recover how health impacts vary with mammography behavior. Furthermore, it rarely engages with the possibility of overdiagnosis as a health impact, perhaps because individual-level data on mammography behavior that follow individuals in a randomized or natural experiment for long enough to identify overdiagnosis are not widely available.

I examine how health impacts vary with mammography behavior in the Canadian National Breast Screening Study (CNBSS), a trial influential to the USPSTF guidelines. The CNBSS enrolled almost 90,000 participants aged 40-59 between 1980 and 1985. All participants were randomly assigned to an intervention group or a control group. Intervention group participants received access to annual mammograms and clinical breast examinations 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 an initial clinical breast examination followed by usual care in the community, and control group women in their 50s at enrollment received access to annual clinical breast examinations in the initial year and 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 health outcomes for all participants. To the best of my knowledge, the CNBSS is the only trial considered by the meta-analysis that informs the USPSTF guidelines (Nelson et al., 2016) that tracked the actual takeup of mammograms for all participants. These data show that a substantial fraction of women in

[3] Habbema et al. (1986); Tabar et al. (1995); Nyström et al. (2002); Bjurstam et al. (2003); Miller et al. (2014); Moss et al. (2015)

[4] Kelaher and Stellman (2000); Habermann et al. (2007); Kadiyala and Strumpf (2011, 2016); Finkelstein et al. (2012); Kolstad and Kowalski (2012); Bitler and Carpenter (2016, 2019); Fedewa et al. (2015); Mehta et al. (2015); Ong and Mandl (2015); Lu and Slusky (2016); Zanella and Banerjee (2016); Cooper et al. (2017); Jacobson and Kadiyala (2017); Kim and Lee (2017); Buchmueller and Goldzahl (2018); Einav et al. (2019); Myerson et al. (2019)

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, I can observe two important long-term health outcomes—breast cancer incidence and all-cause mortality—for all participants through linkage to cancer registries that are complete across Canada (Baines et al., 2016) and the Canadian Mortality Database. The CNBSS is the only trial considered by the meta-analysis that informs the USPSTF guidelines that allows for examination of health outcomes for at least 20 years after enrollment for all participants.

The ability to examine long-term health outcomes proves important to my results. In the short term, breast cancer incidence should be larger in the intervention group than it is in the control group because mammograms diagnose breast cancer. In the long term, however, breast cancer incidence should completely "catch up" in the control group if mammography only leads to earlier diagnosis. In the CNBSS, breast cancer incidence remains higher in the intervention group 25 years after the first participants enrolled (Baines et al., 2016), and the difference is statistically significant. This persistent difference is particularly striking given that mammography behavior likely converged between the control and intervention groups after the active study period. Taking as given that breast cancers that led to death at any follow-up length were diagnosed and are thus available in the data, the persistently higher rate of breast cancer in the intervention group indicates overdiagnosis.

Because I am interested in whether there is heterogeneity in overdiagnosis with mammography behavior, I use a heterogeneous 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), Brinch et al. (2017), and Cornelissen et al. (2018). I present the model using simple figures. I view these figures as a contribution to the literature. They motivate and depict where the MTE model that I use allows me to identify treatment effect heterogeneity using a single ancillary assumption beyond the local average treatment effect (LATE) assumptions of Imbens and Angrist (1994). They also make clear that the terminology of "always takers," "compliers," and "never takers" from the LATE literature (Angrist et al., 1996) separates individuals into three groups on the basis of how likely they are to receive the treatment.

First, I find heterogeneous selection: women more likely to receive mammograms are healthier in terms of long-term breast cancer incidence and mortality. They also have higher socioeconomic status and are more likely to practice several other health behaviors seen as beneficial. They are more likely to be nonsmokers, and they have lower body mass index. Furthermore, findings from natural experiments corroborate my first finding in other contexts.

Second, I find treatment effect heterogeneity: the level of breast cancer overdiagnosis is at least 3.5 times higher among women more likely to receive mammograms such that at least 36% of their cancers are overdiagnosed. I also find suggestive evidence that women more likely to receive mammograms are more likely to be harmed by them in terms of long-term mortality. Though the treatment effects on 20-year mortality are not statistically different from zero or from each other, the implied lower bound on the treatment effect among women more likely to receive mammograms

is large, such that at least 4.9% of deaths among such women would not have occurred otherwise. I am not aware of any other research that has sought to estimate whether treatment effects on breast cancer incidence or mortality vary with mammography behavior. Even if it did seek to do so, the ability of other research to do so would be limited by available follow-up data.

My second finding may be counterintuitive, because we might expect that women will only receive mammograms if mammograms will improve their health outcomes. However, given the available evidence, it is difficult for women and their doctors to predict the individual-level impacts of mammography. Furthermore, women and their doctors might make mammography decisions based on factors beyond impacts on breast cancer diagnosis and mortality. One such factor, articulated by the personal health columnist from the *New York Times* (Brody, 2017), is that "Doing something is often more appealing than doing nothing. Many who think this way consider only the beneficial 'what if's' and not the possible downsides of cancer screening tests." Such logic could also extend to procedures pursued after a breast cancer diagnosis. Although virtually all women with breast cancer during the active study period had at least part of a breast removed, I find suggestive evidence that women more likely to receive mammograms were more likely to have an entire breast removed, providing a potential mechanism for increased mortality in response to overdiagnosis.

Given my first finding that women more likely to receive mammograms have higher socioeconomic status, Welch and Fisher (2017) provide a rationale for my second finding that women more likely to receive mammograms are more likely to be overdiagnosed by them. They hypothesize that women of higher socioeconomic status are exposed to increased "observational intensity" such that "they are likely to be screened more often and by means of such tests...that can detect smaller abnormalities, undergo more follow-up testing, and undergo more biopsies, and they may be served by health systems that have a lower threshold for labeling results as abnormal." It is well-known that women of higher socioeconomic status have higher rates of breast cancer (Hakama et al., 1982; Robert et al., 2004; Reynolds et al., 2005; Brown et al., 2007), even though they are generally healthier in terms of other health outcomes (Pappas et al., 1993; Cutler and Lleras-Muney, 2010; National Center for Health Statistics, 2012). My findings provide a potential mechanism for the empirical relationship between socioeconomic status and breast cancer, whereby increased "observational intensity" among women of higher socioeconomic status leads to greater overdiagnosis.

Knowledge about individual-level breast cancer risk could provide an alternative rationale for my second finding that women more likely to receive mammograms are more likely to be overdiagnosed by them. Women more likely to receive mammograms could possess knowledge that they have a higher risk of breast cancer, despite my first finding that these women are healthier on several other dimensions. These women could therefore be more likely to be diagnosed with breast cancer. As one response to this concern, I take a conservative approach to sample selection in my main analysis sample. The CNBSS conducted extensive baseline surveys and clinical exams. I use variables collected through these means to exclude individuals with a family history of breast cancer as well as other individuals with potential knowledge of increased breast cancer risk. I also examine characteristics of the breast cancers detected during the active study period. I find suggestive evidence that the breast cancers detected in women more likely to receive mammograms

are smaller and less invasive, which is consistent with that idea that women more likely to get mammograms might not actually be sicker; instead, they might be members of the "worried well."

The combination of my first and second findings implies that the current mammography guidelines conflate the women most likely to receive mammograms with the women most likely to benefit from them. These guidelines could therefore have unintended consequences by leading to greater overdiagnosis of breast cancers in healthier women. To mitigate unintended consequences, the guidelines could be further weakened to no longer recommend mammography for any asymptomatic women in their 40s. Such a change would not affect the behavior of women who do not receive mammograms under the current guidelines, but it could curtail overdiagnosis for some women who receive mammograms under the current guidelines.

In the next section, I replicate published results from the CNBSS. In Section 3, I present the model. In Section 4, I arrive at my two main findings. I show that my findings are robust to a wide variety of alternative specifications in Section 5. In Section 6, I discuss the applicability of results from the CNBSS to the current medical environment. I conclude by discussing implications for guidelines and future research in Section 7.

2 Replication of CNBSS Results

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 how the results vary with mammography behavior. The latest results published by the CNBSS investigators show a higher rate of breast cancer in the intervention group in the long term, consistent with overdiagnosis (Baines et al., 2016), and the difference is statistically significant. The rate of all-cause mortality is also higher in the intervention group in the long term (Miller et al., 2014), but the difference is not statistically significant. The rate of breast cancer mortality is slightly lower in the intervention group in the long term (Miller et al., 2014), but the difference is not statistically significant either. In terms of statistical significance, the long-term results are consistent with results published at earlier follow-up lengths (Miller et al., 1992a,b, 1997, 2000, 2002, 2014).

I can exactly replicate the published long-term all-cause and breast cancer mortality results, and I can closely replicate the breast cancer incidence results.[5] The breast cancer incidence results include invasive breast cancer as well as non-invasive ductal carcinoma in situ (DCIS). DCIS tumors are considered to be of ultralow risk, but the word "carcinoma" causes alarm, which has prompted proposals to rename DCIS tumors "indolent lesions of epithelial origin (IDLE)" (Esserman and Varma, 2019). Since many DCIS tumors can only be diagnosed by mammograms, it is important that they are included in my analysis of overdiagnosis.

In the results that serve as the foundation for my analysis, to increase the relevance of my findings to the USPSTF guidelines for women in their 40s, I depart from the latest published results in four ways. First, I only include women aged 40-49 at enrollment in my main analysis sample, and I examine robustness among women aged 50-59 at enrollment. Second, because the

[5] In the Baines et al. (2016) calculation of overdiagnosis 25 years after the first CNBSS participants enrolled, the difference in breast cancer incidence between the intervention and control groups is 0.41% (=7.43% - 7.02%). In my replication, the difference is 0.44% (=7.51% - 7.07%). Both differences are statistically significant.

USPSTF guidelines are intended for asymptomatic women without a genetic predisposition for breast cancer, and because I aim to exclude women with potential knowledge of increased breast cancer risk, I exclude women from my main analysis sample if they report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms. I also exclude women if a nurse found abnormalities or referred them for review. 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, when analyzing mortality, I focus on all-cause mortality, because it is less subjective than breast cancer mortality, and because it can capture a wider range of collateral harms. 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 calendar date that reflects various follow-up lengths. I also examine robustness at earlier follow-up lengths. The mortality results, which are not statistically significant in any sample, vary in sign across robustness samples. However, the breast cancer results are consistent with overdiagnosis in all of the samples that I consider.

3 Model

I use an MTE model to allow for selection and treatment effect heterogeneity within the CNBSS. As the foundation for the model, I make only stylistic changes to the model used by Heckman and Vytlacil (2005)[6] to ensure that the model assumes no more than the LATE assumptions of Imbens and Angrist (1994), as proven by Vytlacil (2002). I present implications of the model using simple figures. The figures motivate the identification of selection heterogeneity under the LATE assumptions and the identification of treatment effect heterogeneity under a single ancillary assumption beyond the LATE assumptions.

3.1 First Stage: Mammography

In the context of the CNBSS, I use "treatment" to refer to mammography, which I represent with D. I set $D = 1$ if a participant receives a mammogram in at least one year during the active study period after the enrollment year. I set $D = 0$ otherwise.

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:

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

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:

[6] One stylistic change is that I do not condition on an optional covariate vector. The absence of the covariate vector simplifies the exposition and emphasizes the role of the *unobserved* net cost of treatment.

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

A.2. (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.3. (Instrument Relevance) $\mu_D(Z)$ is a nondegenerate random variable.

Under A.1, 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, I interpret the guidelines such that they recommend mammography for women with a net benefit of treatment $V_T - V_U$ that is greater than zero. As I show for completeness in Appendix B, given $V_T - V_U$ greater than 0 in (2), A.2 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.3 holds, then the observed threshold is different for the control and intervention groups, resulting in two special cases of the treatment equation:

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

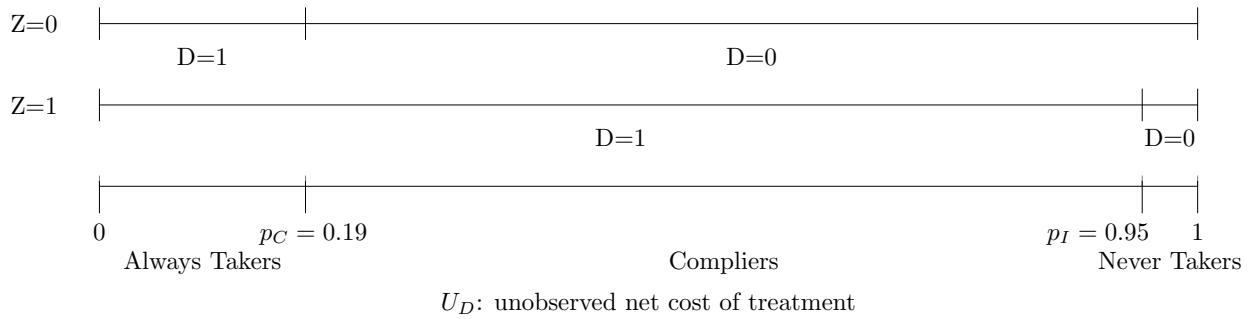
$$D = 1\{U_D \leq p_I\} \qquad \text{where } p_I = P\big(D = 1 \mid Z = 1\big), \tag{5}$$

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 illustrate implications of the first stage of the model in Figure 1 using statistics from the CNBSS. In my main analysis sample, 19% of control group participants and 95% of intervention group participants receive mammograms, so $p_C = 0.19$ and $p_I = 0.95$. By (4) and (5), p_C partitions the control group into the two ranges depicted in the top line, and p_I partitions the intervention group into the two ranges depicted in the middle line. Together, p_C and p_I partition the control and intervention groups into the three ranges depicted in the bottom line. I label the ranges using terminology from Imbens and Angrist (1994) in which "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 treatment regardless of random assignment.

The depiction in Figure 1 emphasizes that there is an ordering from always takers to compliers to never takers, which has been shown previously (Imbens and Rubin, 1997; Vytlacil, 2002). In the CNBSS, this ordering reflects mammography behavior within the trial. Always takers are the

Figure 1: Always Takers Are More Likely To Receive Mammograms Than Compliers, Who Are More Likely to Receive Mammograms Than Never Takers: Ranges of U_D



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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review.

most likely to receive mammograms (they receive them with probability 1), followed by compliers (they receive them with the probability of assignment to intervention), followed by never takers (they receive them with probability 0).

3.2 Second Stage: Health Outcomes

I relate a health outcome Y, such as breast cancer incidence or all-cause mortality, to mammography D as follows:

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

Y_T represents the potential treated outcome and Y_U represents the potential untreated outcome, 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. (Some Treated and Untreated) $0 < \mathrm{P}(D = 1) < 1$.

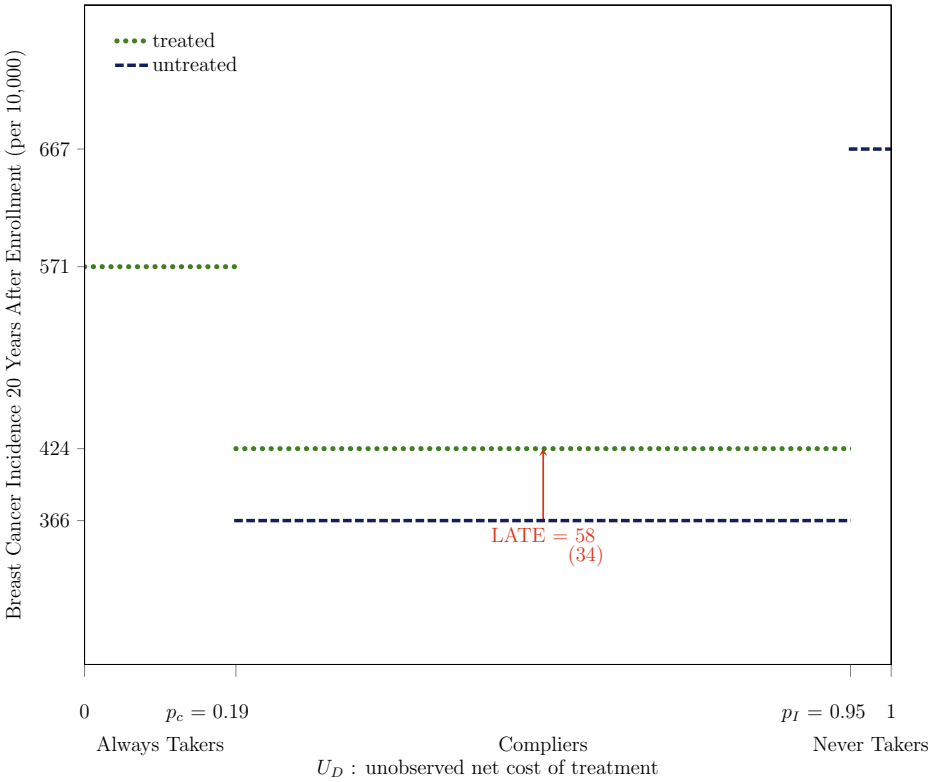
A.5. (Finite Average Outcomes) The values of $E[Y_T]$ and $E[Y_U]$ are finite.

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.3–A.5, assumes no more than the LATE assumptions. This claim follows because my presentation of the model differs only stylistically from the presentation of the model by Heckman and Vytlacil (2005), who invoke the proof in Vytlacil (2002) to claim that the model assumes no more than the LATE assumptions.

I illustrate implications of the second stage of the model using statistics from my main analysis sample in Figure 2. The horizontal axis depicts implications of the first stage as in Figure 1. The

Figure 2: Breast Cancer Incidence for Always Takers, Compliers, and Never Takers



Note. Bootstrapped standard errors in parentheses. The outcome is breast cancer incidence, measured 20 years after enrollment for all participants, based on initial diagnosis and 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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review.

vertical axis depicts implications of the second stage in which the outcome is breast cancer incidence. It is possible to identify some individuals as always takers because they receive mammograms despite assignment to the control group, and it is possible to identify other individuals as never takers because they do not receive mammograms 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 make it possible to calculate the average outcomes for always takers, compliers, and never takers depicted in Figure 2, as I show algebraically and graphically in Appendix C. These derivations yield the same values as the derivations by Imbens and Rubin (1997), Katz et al. (2001), Abadie (2002), and Abadie (2003), which rely on the LATE assumptions. The statistics that they yield are useful because they allow for comparisons across three groups, not just the two groups generated by intervention and control.

The depiction in Figure 2 makes clear that the LATE represents the average treatment effect on compliers but that always and never takers make up sizeable fractions of the sample. Always takers always receive mammograms within the experiment, so it is not possible to derive what their average breast cancer incidence would be if they had *not* received mammograms during the experiment

without ancillary assumptions. Similarly, never takers never receive mammograms within the experiment, so it is not possible to derive what their average breast cancer incidence would be if they *had* received mammograms during the experiment without ancillary assumptions. However, the average breast cancer incidence rates that can be derived for always and never takers appear very different from the average breast cancer incidence rates that can be derived for compliers. By 20 years after enrollment, 5.71% of always takers have been diagnosed with breast cancer, as compared to 4.24% of treated compliers and 3.66% of untreated compliers. In contrast, 6.67% never takers have been diagnosed with breast cancer, which is higher than any of the other reported rates. These differences provide the variation that I use as a starting point to identify selection and treatment effect heterogeneity.

3.3 Definitions of Selection and Treatment Effect Heterogeneity in the Model

I define selection and treatment effect heterogeneity on Y along the unobserved net cost of treatment U_D using functions from the MTE literature (see Carneiro and Lee, 2009; Brinch et al., 2017):

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

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 what I refer to as "selection heterogeneity" along the unobserved net cost of treatment U_D. Selection heterogeneity generalizes the concept of "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:

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

Unlike selection bias, selection heterogeneity does not depend on the fraction of individuals assigned to the intervention group, a parameter explicitly chosen as part of the trial design.[7] Furthermore, selection bias is not identified without ancillary assumptions. In contrast, a different special case of selection heterogeneity is identified without ancillary assumptions. By generating exogenous variation in the fraction of participants who receive treatment, thereby making untreated compliers distinguishable from never takers, randomization makes identification possible.

[7] I 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 $\omega(p, p_L, p_H) = 1\{p_L \leq p < p_H\}/(p_H - p_L)$:

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

This weighted integral depends on the probability of assignment to the intervention group $\mathrm{P}(Z=1)$, a parameter explicitly chosen as part of the trial design.

The second function is the "marginal treatment effect (MTE)" function of Heckman and Vytlacil (1999, 2001, 2005). It defines treatment effect heterogeneity along the unobserved net cost of treatment U_D. In the CNBSS, the MTE function characterizes how the impact of mammography on a health outcome changes as women become less likely to receive mammograms.

The third function, which I refer to as the "marginal treated outcome (MTO)" function, characterizes the sum of selection and treatment effect heterogeneity along the unobserved net cost of treatment U_D. It is tempting to assert that there should be no material distinction between treated and untreated outcomes. However, the treatment effect is defined as the treated outcome minus the untreated outcome, not the untreated outcome minus the treated outcome. The treatment effect has magnitude *and* direction, which is why I represent the LATE with an arrow in Figure 2.

4 Findings

Applying the model to the CNBSS, I identify and estimate how selection and the treatment effect vary with mammography behavior. 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 in terms of long-term breast cancer incidence and all-cause mortality. Baseline covariates that measure socioeconomic status and health behaviors, as well as results from the literature, corroborate this finding. The first finding informs an ancillary assumption that I impose to identify treatment effect heterogeneity. Under the ancillary assumption, I find treatment effect heterogeneity: the 20-year level of overdiagnosis is at least 3.5 times higher among women more likely to receive mammograms, such that at least 36% of their cancers are overdiagnosed.

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 ($p_I < U_D \leq 1$) using the following test statistic:

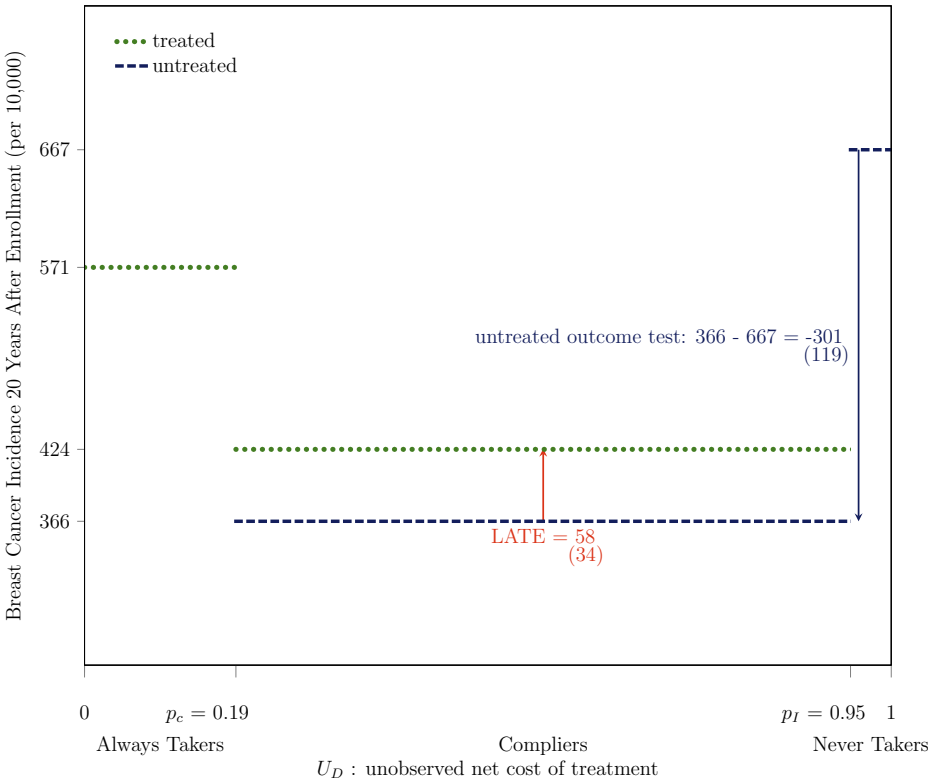
$$\mathrm{E}[Y_U \mid p_C < U_D \leq p_I] - \mathrm{E}[Y_U \mid p_I < U_D \leq 1] = \int_0^1 (\omega(p, p_C, p_I) - \omega(p, p_I, 1)) \,\mathrm{MUO}(p) \,\mathrm{d}p, \quad (9)$$

where $\omega(p, p_L, p_H) = 1\{p_L \leq p < p_H\}/(p_H - p_L)$. The test of the null hypothesis that this test statistic is equal to zero is equivalent or similar to tests proposed by Bertanha and Imbens (2014), Guo et al. (2014), and Black et al. (2017), which are generalized by Mogstad et al. (2018).[8] 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). Identification

[8]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 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. (2017) propose this test as one of two tests for "selection," which they do not define.

stems from randomization, which makes untreated compliers distinguishable from never takers.

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



Note. Bootstrapped standard errors in parentheses. The outcome is breast cancer incidence, measured 20 years after enrollment for all participants, based on initial diagnosis and 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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review.

Applying the untreated outcome test to my main analysis sample, I find selection heterogeneity on breast cancer incidence and all-cause mortality. As shown in Figure 3, the untreated outcome test statistic indicates that the long-term breast cancer incidence rate among untreated compliers was 3.01 percentage points lower than the breast cancer incidence rate among never takers, which was 6.67%. The test statistic is statistically different from zero,[9] so the untreated outcome test rejects selection homogeneity on breast cancer incidence. The test also rejects selection homogeneity on all-cause mortality. As shown in Figure D2, the 20-year all-cause mortality rate was 4.28% for untreated compliers and 9.90% for never takers. The 5.62 percentage point difference in all-cause mortality between these two groups is meaningfully large, and it is statistically different from zero. All-cause mortality and breast cancer incidence are both measures of health, and compliers are more likely to receive mammograms than never takers. Therefore, the selection heterogeneity that I find indicates that women more likely to receive mammograms are healthier.

[9]For inference, I bootstrap the test statistic using 200 replications, and I report the standard deviation as the standard error.

4.1.1 Baseline Covariates Corroborate Selection Heterogeneity

The untreated outcome test shows selection heterogeneity based on the comparison of long-term health outcomes in the absence of mammograms for compliers and never takers. I do not observe long-term health outcomes 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 whether the selection heterogeneity that I find also applies over the range of the unobserved net cost of treatment U_D from always takers to compliers. I obtain average covariates for always takers, compliers, and never takers in the same way that I obtain average outcomes but I combine the averages for treated and untreated compliers, as I discuss at the end of Appendix C.

Table 1: Baseline Covariates Corroborate Selection Heterogeneity:
Women More Likely to Receive Mammograms Have Higher Socioeconomic Status
and Are More Likely to Practice Other Health Behaviors Seen as Beneficial

	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 and 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)
Any 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. 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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review. 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. These patterns are consistent with an extensive literature that shows a negative correlation between socioeconomic status and health (see Pappas et al. (1993); Cutler and Lleras-Muney (2010); National Center for Health Statistics (2012)). Measures of baseline health behavior suggest a potential mechanism: women more likely to receive mammograms are more likely to practice other health behaviors seen as beneficial. As shown, smoking status, body mass index, and breast self-

examination vary monotonically from always takers to compliers to never takers, and many of the differences are statistically significant. Overall, analysis of baseline covariates corroborates the selection heterogeneity that I find from compliers to never takers. It also supports extension of the finding such that in the absence of mammograms, always takers would have the best health outcomes, followed by compliers, followed by never takers.

4.1.2 Findings from Related Literature Corroborate Selection Heterogeneity

Findings from the related literature on natural experiments also corroborate the selection heterogeneity that I find. Analyzing mammography takeup before and after age 40 in the United States in recent years, Einav et al. (2019) cannot observe long-term health outcomes, but they predict the long-term cancer incidence of women who did not receive mammograms with a clinical model. Re-cast in terms of the MTE model, their predictions imply that compliers are less likely to have cancer than never takers. They also present findings that imply that compliers are more likely than never takers to invest in their health through prior flu shots and pap tests. Analyzing a national cancer screening program in Korea that generated discontinuities in eligibility, Kim and Lee (2017) find that among individuals who were not screened through the program, cancer incidence was lower for compliers than never takers six years afterward. Furthermore, they show that individuals who were screened were healthier in terms of body mass index, blood glucose, and cholesterol. Other evidence from Goldman and Smith (2002); Berrigan et al. (2003); Friel et al. (2005); Brookhart et al. (2007); Cutler and Lleras-Muney (2010); Cutler et al. (2011); Myerson et al. (2018) and Oster (2018) shows that individuals often simultaneously select into several health behaviors, consistent with my finding that women more likely to receive mammograms are also more likely to practice other health behaviors seen as beneficial.

4.2 Treatment Effect Heterogeneity: Women More Likely to Receive Mammograms Experience Higher Levels of Overdiagnosis

To identify treatment effect heterogeneity, I impose an ancillary assumption that builds on the selection heterogeneity that I find on breast cancer incidence as well as corroborating evidence from baseline covariates and related literature. In the CNBSS, the ancillary assumption implies that average health, measured by breast cancer incidence or all-cause mortality in the absence of mammograms, varies monotonically from always takers to compliers to never takers. I impose

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), M.1 imposes a related weak monotonicity in the second stage. Empirical selection heterogeneity determines the direction of weak monotonicity. In the CNBSS, M.1 implies that always takers would be healthier than compliers in the absence of mammograms because compliers are healthier than never takers in the absence of mammograms.

Brinch et al. (2017) impose M.1 in conjunction with an analogous assumption on the MTO function. I emphasize that either of the Brinch et al. (2017) assumptions is sufficient to test for

treatment effect heterogeneity. I only impose M.1 in the CNBSS because I can find empirical support for it using baseline measures of socioeconomic status and health behavior, which corroborate selection heterogeneity. In contrast, alternative assumptions on the MTO or MTE functions entail assumptions about treatment effect heterogeneity. I prefer not to identify treatment effect heterogeneity by making assumptions about treatment effect heterogeneity.

M.1 yields a one-sided bound on the average treatment effect for always takers that is of interest in its own right. It is well-known that it is possible to estimate bounds on the average treatment effect of always takers using bounds that arise from the natural range of outcomes (Robins, 1989; Manski, 1990; Balke and Pearl, 1997) or from ancillary assumptions (Imbens and Rubin, 1997). The ancillary assumptions made by Olsen (1980), Heckman (1979), and Brinch et al. (2017), discussed by Kline and Walters (2019), also imply bounds on the average treatment effect for always takers, but those assumptions are stronger, and it is more difficult to motivate them in the context of the CNBSS.

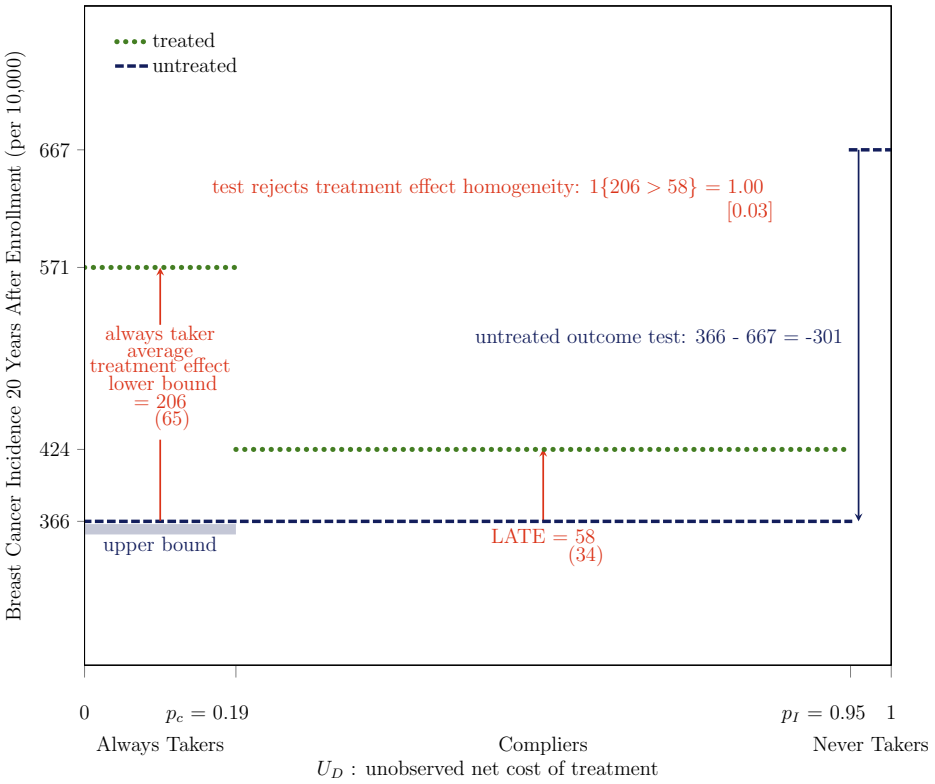
Imposing M.1, I test the null hypothesis of treatment effect homogeneity using the following decision rule, which has an outcome that is equal to 1 if the test rejects treatment effect homogeneity and 0 otherwise:

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

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. As illustrated in Figure 4, under M.1, the average outcome for untreated compliers $\mathrm{E}[Y_U \mid p_C < U_D \leq p_I]$ is an *upper* bound on the average untreated outcome of always takers $E[Y_U \mid 0 \leq U_D \leq p_C]$, which is not observed. The average treatment effect for always takers is equal to the average treated outcome of always takers minus the average untreated outcome of always takers. Therefore, the upper bound on the average untreated outcome 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 average treatment effect for compliers, the LATE. If the lower bound on the average treatment effect for always takers is strictly greater than the LATE, then the 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.

Within the CNBSS, the test rejects treatment effect homogeneity. As depicted in Figure 4, the lower bound on the always taker average treatment effect is strictly greater than the LATE, so the treatment effect on always takers is strictly greater than the average treatment effect on compliers. Therefore, the decision rule in (10) yields a value of one. For inference, I repeat the test in 200

Figure 4: Test Rejects Treatment Effect Homogeneity on Breast Cancer Incidence at 3% Level: Overdiagnosis is at Least 3.5 Times Higher Among Women More Likely to Receive Mammograms At Least 36% (= 206/571) of Their Cancers are Overdiagnosed



Note. Bootstrapped standard errors in parentheses and p-values in brackets. The outcome is breast cancer incidence, measured 20 years after enrollment for all participants, based on initial diagnosis and 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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review. Some differences between statistics might not appear internally consistent because of rounding.

bootstrap samples, and the decision rule yields a value of one in 97% of them. Therefore, the test rejects treatment effect homogeneity at the 3% level.

The magnitude of the treatment effect heterogeneity is meaningful, as are the magnitudes of the treatment effects themselves. As depicted in Figure 4, the magnitude of the treatment effect on compliers, the LATE, indicates that by 20 years after enrollment, breast cancer incidence among compliers who received mammograms during the active study period was 0.58 percentage points higher than it would have been otherwise. To put this magnitude in context, 20-year breast cancer incidence was 424 per 10,000 among treated compliers, so the LATE indicates that 14% (=0.58/4.24) of breast cancers, almost 1 in 7 breast cancers, were overdiagnosed. Turning to always takers, the lower bound on the treatment effect for always takers indicates that breast cancer incidence among always takers who received mammograms during the active study period was at least 2.06 percentage points higher than it would have been otherwise. Thus, the treatment effect for always takers was at least 3.5 (=2.06/0.58) times higher than it was for compliers. Always takers

are more likely to receive mammograms than compliers. Therefore, I summarize the treatment effect heterogeneity that I find by saying that the 20-year level of overdiagnosis is at least 3.5 times higher among women more likely to receive mammograms. Furthermore, among the women more likely to receive mammograms, the always takers, the 20-year breast cancer incidence rate was 5.71%, so at least 36% (=2.06/5.71) of their breast cancers were overdiagnosed.

The rates of overdiagnosis that I estimate within my main analysis sample, 36% among always takers and 14% among compliers, fall squarely within the range of overdiagnosis estimates from literature. Estimates in the literature vary in their data sources, their identification strategies, the types of breast cancers that they consider, and the denominators that they use to calculate overdiagnosis rates. Within the CNBSS, Miller et al. (2014) reports an overdiagnosis rate of 22%, and Baines et al. (2016) report several different overdiagnosis rates that vary from 5% to 48%. In other contexts, which include clinical trials as well as natural experiments created by population screening programs, estimates vary widely.[10]

I do not provide an estimate of overdiagnosis among never takers. In the long term, never takers could receive mammograms (the term "never taker" gets its meaning within the active study period), so never takers could be overdiagnosed or underdiagnosed. I could potentially estimate whether they are overdiagnosed or underdiagosed on average by making making additional assumptions. However, assumptions analogous to M.1 on the MTO and MTE functions are either difficult to defend or uninformative, so I refrain from imposing them.[11]

The average treatment effects on always takers and compliers are arguably more informative about the impact of a realistic future change in the USPSTF guidelines than the average treatment effect on never takers. For example, suppose that the USPSTF changes it recommendation in the future such that it does not recommend mammography for any asymptomatic women in their 40s. Such a policy change would divide women into always takers, compliers, and never takers. Always takers would be those who receive mammograms under the current and future policy, compliers would be those who receive mammograms only under the current policy, and never takers would be those who do not receive mammograms under the current or future policy. The groups of always takers, compliers, and never takers generated by the policy change need not correspond to the groups of always takers, compliers, and never takers generated by the CNBSS. Nonetheless, it seems

[10] Etzioni et al. (2002); Pohl and Welch (2005); Zackrisson et al. (2006); Jørgensen and Gøtzsche (2009); Bleyer and Welch (2012); Marmot et al. (2012); Baum (2013); Duffy and Parmar (2013); Biller-Andorno and Jüni (2014); Helvie et al. (2014); Miller et al. (2014); Patz et al. (2014); Welch and Passow (2014); Harding et al. (2015); Baines et al. (2016); McCaffery et al. (2016); Nelson et al. (2016); Welch et al. (2016); Jørgensen et al. (2017); Lannin and Wang (2017); Raffle and Gray (2019)

[11] Specifically, weak monotonicity of the MTO function would imply that the sum of selection and treatment effect heterogeneity is monotonic from always takers to never takers to compliers, but such an assumption would be difficult to defend in the context of the CNBSS. Within the CNBSS, my two main findings show that 1) selection heterogeneity is increasing and 2) treatment effect heterogeneity is decreasing along the unobserved net cost of treatment U_D. Therefore, it is unclear if their *sum* should be decreasing or increasing. Baseline covariates only inform selection heterogeneity; they do not inform the sum of selection and treatment effect heterogeneity. It could be more palatable to impose weak monotonicity of the MTE function. However, alone, such an assumption would not identify a treatment effect on never takers. In conjunction with M.1, such an assumption would imply that the treatment effect is smaller for never takers than it is for compliers, but the smaller treatment effect could be positive or negative, so it would not separate overdiagnosis from underdiagnosis.

reasonable that the never takers within the CNBSS, who did not get mammograms even after being randomized into the intervention group, would also be never takers under a policy change that did not recommend mammograms. Since such a policy change would not affect mammograms for never takers, average treatment effects on always takers and compliers are arguably more informative.

4.2.1 Breast Cancer Characteristics Corroborate Treatment Effect Heterogeneity

One concern with my finding of treatment effect heteterogeneity, which shows that women more likely to receive mammograms are more likely to be overdiagnosed by them, is that M.1 does not actually hold, such that always takers would actually have higher breast cancer incidence than compliers in the absence of mammograms. This could be the case, for example, if always takers receive mammograms because they know that they have a higher risk of breast cancer than compliers, despite appearing healthier on other dimensions. To address this concern, in addition to selecting the sample to exclude women who could have knowledge that they have a higher risk of breast cancer, I compare average characteristics of the breast cancers detected among always takers and treated compliers during the active study period.

As shown in Table 2, I find suggestive evidence that breast cancers detected among always takers are smaller and less invasive than breast cancers detected among treated compliers. One potential explanation for this evidence is selection heterogeneity such that always takers with breast cancer are healthier than compliers with breast cancer, consistent with my main finding of selection heterogeneity such that women more likely to receive mammograms are healthier. A second potential explanation is that mammography has a larger treatment effect on breast cancer diagnosis for always takers relative to compliers such that given the same or better underlying health, always takers are more likely to be diagnosed with breast cancer. The second explanation is consistent with my finding of treatment effect heterogeneity such that women more likely to receive mammograms are more likely to be overdiagnosed by them.

Table 2: Suggestive Evidence that Women More Likely to Receive Mammograms Have Breast Cancers That Are Smaller and Less Invasive and Undergo More Aggressive Procedures

	Means		Difference in Means
	(1)	(2)	(1) - (2)
	Always Takers	Treated Compliers	
Tumor Size Among Breast Cancers (in mm)	13	18	-5
	(2)	(3)	(4)
Share of Invasive Breast Cancer Among Breast Cancers (%)	73	75	-2
	(9)	(7)	(13)
Share of Mastectomy Among Breast Cancers with Mastectomy or Lumpectomy (%)	45	23	22
	(9)	(7)	(13)

Note. Bootstrapped standard errors in parentheses. All outcomes are restricted to those years for which treatment is defined during the active study period. Lumpectomy is a procedure that involves partial removal of the breast, and mastectomy is a more aggressive procedure that involves complete removal of the breast. 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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review.

5 Robustness

I examine the robustness of my two main findings by estimating my main specification with alternative outcomes, alternative sample restrictions, alternative definitions of mammography, and alternative follow-up lengths. To facilitate comparisons with my main specification, I summarize important statistics from Figure 4 in Table 3. A specification shows my two main findings — selection and treatment effect heterogeneity such that women more likely to receive mammograms are healthier and experience a larger treatment effect from mammograms — 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 (5).

Table 3: Summary of Findings Depicted in Figure 4
and Robustness to Alternative Outcomes, Sample Restrictions, and Definitions of Mammography

	N	(1) Untreated Outcome Test	(2) Always Taker Average Treatment Effect Lower Bound	(3) Local Average Treatment Effect LATE	(4) Lower Bound on Treatment Effect Ratio (2)/(3)	(5) Test Rejects Treatment Effect Homogeneity 1{(2) > (3)}
Main Specification						
Outcome is breast cancer incidence, sample is main analysis sample, treatment is defined as mammogram in at least one active study period after enrollment						
Breast cancer incidence	19,505	-301	206	58	3.5	1.00
		(119)	(65)	(34)	(12)	[0.03]
Alternative Outcomes						
All-cause mortality	19,505	-562	22	-13	-	1.00
		(147)	(59)	(38)		[0.36]
Breast cancer mortality	19,505	-43	30	-12	-	1.00
		(47)	(25)	(13)		[0.23]
Alternative Sample Restrictions						
All excluded participants aged 40-49 at enrollment	30,925	-1,237	309	79	3.9	1.00
		(147)	(48)	(43)	(35)	[0.00]
All participants aged 40-49 at enrollment	50,430	-826	298	69	4.3	1.00
		(107)	(40)	(31)	(31)	[0.00]
All participants aged 50-59 at enrollment	39,405	-1,555	419	39	10.7	1.00
		(140)	(49)	(34)	(89)	[0.00]
All participants	89,835	-1,156	332	55	6.0	1.00
		(96)	(31)	(21)	(95)	[0.00]
Alternative Definitions of Mammography						
At least two active study period years after enrollment	19,505	-341	239	54	4.5	1.00
		(95)	(95)	(32)	(16)	[0.03]
At least three active study period years after enrollment	19,505	-330	167	55	3.0	1.00
		(73)	(145)	(32)	(9)	[0.19]
All active study period years after enrollment	19,505	-178	158	64	2.5	1.00
		(61)	(190)	(38)	(10)	[0.31]

Note. Bootstrapped standard errors in parentheses and p-values in brackets, based on 200 bootstrap samples, which yields some p-values of zero. All outcomes are measured 20 years after enrollment per 10,000 participants for all participants, based on initial occurrence and the exact calendar date of enrollment. In the main specification, 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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review. Values in column (4) are omitted (-) if the ratio of column (2) to column (3) does not yield a lower bound on the treatment effect ratio, which occurs when the LATE is negative. Some differences between statistics might not appear internally consistent because of rounding.

5.1 Alternative Outcomes

I have shown that my first finding of selection heterogeneity, which shows that women more likely to receive mammograms are healthier, holds in terms of breast cancer incidence and all-cause mortality.

In Table 3, I also examine breast cancer mortality for comparison to the literature, even though it is subject to more classification error than all-cause mortality. The untreated outcome test statistic in column (1) shows suggestive evidence that women more likely to receive mammograms are also healthier in terms of breast cancer mortality.

Turning to my second finding, column (5) shows suggestive evidence of treatment effect heterogeneity such that women more likely to receive mammograms experience greater harms from them, as measured in terms of all-cause and breast cancer mortality. I illustrate the results for all-cause mortality in Figure D2. The always taker average treatment effect lower bound is greater than the LATE, but the difference is not statistically significant. The lack of statistical significance is unsurprising given that the LATE itself is not statistically significant. Consistent with imprecision, the sign of the LATE on all-cause mortality is negative in my main analysis sample, even though the latest results published by CNBSS investigators (Miller et al., 2014) would be consistent with a positive LATE. The finding of treatment effect heterogeneity holds regardless of the sign of the LATE.

The magnitude of the lower bound on the average treatment effect on mortality for always takers is notable. It indicates that always takers experience at least an additional 22 deaths per 10,000 participants when they receive mammograms, which suggests that at least 4.9% (= 22/451) of their deaths would not have occurred otherwise. For comparison, the World Health Organization estimates the number of road traffic deaths in the entire U.S. population each year at 1.1 per 10,000 people (World Health Organization, 2015). Therefore, the lower bound on the average treatment effect for always takers, which is measured over a 20-year period, is comparable to the rate of road traffic deaths over a period of the same length.

Why might women more likely to receive mammograms be more likely to experience harm from them, as measured in terms of all-cause mortality and breast cancer mortality? As shown in the first two rows of Table 2, I find suggestive evidence that women more likely to receive mammograms have breast cancers that are smaller and less invasive. Virtually all women diagnosed with breast cancer during the active study period underwent lumpectomy or mastectomy. Whereas lumpectomy involves only partial removal of the breast, mastectomy is a more aggressive procedure that involves complete removal of the breast. The third row of Table 2 shows that, among women with breast cancer who underwent either of these procedures during the active study period, 45% of always takers underwent the more aggressive procedure of mastectomy, compared to only 23% of compliers. These results suggest that women more likely to receive mammograms may receive more aggressive treatment for smaller, less invasive breast cancers. These aggressive treatments could lead to increased collateral harms in the form of all-cause and breast cancer mortality.

5.2 Alternative Sample Restrictions

In the rest of Table 3, I examine the robustness of my findings in terms of breast cancer incidence. I consider alternative sample restrictions that include all *excluded* participants aged 40-49 at enrollment (those who report any breast cancer in their family, any previous false-positive breast cancer diagnosis, any other breast disease, or any symptoms and those for whom a nurse found

abnormalities or referred them for review), all participants aged 40-49 at enrollment, all participants aged 50-59 at enrollment, and all participants. My first and second findings hold in all of the reported samples. Furthermore, the lower bound on the ratio of treatment effects for always takers and compliers is always at least 3.5. Hence, the 20-year level of overdiagnosis is at least 3.5 times higher among always takers than it is among compliers, regardless of the sample.

5.3 Alternative Definitions of Mammography

In the main specification, 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. Since I start with a broad definition of mammography, I assess robustness to narrower definitions of mammography in Table 3. The results for selection and treatment effect heterogeneity yield the same conclusions as those from the main specification, although the test for treatment effect homogeneity is not statistically significant under the two narrowest definitions. The two narrowest definitions are arguably too extreme because they require that "treated" participants must receive mammograms in three or more active study period years after enrollment, so it is notable that the results yield the same conclusions.

I cannot examine robustness to definitions that include mammography after the active study period because such information was not collected. However, I know that breast cancer screening programs began in British Columbia in 1988 and in other Canadian provinces in the 1990s (Baines et al., 2016). Given the greater availability of mammography through such programs, mammography behavior in the control and intervention groups likely converged over time. Given likely convergence in mammography behavior, results from the CNBSS likely reflect the effect of starting mammography earlier, as opposed to the effect of ever receiving mammography. My ability to find heterogeneous selection and treatment effects in the face of likely long-term attenuation speaks to the robustness of my results.

5.4 Alternative Follow-up Lengths

In the main specification, breast cancer incidence is measured 20 years after enrollment. Table E1 summarizes results for breast cancer incidence at all earlier annual follow-up lengths. The untreated outcome test statistic is negative at all follow-up lengths, consistent with selection heterogeneity such that women more likely to receive mammograms are healthier. Furthermore, the test rejects treatment effect homogeneity at all follow-up lengths, consistent with treatment effect heterogeneity such that women more likely to receive mammograms experience higher levels of overdiagnosis.

Whether overdiagnosis can be estimated in the short term is controversial due to the concept of lead time. Lead time refers to "the time from detection of preclinical cancer by screening to detection of clinical (symptomatic) cancer in the absence of screening" (Baker et al., 2014). Short follow-up lengths might not allow for enough lead time, such that excess breast cancer detection in the intervention group could just reflect lead time instead of overdiagnosis. However, once there is evidence of overdiagnosis in the long term, estimates from the short term can also be interpreted as estimates of overdiagnosis (Zahl et al., 2013; Baines et al., 2016). As shown in Table E1, the LATE is positive and statistically significant in the first year, and it is still statistically significant

at longer follow-up lengths, consistent with overdiagnosis. Consequently, my findings at earlier follow-up lengths could also reflect overdiagnosis.

6 Discussion

The active study period of the CNBSS took place in the 1980s, so it is important to assess whether my findings are still applicable in the current environment. It is plausible that my first finding, which shows that women more likely to receive mammograms are healthier, is still applicable, especially given recent evidence from natural experiments (Einav et al., 2019; Oster, 2018). It is also plausible that my second finding, which shows that women more likely to receive mammograms have a higher level of overdiagnosis, is still applicable. Recent research does not inform current rates of overdiagnosis since recent trials lack long-term follow-up data. The threat of overdiagnosis, however, remains present. As mammograms become increasingly accurate, they could identify even smaller tumors that would never become life-threatening. At the same time, existing breast cancer treatments have likely become less harmful, so the impact of overdiagnosis on mortality may have decreased. However, new breast cancer drugs have also been developed. As drugs become more effective at treating advanced cancers, there is less of a need to screen women before they develop symptoms, especially given potential overdiagnosis.

In the current 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. In 2015, 64% of U.S. women aged 40 and older received a mammogram within the previous two years (National Health Interview Survey, 2017). 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. The active study period of the CNBSS was not particularly recent, but changes in environment are an inherent limitation of any long-term analysis. The current USPSTF guidelines consider previous findings from the CNBSS, which are based on a comparison of average outcomes between the intervention and control groups. Since I uncover treatment effect heterogeneity in the CNBSS based on mammography behavior, my findings could also be useful for future guidelines.

7 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 behavior and thus health outcomes in practice. To do so, I examine relationships between behavior and health outcomes within existing clinical trial data. Specifically, I examine relationships between mammography behavior and health outcomes in the CNBSS, an influential and extensive trial on mammography.

My first finding shows heterogeneous selection: women more likely to receive mammograms are healthier and of higher socioeconomic status. My second finding shows treatment effect heterogeneity: women more likely to receive mammograms are more likely to be overdiagnosed with breast cancer by them. This result is statistically significant, consistent with a growing consensus in the

literature that shows a positive treatment effect on overdiagnosis. I also find that women more likely to receive mammograms may be more likely to face a higher rate of 20-year all-cause mortality from them. While this result is not statistically significant, such imprecision is unsurprising, as there is no consensus in the literature on the simpler matter of the sign of the treatment effect on all-cause mortality. Within the CNBSS data, I find suggestive evidence of a potential mechanism for increased all-cause mortality: women more likely to receive mammograms have breast cancers that are smaller and less invasive, but they pursue more aggressive procedures, which could increase collateral harm.

Growing concern about overdiagnosis and collateral harm from mammography for women in their 40s has prompted changes in screening guidelines around the world. The Candian guidelines, which changed in 2018, do not recommend mammography for any asymptomatic women in their 40s (Klarenbach et al., 2018). The Swiss Medical Board took steps to limit screening programs in 2014 (Biller-Andorno and Jüni, 2014), and in 2016 the French Minister of health released results of an independent review that recommended an end to screening or radical reforms (Barratt et al., 2018). As of 2018, neither the Swiss nor French screening guidelines recommend regular mammography screening for asymptomatic women in their 40s (Ebell et al., 2018).

Had I found that women more likely to receive mammograms were also more likely to benefit from them, then my findings would have supported the current USPSTF guidelines for women in their 40s. Instead, my findings support a further weakening of the guidelines such that they no longer recommend regular mammography for any asymptomatic women in their 40s. Under such revised guidelines, women who do not receive mammograms under the current guidelines would likely be unaffected. However, some women who receive mammograms under the current guidelines would likely not receive them under revised guidelines. My findings indicate that these women would benefit from a reduction in overdiagnosis.

Even under weaker guidelines, it is likely that some asymptomatic women in their 40s would still obtain mammograms, and it is logical for guidelines to leave room for such behavior, as mammography may still be appropriate for some asymptomatic women in their 40s. Mammography can offer benefits to women that are not captured by breast cancer diagnosis or all-cause mortality, such as the peace of mind that comes with a negative breast cancer diagnosis. As articulated by a clinical nurse,"Being preoccupied with saving one's life produces a myopia, in which other worries unrelated to one's possibly imminent death fall away." (Brown, 2017). For asymptomatic women considering whether to receive mammograms even under revised guidelines that do not recommend them, my findings can help to inform the magnitudes of the impacts on two key health outcomes. For women in their 40s who see or feel abnormalities in their breasts, it is important to note that screening guidelines apply only to asymptomatic women, so guideline changes should not affect the mammography behavior of women with symptoms.

Beyond the context of mammograms, my findings support the need for clinical trials to collect data on behavior and for those 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 considered in the development of guidelines. Whenever the USPSTF determines that "there is at least moderate certainty that the net benefit is small," it 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). These C recommendations presuppose 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, these guidelines are 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 enhance the ability of future guidelines to target treatments toward individuals most likely to benefit from them.

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.

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

■

Appendix B Derivation of the Treatment Equation

Treatment D is given by

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

where the last equality follows from

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

Appendix C Derivations of Average Outcomes and Covariates

These derivations yield the same values as the derivations by Imbens and Rubin (1997), Katz et al. (2001), Abadie (2002), and Abadie (2003), which rely on the LATE assumptions. 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.2 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:

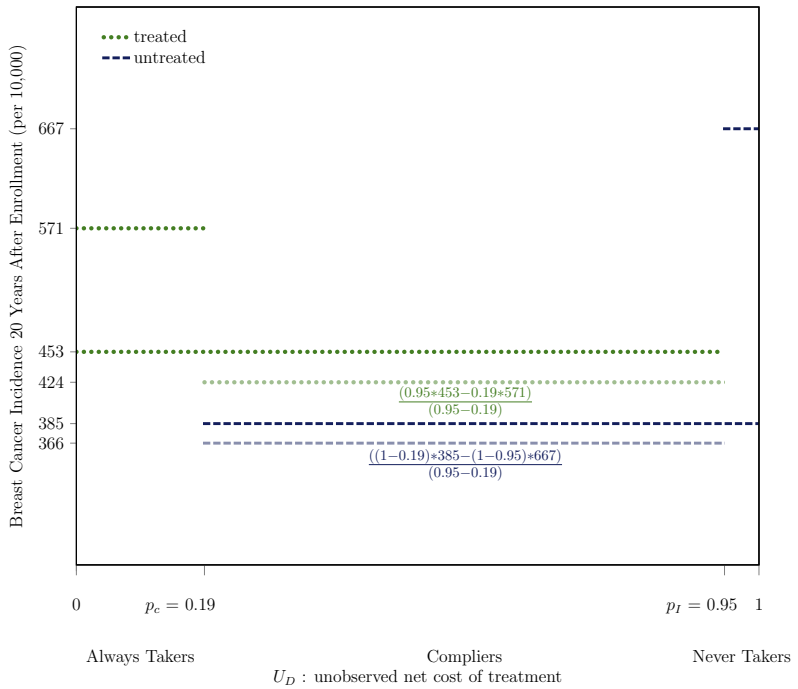
$$\begin{aligned}
\mathrm{E}[Y \mid D = 1, Z = 0] &= \mathrm{E}[Y_U + D(Y_T - Y_U) \mid D = 1, Z = 0] && \text{(by (6))} \\
&= \mathrm{E}[Y_T \mid D = 1, Z = 0] \\
&= \mathrm{E}[Y_T \mid 0 \leq U_D \leq p_C, Z = 0] && \text{(by (4), where } p_C = P(D = 1|Z = 0)) \\
&= \mathrm{E}[g_T(U_D, \gamma_T) \mid 0 \leq U_D \leq p_C, Z = 0] && \text{(by (7))} \\
&= \mathrm{E}[g_T(U_D, \gamma_T) \mid 0 \leq U_D \leq p_C] && (Z \perp (U_D, \gamma_T) \text{ by A.2}) \\
&= \mathrm{E}[Y_T \mid 0 \leq U_D \leq p_C].
\end{aligned}$$

In the CNBSS, I obtain an average treated outcome of always takers of 571 breast cancers per 10,000 women, which I plot over the relevant range ($0 \leq U_D \leq p_C$) 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. 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 453 breast cancers 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:

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

Figure C1: Derivation of Average Breast Cancer Incidence for Treated and Untreated Compliers (Lighter Shading)



Note. Bootstrapped standard errors in parentheses. The outcome is breast cancer incidence, measured 20 years after enrollment for all participants, based on initial diagnosis and 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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review.

In the CNBSS, the average treated outcome of compliers is 424 breast cancers 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. 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 667 breast cancers 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 385 breast cancers per 10,000 women, and I plot it over the 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:

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

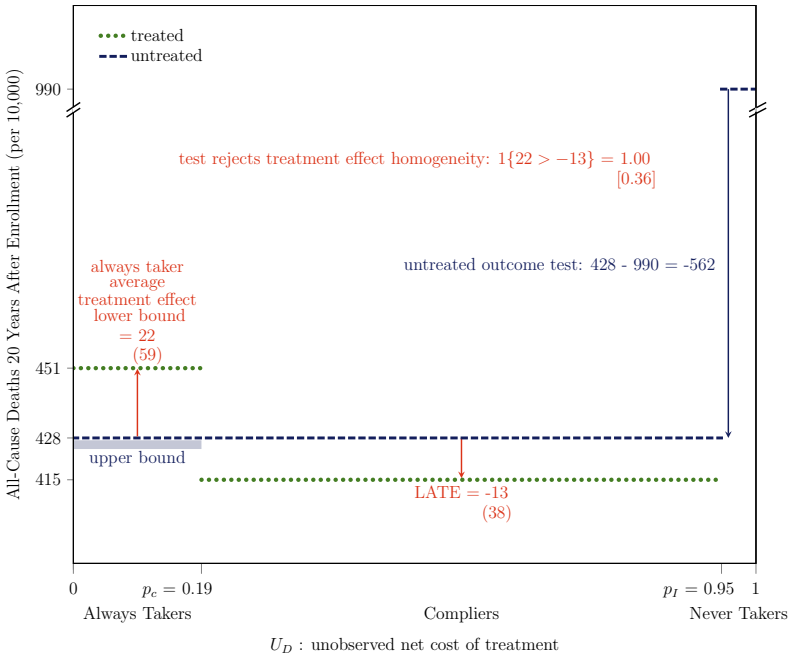
In the CNBSS, the average untreated outcome of compliers is 366 breast cancers per 10,000 women, which I plot over the relevant range ($p_C < U_D \leq p_I$) in Figure C1, using a dashed line to indicate that it represents an average untreated outcome.

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 by weighting the average covariate vectors for the treated and untreated compliers by the probabilities of being assigned to the intervention and control groups:

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

Appendix D Alternative Outcome: All-Cause Mortality

Figure D2: Untreated Outcome Test Rejects Selection Homogeneity on All-Cause Mortality: Women More Likely to Receive Mammograms are Healthier and Test Rejects Treatment Effect Heterogeneity on All-Cause Mortality at 36% Level: Women More Likely to Receive Mammograms Experience Greater Harm From Them At Least 4.9% (= 22/451) of Their Deaths Would Not Have Occurred Otherwise



Note. Bootstrapped standard errors in parentheses, and p-value in brackets. The outcome is all-cause mortality, 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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review. Some differences between statistics might not appear internally consistent because of rounding.

Appendix E Alternative Follow-up Lengths

Table E1: Summary of Findings Depicted in Figure 4
and Robustness to Alternative 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) Lower Bound on Treatment Effect Ratio (2)/(3)	(5) Test Rejects Treatment Effect Homogeneity 1{(2) > (3)}
Main specification: 20	19,505	-301	206	58	3.5	1.00
		(119)	(65)	(34)	(12)	[0.03]
19	19,505	-269	196	52	3.8	1.00
		(114)	(63)	(33)	(30)	[0.03]
18	19,505	-311	210	54	3.9	1.00
		(113)	(60)	(31)	(14)	[0.01]
17	19,505	-322	214	49	4.4	1.00
		(112)	(59)	(32)	(66)	[0.01]
16	19,505	-342	232	56	4.1	1.00
		(110)	(58)	(31)	(21)	[0.00]
15	19,505	-381	211	84	2.5	1.00
		(110)	(55)	(29)	(3)	[0.02]
14	19,505	-404	201	80	2.5	1.00
		(110)	(52)	(27)	(2)	[0.03]
13	19,505	-431	223	75	3.0	1.00
		(110)	(51)	(26)	(3)	[0.02]
12	19,505	-443	191	64	3.0	1.00
		(109)	(47)	(25)	(4)	[0.03]
11	19,505	-423	195	55	3.5	1.00
		(109)	(46)	(24)	(6)	[0.01]
10	19,505	-419	200	47	4.2	1.00
		(107)	(45)	(22)	(33)	[0.00]
9	19,505	-413	192	34	5.6	1.00
		(103)	(42)	(21)	(32)	[0.00]
8	19,505	-409	175	35	5.1	1.00
		(100)	(40)	(20)	(39)	[0.00]
7	19,505	-393	177	46	3.9	1.00
		(95)	(36)	(17)	(17)	[0.00]
6	19,505	-412	185	50	3.7	1.00
		(95)	(34)	(15)	(8)	[0.00]
5	19,505	-382	180	45	4.0	1.00
		(90)	(32)	(14)	(4)	[0.00]
4	19,505	-393	152	46	3.3	1.00
		(91)	(29)	(13)	(2)	[0.00]
3	19,505	-354	104	37	2.8	1.00
		(85)	(23)	(12)	(4)	[0.01]
2	19,505	-337	63	25	2.5	1.00
		(82)	(18)	(10)	(3)	[0.04]
1	19,505	-342	35	20	1.8	1.00
		(82)	(12)	(7)	(1)	[0.10]

Note. Bootstrapped standard errors in parentheses, and p-values in brackets. Statistical significance for all specifications is based on 200 bootstrap samples, including specifications that have a p-value value of zero. The outcome is breast cancer incidence per 10,000 participants. In the main specification, breast cancer incidence is measured 20 years after enrollment for all participants, based on initial diagnosis and 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. The main analysis sample includes women aged 40-49 at enrollment and excludes those who report any breast cancer in their family, any previous breast cancer diagnosis, any other breast disease, or any symptoms, as well as those for whom a nurse found abnormalities or referred them for review. Some differences between statistics might not appear internally consistent because of rounding.

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.

Aidan, J. C., N. R. Priddee, and J. J. McAleer (2013). Chemotherapy causes cancer! a case report of therapy related acute myeloid leukaemia in early stage breast cancer. *The Ulster medical journal 82*(2), 97.

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., 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.

Baker, S. G., P. C. Prorok, and B. S. Kramer (2014). Lead time and overdiagnosis.

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*(439), 1171–1176.

Barratt, A., K. J. Jørgensen, and P. Autier (2018). Reform of the national screening mammography program in france. *JAMA internal medicine 178*(2), 177–178.

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

Berrigan, D., K. Dodd, R. P. Troiano, S. M. Krebs-Smith, and R. B. Barbash (2003). Patterns of health behavior in us adults. *Preventive medicine 36*(5), 615–623.

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

Biller-Andorno, N. and P. Jüni (2014). Abolishing mammography screening programs? a view from the swiss medical board. *New England Journal of Medicine 370*(21), 1965–1967.

Bitler, M. and C. Carpenter (2019, August). Effects of direct care provision to the uninsured: Evidence from federal breast and cervical cancer programs. Working Paper 26140, 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 (2017, March). Simple tests for selection: Learning more from instrumental variables. Working Paper 6932, CESifo.

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.

Brody, J. E. (2017, July). With cancer screening, better safe than sorry? *New York Times*.

Brookhart, M. A., A. R. Patrick, C. Dormuth, J. Avorn, W. Shrank, S. M. Cadarette, and D. H. Solomon (2007). Adherence to lipid-lowering therapy and the use of preventive health services: an investigation of the healthy user effect. *American journal of epidemiology 166*(3), 348–354.

Brown, S. B., D. J. Hole, and T. G. Cooke (2007). Breast cancer incidence trends in deprived and affluent scottish women. *Breast cancer research and treatment 103*(2), 233–238.

Brown, T. (2017, October). Breast cancer is serious. pink is not. *New York Times*.

Buchmueller, T. C. and L. Goldzahl (2018, February). The effect of organized breast cancer screening on mammography use: Evidence from france. Working Paper 24316, 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.

Cooper, G. S., T. D. Kou, A. Dor, S. M. Koroukian, and M. D. Schluchter (2017). Cancer preventive services, socioeconomic status, and the affordable care act. *Cancer 123*(9), 1585–1589.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2018). Who benefits from universal child care? estimating marginal returns to early child care attendance. *Journal of Political Economy 126*(6), 2356–2409.

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

Cutler, D. M., A. Lleras-Muney, and T. Vogl (2011). Socioeconomic status and health: Dimensions and mechanisms. In *The Oxford Handbook of Health Economics*.

Duffy, S. W. and D. Parmar (2013). Overdiagnosis in breast cancer screening: the importance of length of observation period and lead time. *Breast Cancer Research 15*(3), R41.

Early Breast Cancer Trialists' Collaborative Group (2005). Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *The Lancet 366*(9503), 2087–2106.

Ebell, M. H., T. N. Thai, and K. J. Royalty (2018). Cancer screening recommendations: an international comparison of high income countries. *Public Health Reviews 39*(7).

Einav, L., A. Finkelstein, T. Oostrom, A. Ostriker, and M. R. Cullen (2019, August). Screening and selection: The case of mammograms. Working Paper 26162, National Bureau of Economic Research.

Esserman, L. J. and M. Varma (2019). Should we rename low risk cancers? *BMJ 364*, k4699.

Etzioni, R., D. F. Penson, J. M. Legler, D. Di Tommaso, R. Boer, P. H. Gann, and E. J. Feuer (2002). Overdiagnosis due to prostate-specific antigen screening: lessons from us prostate cancer incidence trends. *Journal of the National Cancer Institute 94*(13), 981–990.

Fedewa, S. A., M. Goodman, W. D. Flanders, X. Han, R. A. Smith, E. M. Ward, C. A. Doubeni, A. G. Sauer, and A. Jemal (2015). Elimination of cost-sharing and receipt of screening for colorectal and breast cancer. *Cancer 121*(18), 3272–3280.

Finkelstein, A. F., S. Taubman, B. J. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. L. Allen, K. Baicker, and the Oregon Health Study Group (2012). The Oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics 127*(3), 1057–1106.

Friel, S., J. Newell, and C. Kelleher (2005). Who eats four or more servings of fruit and vegetables per day? multivariate classification tree analysis of data from the 1998 survey of lifestyle, attitudes and nutrition in the republic of ireland. *Public Health Nutrition 8*(2), 159–169.

Goldman, D. P. and J. P. Smith (2002). Can patient self-management help explain the ses health gradient? *Proceedings of the National Academy of Sciences 99*(16), 10929–10934.

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.

Habermann, E. B., B. A. Virnig, G. F. Riley, and N. N. Baxter (2007). The impact of a change in medicare reimbursement policy and hedis measures on stage at diagnosis among medicare hmo and fee-for-service female breast cancer patients. *Medical care 45*(8), 761–766.

Hakama, M., T. Hakulinen, E. Pukkala, E. Saxen, and L. Teppo (1982). Risk indicators of breast and cervical cancer on ecologic and individual levels. *American journal of Epidemiology 116*(6), 990–1000.

Harding, C., F. Pompei, D. Burmistrov, H. G. Welch, R. Abebe, and R. Wilson (2015, 09). Breast Cancer Screening, Incidence, and Mortality Across US CountiesBreast Cancer Screening, Incidence, and Mortality Across US CountiesBreast Cancer Screening, Incidence, and Mortality Across US Counties. *JAMA Internal Medicine 175*(9), 1483–1489.

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.

Helvie, M. A., J. T. Chang, R. E. Hendrick, and M. Banerjee (2014). Reduction in late-stage breast cancer incidence in the mammography era: implications for overdiagnosis of invasive cancer. *Cancer 120*(17), 2649–2656.

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

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.

Jacobson, M. and S. Kadiyala (2017). When guidelines conflict: A case study of mammography screening initiation in the 1990s. *Women's Health Issues 27*(6), 692–699.

Jørgensen, K. J. and P. C. Gøtzsche (2009). Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. *BMJ 339*.

Jørgensen, K. J., P. C. Gøtzsche, M. Kalager, and P.-H. Zahl (2017). Breast cancer screening in denmark: a cohort study of tumor size and overdiagnosis. *Annals of internal medicine 166*(5), 313–323.

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.

Kadiyala, S. and E. C. Strumpf (2011). Are united states and canadian cancer screening rates consistent with guideline information regarding the age of screening initiation? *International Journal for Quality in Health Care 23*(6), 611–620.

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.

Kelaher, M. and J. M. Stellman (2000). The impact of medicare funding on the use of mammography among older women: implications for improving access to screening. *Preventive medicine 31*(6), 658–664.

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.

Klarenbach, S., N. Sims-Jones, G. Lewin, H. Singh, G. Thériault, M. Tonelli, M. Doull, S. Courage, A. J. Garcia, and B. D. Thombs (2018). Recommendations on screening for breast cancer in women aged 40–74 years who are not at increased risk for breast cancer. *CMAJ: Canadian Medical Association Journal 190*(49), E1441.

Kline, P. and C. R. Walters (2019). On heckits, LATE, and numerical equivalence. *Econometrica 87*(2), 677–696.

Kolstad, J. T. and A. E. Kowalski (2012). The impact of health care reform on hospital and preventive care: evidence from massachusetts. *Journal of Public Economics 96*(11-12), 909–929.

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). Reconciling seemingly constradictory results from the Oregon health insurance experiment and the Massachusetts health reform. Working Paper 24647, 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.

Kowalski, A. E. (2018b, September). Behavior within a clinical trial and implications for mammography guidelines. Working Paper 25049, National Bureau of Economic Research.

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.

Lu, Y. and D. J. Slusky (2016). The impact of women's health clinic closures on preventive care. *American Economic Journal: Applied Economics 8*(3), 100–124.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

Marmot, M., D. Altman, D. Cameron, J. Dewar, S. Thompson, and M. Wilcox (2012). The benefits and harms of breast cancer screening: an independent review. *Lancet 380*, 1778–86.

Martin, M. G., J. S. Welch, J. Luo, M. J. Ellis, T. A. Graubert, and M. J. Walter (2009). Therapy related acute myeloid leukemia in breast cancer survivors, a population-based study. *Breast cancer research and treatment 118*(3), 593–598.

McCaffery, K. J., J. Jansen, L. D. Scherer, H. Thornton, J. Hersch, S. M. Carter, A. Barratt, S. Sheridan, R. Moynihan, J. Waller, et al. (2016). Walking the tightrope: communicating overdiagnosis in modern healthcare. *Bmj 352*, i348.

Mehta, S. J., D. Polsky, J. Zhu, J. D. Lewis, J. T. Kolstad, G. Loewenstein, and K. G. Volpp (2015). Aca-mandated elimination of cost sharing for preventive screening has had limited early impact. *The American journal of managed care 21*(7), 511.

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.

Myerson, R. M., D. Lakdawalla, L. D. Colantonio, M. Safford, and D. Meltzer (2018, February). Effects of expanding health screening on treatment - what should we expect? what can we learn? Working Paper 24347, National Bureau of Economic Research.

Myerson, R. M., R. Tucker-Seeley, D. Goldman, and D. N. Lakdawalla (2019, September). Does medicare coverage improve cancer detection and mortality outcomes? Working Paper 26292, National Bureau of Economic Research.

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

National Health Interview Survey (2017). Use of mammography among women aged 40 and over, by selected characteristics : United states, selected years 1987 – 2015. https://www.cdc.gov/nchs/data/hus/2017/070.pdf.

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.

Ong, M.-S. and K. D. Mandl (2015). National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at $4 billion a year. *Health Affairs 34*(4), 576–583.

Oster, E. (2018, November). Behavioral feedback: Do individual choices influence scientific results? Working Paper 25225, National Bureau of Economic Research.

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.

Patz, E. F., P. Pinsky, C. Gatsonis, J. D. Sicks, B. S. Kramer, M. C. Tammemägi, C. Chiles, W. C. Black, and D. R. Aberle (2014). Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA internal medicine 174*(2), 269–274.

Pohl, H. and H. G. Welch (2005). The role of overdiagnosis and reclassification in the marked increase of esophageal adenocarcinoma incidence. *Journal of the National Cancer Institute 97*(2), 142–146.

Praga, C., J. Bergh, J. Bliss, J. Bonneterre, B. Cesana, R. C. Coombes, P. Fargeot, A. Folin, P. Fumoleau, R. Giuliani, et al. (2005). Risk of acute myeloid leukemia and myelodysplastic syndrome in trials of adjuvant epirubicin for early breast cancer: correlation with doses of epirubicin and cyclophosphamide. *Journal of clinical oncology 23*(18), 4179–4191.

Raffle, A. E. and J. M. Gray (2019). *Screening: evidence and practice*. Oxford University Press, USA.

Reynolds, P., S. E. Hurley, A.-T. Quach, H. Rosen, J. Von Behren, A. Hertz, and D. Smith (2005). Regional variations in breast cancer incidence among california women, 1988–1997. *Cancer Causes & Control 16*(2), 139.

Robert, S. A., I. Strombom, A. Trentham-Dietz, J. M. Hampton, J. A. McElroy, P. A. Newcomb, and P. L. Remington (2004). Socioeconomic risk factors for breast cancer: distinguishing individual-and community-level effects. *Epidemiology*, 442–450.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

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.

Welch, H. G. and W. C. Black (2010). Overdiagnosis in cancer. *Journal of the National Cancer Institute 102*(9), 605–613.

Welch, H. G. and E. S. Fisher (2017). Income and cancer overdiagnosis — when too much care is harmful. *New England Journal of Medicine 376*(23), 2208–2209. PMID: 28591536.

Welch, H. G. and H. J. Passow (2014, 03). Quantifying the Benefits and Harms of Mammography. *JAMA Internal Medicine 174*(3), 448–454.

Welch, H. G., P. C. Prorok, A. J. O'Malley, and B. S. Kramer (2016). Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness. *New England Journal of Medicine 375*(15), 1438–1447. PMID: 27732805.

World Health Organization (2015). Global status report on road safety 2015. http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/. Online.

Zackrisson, S., I. Andersson, L. Janzon, J. Manjer, and J. P. Garne (2006). Rate of over-diagnosis of breast cancer 15 years after end of malmö mammographic screening trial: follow-up study. *Bmj 332*(7543), 689–692.

Zahl, P., K. J. Jørgensen, and P. Gøtzsche (2013). Overestimated lead times in cancer screening has led to substantial underestimation of overdiagnosis. *British journal of cancer 109*(7), 2014.

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