Preventive Care Through a Richer Policy Space: A Machine Learning Approach to Breast Cancer Screening

N. Meltem Daysal* Sendhil Mullainathan† Ziad Obermeyer‡
Suproteem Sarkar§ Mircea Trandafir¶

December 9, 2020

We study the welfare impact and distributional effects of targeting preventive care policy through machine learning. As the ultimate value of a preventive intervention is often tied to underlying health risks, machine predictions can help individualize value measures that are typically compared using averages. Focusing on preventive mammography, which aims to find breast cancers when they are easy to treat, we train and evaluate machine learning models for predicting cancer risk. On Danish administrative data, we find that reallocating cancer screens using these models, compared to policies that raise minimum invitation ages, could catch 13% more large tumors or reduced screening by 25%. Leveraging the staggered rollout of invited mammography programs across Denmark, we also find that screening high-risk women could diminish the long-term incidence of large tumors—while reducing the overdiagnosis of cancers that would likely not become deadly.

1 Introduction

Preventive healthcare interventions can save lives, but the cost-effectiveness of these measures—which include immunizations, counseling, and cancer screens—can vary considerably (Maciosek et al., 2010). While some preventive measures, including childhood immunization, pay for themselves (Bärnighausen et al., 2014), for others the average cost of the intervention may exceed the eventual average gain. However, evaluating cost-effectiveness on averages may ignore key context about the individuals these interventions target. The value of a disease screen varies across patients (Mullainathan and Obermeyer, 2019), and those who receive preventive care may not be representative of the intended population (Lopez-Cevallos and Chi, 2010; Einav et al., 2020). Moreover, follow-on procedures as a result of preventive screening can often be delivered to those

*University of Copenhagen, meltem.daysal@econ.ku.dk
†University of Chicago, sendhil.mullainathan@chicagobooth.edu
‡University of California, Berkeley, zobermeyer@berkeley.edu
§Harvard University, suproteemsarkar@g.harvard.edu
¶University of Southern Denmark, mircea.trandafir@sam.sdu.dk

for whom the value of treatment is low (Welch et al., 2011). In many cases, the value of a preventive measure is related to underlying health risks—e.g. how likely is someone to contract cancer, develop heart disease, or experience further vision loss? While numerous factors may contribute to risk, clinicians may only be able to identify a handful to use for making screening decisions. We suggest that machine learning can help better measure these risks and improve how preventive care is targeted.

Cancer screening is a preventive intervention that aims to find potentially deadly cancers when they are easier to treat. Mammography screening is a common class of cancer screen—in 2015, more than 67% of American women over the age of 50 had a breast cancer screen in the past two years (NCHS, 2019), and more than 90% had a screen at some point in their lives (Clarke et al., 2019). However, widespread mammography screening has faced pushback for over-testing and overdiagnosis (e.g. Bleyer and Welch, 2012). Most modern breast cancer screening guidelines are predominantly based on a woman's age, an easily-measurable risk factor. In this light, many policy recommendations that attempt to address over-testing consider the relative benefits of adjusting recommended age ranges. For example, the US Preventive Services Task Force raised the minimum recommended age for screening from 40 to 50 (Nelson et al., 2009), while Einav et al. (2020) estimate the effects of raising the minimum recommended age from 40 to 45. However, these coarse changes to screening guidelines may also lead to under-testing—some of the high-risk younger women these policies drop may be precisely those women for whom it may be of high value to catch and treat cancers early. This challenge suggests that the set of policy levers considered for forming screening guidelines, which are often restricted to shifting recommended age ranges, may be too coarse.

We argue that machine learning, and its potential to provide more targeted measurements of cancer risk, may help better target these screening policies to women for whom screens are of higher value. Using Danish administrative and health data from 1990-2018, we train machine learning models on medical and prescription histories to predict the incidence of tumors that could become deadly. In Section 4, we predict the probability finding a malignant tumor through screening in a population of women universally invited to pursue mammography, and evaluate the value of targeting screens using this model. We also consider the relative value of predicting on other target variables, including invasive cancers found, large tumors found, and metastatic cancers found. Ranking women by their risk scores, we find that our algorithms can address under-testing and reduce over-testing compared to other suggested policies—including those that modify recommended age ranges. Compared to raising the minimum screening age by five years, for example, we find our models could reallocate the same number of screens and catch 13 percent more large tumors, or reduce screens by 25 percent and catch the same number of tumors. Under this age-based policy shift's implied value of catching cancers early, our model's reallocation of screens would lead to a net benefit of $64 million per million person-years, out of which $43 million would come from addressing the under-testing problem.

Challenges to this population-based approach include selection into screening and unobserved

changes in cancer risk for those not regularly screened. For example, some women living under a universal screening policy might have also elected to receive a mammogram absent a recommendation. This channel could mean our reported welfare gains from risk-score based invitations in Section 4 are too high, as some tumors might have been found even without encouragement to screen. Furthermore, the long-run impact of universal mammography might lead to fewer tumors caught at later stages for low-risk women. This effect could mean our estimated welfare gains from Section 4 are too low, as more low-risk women might develop later-stage tumors under a risk-based screening policy than under a universal screening policy. These issues call for measurement on a counterfactual, unscreened population against which we can compare health outcomes. Leveraging the staggered rollout of invited screening policies across Danish municipalities, we consider the effect of inviting risk-segmented subpopulations of women to screen on long-term population health outcomes in Section 5. We find that that invited mammography programs increased the incidence of small tumors and reduced the incidence of large tumors over time, with the large tumor reduction concentrated among those our algorithm labeled as high risk. These results suggest that targeted screening policies can reduce the number of tumors that become large, and allocate screens away from low risk women for whom screens yield more overdiagnosis

2 Background and Previous Literature

2.1 Breast Cancer Screening

Breast cancer is the most prevalent cancer in women worldwide. In 2018, an estimated 2.1 million new breast cancer diagnoses were made globally. More than 30% of these cases occurred in the European Union and the United States. In the same year 626,679 women died of breast cancer, making it the leading cause of cancer death in women throughout the world. Breast cancer is also the second most common cause of cancer death among women in the EU and the US, after lung cancer.[1]

Breast cancer screening is the medical screening of asymptomatic women for breast cancer with the aim of detecting cancers early and of improving health outcomes. The most basic form of screening includes a clinical breast exam—a physical examination of the breast by a health care provider. Although clinical breast exams were widely recommended in the past, currently the standard screening method recognized by expert organizations, including the American Cancer Society and the European Commission, is mammography. During a mammography screening, a radiologist takes low-dose X-ray pictures of a woman's breasts from one or two angles (frontal and profile) and interprets the images to form diagnoses.[2]

[1]These statistics come from the Global Cancer Observatory, owned by the International Agency for Research on Cancer. The estimates are based on the most recent data available from population-based cancer registries, the World Health Organization, or on publicly available information online. For more information see https://gco.iarc.fr/today/home.

[2]Mammography is also used as a diagnostic tool. Diagnostic mammograms are generally offered to women with

Virtually all developed countries have policies concerning coverage of mammography screening. In Europe, mammography screening is provided through organized screening programs (Altobelli and Lattanzi, 2014). During the 1980s and 1990s, mammography screening was offered by a handful of EU Member States—primarily at the local level. The increase in organized screening programs followed the European Commission's 2003 recommendation to implement population-based nationwide screening for women aged 50 to 69 years old. By 2014, 22 of the then 28 EU Member States had established population-based nationwide breast cancer screening programs. Although there are national variations in the details, these programs share key features across target ages, screening intervals, and screening methods. In a typical program, women between the ages of 50 and 69 are sent an invitation letter every two years to receive mammography screening free of charge. Most countries rely on digital mammography and use double reading of normal mammograms.

The United States does not have a universal breast cancer screening program. Instead, mammography screening is tied to insurance status. For women with private health insurance, subsidized breast cancer screening was initially implemented through state laws that required private health insurance plans to include screening mammograms as a covered benefit (Bitler and Carpenter, 2016). Even though more than 42 states enacted these laws between 1987 and 2000, there was substantial variation in target age groups across states, as well as in the frequency of screenings. The 2010 Affordable Care Act streamlined these differences through a national policy where all private insurance plans are required to cover the full cost of screening mammograms for eligible women. Mammography screening for uninsured women is primarily organized through the National Breast Cancer and Cervical Cancer Early Detection Program, a federal program that provided earmarked funds to states to provide cancer screening to uninsured low-income women. The program was rolled out across states between 1991 and 1999. The current guidelines by the US Preventive Services Task Force recommends biennial screening mammography for women aged 50 to 74 years.

These widespread screening recommendations have recently come under scrutiny because of the potential harms that may outweigh the benefits. In particular, mammography screening can result in high false-positive rates and overdiagnosis, defined as the diagnosis of a cancer that would not have been detected during the remaining lifetime of the woman in the absence of screening (Bleyer and Welch, 2012; Løberg et al., 2015; Welch et al., 2016). Recent reviews put the cumulative rate of false positives over the recommended screening age range at around 20% in Europe and 30% in the US (Hofvind et al., 2012; Hubbard et al., 2011). The overdiagnosis rate is more difficult to calculate in observational studies because it depends on assumptions about the evolution of detected cancers in the absence of screening. The current best estimates come from two randomized controlled trials: a 25-year follow-up study using Canadian data that finds an overdiagnosis rate of 22% from mammography screening and breast exams as compared with

previous abnormal screening mammograms or with family histories of breast cancer, as well as in the presence of certain symptoms (e.g., lumps in the breast, changes in the breast structure or the nipple).

physical breast exams only (Miller et al., 2014), and a 15-year follow-up study using Swedish data that finds an overdiagnosis rate of 24% (Welch et al., 2006, who correct the calculations in Zackrisson et al., 2006).

In response to these concerns, there are now global calls for changes to breast cancer screening policies. Academic researchers and policymakers have proposed and examined the effects of various such changes to current practices. For example, recent academic work studies proposals to raise the eligibility age for screening recommendations (e.g., Einav et al., 2020). In a more radical approach, the members of the Swiss Medical Board, an agency that assesses medical cost-effectiveness, contemplate the effects of abolishing mammography programs altogether (Biller-Andorno and Jüni, 2014). As a middle-of-the-road alternative, the United States Preventive Services Task Force no longer recommends regular mammography for women between the ages of 40 and 49, and instead advises that "the decision to start screening mammography in women prior to age 50 years should be an individual one" (Siu, 2016).

2.2 Breast Cancer Screening in Denmark

Since the establishment of the National Cancer Register in 1943, breast cancer remains as the most common cancer among women in Denmark (Olsen et al., 2003). Despite this fact, Denmark has been slow to introduce breast cancer screening programs. A number of early adopters of screening programs in the European Member States conducted randomized control trials on the effectiveness of mammography screening during the 1980s. Following the results of these trials, which suggested that screening could reduce breast cancer mortality, Denmark started a number of regional screening programs during the 1990s.

The first program started in the municipality of Copenhagen in 1991, followed by Funen County in 1993, and the Frederiksberg municipality in 1994.[3] Across all regional programs, the target group consisted of women aged 50 to 69 (approximately 20 percent of Danish women in the targeted age range) and the screening interval was two years. Eligible women in the target age range received an invitation that included information about the screening program in general, how a mammography screening is conducted, and a (changeable) date for screening.[4] Women born in January received a date in the first two months of the invitation round, women born in February were assigned a date in the next two months, and so on (Jacobsen et al., 2017). Those who did not respond to the first invitation received two reminders or another invitation (Lynge et al., 2017). The protocol for these early programs, until 2001, included two-view mammography at the first examination, followed by one-view (for women with fatty breast tissue) or a two-view (for women with mixed/dense tissue) mammography in future invitation rounds. Starting in 2001,

[3] The Frederiksberg screening program was merged with the Copenhagen program in 1996.

[4] The screening eligibility criteria differed between Copenhagen and the Funen county. In Copenhagen, initially only women who were treated for breast cancer were deemed ineligible for screening. As a result, 95% of the women in the target age range were sent invitations. In Funen, ineligible women included those who were diagnosed with breast cancer in the last five years as well as women with a prior benign breast lesion. Based on these, the share of ineligible women in the target group was 12%. In both settings, women who wanted to opt out of screening were excluded from future invitation rounds (Domingo et al., 2013).

all examinations included two-view mammography. The screening method was based on analog mammography until 2006, when the programs switched to digital mammography. The images were always read by two radiologists.

Following the EU recommendations in 2003 and the results of a study that showed reductions in breast cancer mortality in Copenhagen after the introduction of the screening program (Olsen et al., 2005), the Danish Ministry of Health ruled that all regions start screening programs. The universal nationwide screening began in 2007 and the rollout across the Danish regions was complete by the end of the decade.[5] The national screening program adopted all the main features of the regional programs (Lynge et al., 2017).

Key considerations in screening programs are the coverage and participation rates.[6] According to the European guidelines for quality assurance, screening programs should have a participation rate of at least 70% in order to be rated as acceptable while rates over 75% are desirable (Euler-Chelpin et al., 2006). Several independent studies examined coverage and participation in the Danish screening programs. Those investigating the regional programs find that the coverage rate in Copenhagen and Frederiksberg started at around 70% in the first invitation round and declined to 65% by the fourth invitation round, but participation remained stable at around 70% (Euler-Chelpin et al., 2006; Jacobsen et al., 2017). Coverage and participation rates were considerably higher in the county of Funen: coverage started at around 85% in the first round and was still at around 83% in the fourth round, while participation increased from 85% in the first round to almost 94% in the fourth round (Euler-Chelpin et al., 2006; Jacobsen et al., 2017). During the first four rounds of the nationwide screening program, national coverage remained relatively stable at 75–77% but coverage was again lower in the Capital Region (Lynge et al., 2017). Existing evidence suggests that opportunistic screening is rare in Denmark and that the difference in coverage rates between Copenhagen (Capital Region) and other regions cannot be explained by differences in opportunistic screening. Based on data on all diagnostic mammographies performed in Denmark in 2000, Jensen et al. (2005) find that only 3% of women aged 50 to 69 used diagnostic mammography and that take-up of opportunistic screening was not higher among non-participants of organized screening programs (3% in Copenhagen and 1% in Funen).

Despite the widespread incidence of these programs, there is controversy in both the medical community and in public debates concerning the effectiveness of screening programs. Existing concerns mirror the global discussions on false positives and overdiagnosis. Recent evidence suggests that the cumulative false-positive risk after eight completed screens was around 16% in Copenhagen and 7% in Funen (Jacobsen et al., 2015). These estimates are much lower than those found in the US but fall within the range of false positive estimates found in other European countries with similar organized breast cancer screening programs (Hovfind et al., 2004; Puliti et al., 2011; Salas et al., 2011). Estimates of overdiagnosis, on the other hand, largely depend on

[5] Appendix Table X shows the start and end dates corresponding to the first round of the national screening program across different regions.

[6] Coverage is defined as ratio of the number of women screened to the number of women in the target group. Participation refers to the share of screened women among those who are invited for screening.

the methods used. Previous studies based on individually linked data from cohorts of women invited to screening tend to find overdiagnosis rates of 2-3% (Njor et al., 2013). Studies relying on data from fixed age-groups, however, find overdiagnosis rates ranging from 30% to close to 50% (Jørgensen et al., 2009, 2017).

3 Data

We use national administrative data from Denmark over the period 1990–2018. Information on variables related to screening comes from three separate mammography registers: one for the screening program in the Copenhagen and Frederiksberg regions, covering the period 1991–2008; one for the screening program in the Funen region, covering the period 1993–2005; and the Danish Quality Database of Mammography Screening, which covers the nationwide screening program and the period after 2008 (Mikkelsen et al., 2016). Using individual identification numbers, we are able to add information from several other sources of administrative data. First, we obtain from the Danish Breast Cancer Cooperative Group data set, which covers the near-universe of invasive breast cancer diagnoses since 1977, breast cancer diagnoses and detailed data on clinical characteristics (tumor size, cancer stage, number of tumor-positive lymph nodes, etc.), date and type of surgery (mastectomy or lumpectomy), location, date and type of treatments provided (radiation therapy, chemotherapy, or hormone therapy), and follow-up up to five years post diagnosis.

Additional medical history can be obtained from four other registers. The National Patient Register records all hospital visits since 1977 with detailed information such as the hospital and department identifiers, exact date of the visit, type of visit (outpatient, inpatient, or emergency room), date of discharge if applicable, main diagnosis, and date and type of any procedure performed. The National Health Service Register tracks all visits since 1990 to private practitioners that are covered by the public health insurance plan. These include all visits to general practitioners (GP) and to specialists when referred by a GP. The register provides the unique identification number of the practice, the specialty of the physician, the type and cost of the service provided, and the date when the practice submitted the payment request to the national health insurance.[7] The National Prescription Drug Register documents, for all the prescription filled in Denmark since 1995, the personal identification number of the person filling the prescription, the date when the prescription was filled, the identification number of the practice that issued the prescription, the Anatomical Therapeutic Chemical Classification (ATC) code of the drug, the quantity, and its price. Finally, the Causes of Death Register records the exact date, manner, main cause, and up to three contributing causes for all deaths since 1970.

We augment the health data with socioeconomic characteristics from several other registers that cover the entire (relevant) resident population. The Danish Population Register provides a snapshot of the population on January 1, starting from 1980, detailing for each person their age,

[7] While this is not the actual date when the service is provided, Statistics Denmark indicates that it is reasonable to assume that the two dates are relatively close.

municipality of residence, marital status (including the identification number of the spouse, if applicable), immigration status, and the personal identification numbers of the parents for persons born after 1960. Similarly, the Population Education Register provides a yearly snapshot on December 31 of the highest level of education completed and its usual duration in months from 1980 onward. The Labor Force Register describes the position in the labor force at the end of November of each year starting from 1980, as well as the type of employment, occupation, and industry, if applicable. The Income Register includes data provided by the national tax authority on total annual income and its sources (labor income, government transfers, etc.) starting from 1980. The DREAM data base records, for all the persons who received a social benefit or any other type of government transfer since July 1991, the year and week during which they received a government transfer for at least one day, and the type of transfer.

Our sample is constructed as follows. Over the period 1990–2018, we identify in each year all women aged 50-69 who are currently covered or who would have been covered by the current Danish screening invitation guidelines. Our data frame covers a total of 15,830,794 person-years and 1,679,450 women. Out of these observations, 2,387,864 occurred starting in 2009, when mammography screening was instituted nationwide. Using invitation response data from the Danish Quality Database of Mammography Screening, we identify 1,926,679 observations, representing 756,317 women, for which these invitations to screen were accepted. Based on the invasive cancer data from the Danish Breast Cancer Cooperative Group (Christiansen et al., 2016), we also identify the invasive cancer status and cancer characteristics for each of these women. The incidence of invasive cancers found across our entire sample was 289 per 100,000 person-years. Table 1 describes the composition of our data across these categories.

4 Screen Allocation

4.1 Model Training and Comparison

While the ultimate evaluation metric for our approach is a welfare comparison, it is helpful to understand the prediction problem that feeds into these welfare calculations. Cancer is a rare event—the incidence of invasive cancer across women aged 50-69 in our sample is 289 per 100,000 person-years. As a positive result is rare, to evaluate our model's predictive power we consider the area under the curve of the receiver operating characteristic curve (AUC-ROC). [8] We develop our cancer risk models on a sample of 1,926,679 observations of women who received and accepted invitations into screening mammography from 2010-2017. We train gradient-boosted trees and penalized logistic regression models using medical information available at the time of invitation on a randomly selected training set, grouped at the individual level to avoid individual history memorization. We compare model predictions against realized cancer outcomes on a separate

[8] Raw accuracy scores are not the best metrics for assessing model performance when cancer is a rare event. A classifier that predicts no cancer for all women would achieve an accuracy score of 99.7 percent, but would not be helpful for segmenting women by breast cancer risk.

holdout set.

Large tumors, or tumors found to be $\geq$2cm at the time of biopsy, are documented in the medical literature to coincide with mortality rates (e.g. Welch et al., 2016). For the prediction problem of finding large tumors through screening, we achieve an AUC of .712 using gradient-boosted trees and an AUC of .625 using penalized logistic regression. We achieve similar performance using gradient-boosted trees for the prediction tasks of finding malignant tumors (.713 AUC) and any invasive cancer (.717 AUC). Furthermore, these models yield comparable predictions when evaluated against different prediction labels—for each model among the large tumor, malignant tumor, and invasive cancer predictors, test performance declines by a maximum of .020 AUC when evaluated on one of the other two labels. The performance of the gradient-boosted model and logistic model on predicting large tumor incidence is compared in Figure 1. For the women ranked in the top risk decile for each model, the rate of cancer found is 2.43 times higher than the population mean for the gradient-boosted model and 1.86 times higher than the population mean for the penalized logistic regression model.

4.2 Targeting Screening Policy through a Richer Space

We evaluate the yield of allocating screens using risk models on the holdout set of women invited to screen starting in 2010. Under a constant cost to the health system of $100 per mammogram, Figure 2 shows the number of malignant tumors that would have been caught by ranking women by algorithm-based risk and age-based risk for each dollar of mammography conducted. Compared to increasing the minimum screening age by five years, allocating screens using the algorithm's ranking could catch 13% more malignant tumors or reduced screening by 25%. Figure 3 shows how the algorithm would reallocate screens compared to raising the minimum age bt five years, and the sources of value from reallocation across several valuations of catching a cancer early. Under the value implied by raising the minimum age by five years, the model's reallocation would lead to a net benefit of $64 million per million person-years, out of which $43 million would come from addressing under-testing.

5 High-Value Screens

5.1 Long-term Impact of Mammography

While our previous analyses have been static, a key goal of preventive cancer screening is to catch cancers that could eventually become deadly. This suggests that an effective screening policy would decrease the incidence of more developed cancers over time. Leveraging the staggered rollout of invited mammography programs for women aged 50-69 across municipalities, we measure cancer incidence over time. We run differences-in-differences regressions, for individual i

living in municipality m at time t, of the form

$$Y_{imt} = \tau \times \mathbb{1}(\text{invited_screening})_{it} + \alpha_m + \lambda_t + \varepsilon_{imt}$$

for $Y_{it} = \mathbb{1}(\text{large_tumor})_{it}$ and $Y_{it} = \mathbb{1}(\text{small_tumor})_{it}$. Results from this regression for each outcome variable are reported as (1) and (3) in Table 2. As treatment is staggered in this policy setting, it may be more appropriate to run event study regressions that account for leads and lags of treatment and potential challenges from treatment effect heterogeneity (see Callaway and Sant'Anna, 2019; Sun and Abraham, 2018). Following the stacking procedure in Novgorodsky and Setzler (2019), we run event study versions of the previous differences-in-differences specification accounting for treatment timing and overlap, and present these results in Figure 4. In the ten years following invited mammography, large tumor incidence cumulatively decreased by 423 tumors per million person-years, while small tumor incidence cumulatively increased by 901 tumors per million person-years. That these two differences do not offset one-for-one suggests that there was some overdiagnosis arising from invited screening policies, resulting in small tumors found that might not have developed into larger tumors.

5.2 Risk Segmentation and Screen Value

To better understand the relationship between invited mammography, tumor size evolution, and overdiagnosis, we add an additional term accounting for predicted risk to the specification above. Labeling women with risk scores above the median as high risk, and those below as low risk, we run regressions of the form

$$\begin{aligned} Y_{imt} =& \beta \times \mathbb{1}(\text{invited_screening})_{it} \times \mathbb{1}(\text{high_risk})_{it} \\ &+ \pi \times \mathbb{1}(\text{high_risk})_{it} + \tau \times \mathbb{1}(\text{invited_screening})_{it} + \alpha_m + \lambda_t + \varepsilon_{imt} \end{aligned}$$

for $Y_{it} = \mathbb{1}(\text{large_tumor})_{it}$ and $Y_{it} = \mathbb{1}(\text{small_tumor})_{it}$. Results from this regression for each outcome variable are reported as (2) and (4) in Table 2. Following the previous section, to account for staggered treatment timing and treatment overlap, we separate our dataframe into high-risk and low-risk subsets and run stacked event study regressions. Results from these event studies are reported in Figure 5.

While overdiagnosis appears to be present across across the entire population of women affected by the policy, Figure 5 suggests it may have been less of a concern for screens targeted to women deemed to be at higher risk. For women predicted to be low-risk, small tumors found outpace large tumors avoided following the introduction of invited screening. However, for women predicted to be high-risk, the difference between small tumors found and large tumors avoided is smaller. This suggests that in addition to reducing the incidence of large tumors, allocating screens by risk profiles might also help reduce overdiagnosis.

5.3 Screening and Mortality

While large and small tumors provide nearer-term measures of severe cancer incidence that coincide with eventual disease severity and mortality, it is also helpful to understand the long-run effects of screening and catching cancers early on mortality. Two potential mortality outcomes to consider are all-cause and breast cancer-specific mortality—all-cause proponents argue the measure is less biased, while cancer-specific adherents contend all-cause mortality is too restrictive a measure (see Penston, 2011; Steele and Brewster, 2011; Heijnsdijk et al., 2019). Potential reporting biases may affect the fidelity of breast cancer mortality measures as a benchmark in Denmark, though variation in other sources of mortality may weaken the power of all-cause mortality as an outcomes. We report results from both measures using the specification from Section 5.1—with $Y_{it} = \mathbb{1}(\text{all_cause_mortality})$ in (5) and $Y_{it} = \mathbb{1}(\text{breast_cancer_mortality})$ in (6) in Table 3. We also report the risk-interacted results from the specification in Section 5.2 for all-cause (7) and breast cancer (8) mortality in Table 3. The effect of the interaction term between invited mammography and high risk on mortality is negative, suggesting high-risk patients might also have had greater mortality reductions after being invited to screen. This coefficient is statistically significant at $\alpha = 0.05$ for breast cancer mortality, but is not so for all-cause mortality, echoing the discussion of Heijnsdijk et al. (2019) on the additional power required to more precisely measure all-cause effects.

6 Discussion

While the medical and health economics literatures often evaluate health interventions in terms of average effects, the value of care can differ across individuals. In preventive care, we argue that value is closely tied to underlying health risk, and that machine learning can help measure these risks. Using Danish administrative and medical data, we train models to predict cancer risk that could allocate screening more efficiently than policies that would target screening through age-based measures of risk. To address issues of selection into screening and differential evolution of cancer, we leverage the variation in screening decisions from the staggered rollout of mammography invitations across Denmark. We find that targeting screens through risk measures could increase the efficiency of finding deadly tumors and reduce overdiagnosis. More broadly, these results suggest that the set of existing policy levers used to evaluate health interventions may be too narrow—more expressive measures of risk and predicted changes to health outcomes could help target care to those who would benefit from it most.

References

Bärnighausen, Till, David E Bloom, Elizabeth T Cafiero-Fonseca, and Jennifer Carroll O'Brien, "Valuing vaccination," *Proceedings of the National Academy of Sciences*, 2014, *111* (34), 12313–12319.

Biller-Andorno, Nikola and Peter Jüni, "Abolishing mammography screening programs? A view from the Swiss Medical Board," *Obstetrical & Gynecological Survey*, 2014, *69* (8), 474–475.

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

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Available at SSRN 3148250*, 2019.

Christiansen, Peer, Bent Ejlertsen, Maj-Britt Jensen, and Henning Mouridsen, "Danish breast cancer cooperative group," *Clinical epidemiology*, 2016, *8*, 445.

Clarke, Tainya C, Meheret Endeshaw, Denise Duran, and Mona Saraiya, "Breast cancer screening among women by nativity, birthplace, and length of time in the United States," 2019.

Einav, Liran, Amy Finkelstein, Tamar Oostrom, Abigail Ostriker, and Heidi Williams, "Screening and Selection: The Case of Mammograms," *American Economic Review*, December 2020, *110* (12), 3836–3870.

Heijnsdijk, Eveline AM, Marcell Csanádi, Andrea Gini, Kevin ten Haaf, Rita Bendes, Ahti Anttila, Carlo Senore, and Harry J de Koning, "All-cause mortality versus cancer-specific mortality as outcome in cancer screening trials: A review and modeling study," *Cancer medicine*, 2019, *8* (13), 6127–6138.

Hofvind, Solveig, Antonio Ponti, Julietta Patnick, Nieves Ascunce, Sisse Njor, Mireille Broeders, Livia Giordano, Alfonso Frigerio, and Sven Törnberg, "False-Positive Results in Mammographic Screening for Breast Cancer in {Europe}: A Literature Review and Survey of Service Screening Programmes," *Journal of Medical Screening*, September 2012, *19* (1_suppl), 57–66.

Hubbard, Rebecca A., Karla Kerlikowske, Chris I. Flowers, Bonnie C. Yankaskas, Weiwei Zhu, and Diana L. Miglioretti, "Cumulative Probability of False-Positive Recall or Biopsy Recommendation after 10 Years of Screening Mammography," *Annals of internal medicine*, October 2011, *155* (8), 481–492.

Lopez-Cevallos, Daniel F and Chunhuei Chi, "Health care utilization in Ecuador: a multilevel analysis of socio-economic determinants and inequality issues," *Health policy and planning*, 2010, *25* (3), 209–218.

Lynge, Elsebeth, Martin Bak, My von Euler-Chelpin, Niels Kroman, Anders Lernevall, Nikolaj Borg Mogensen, Walter Schwartz, Adam Jan Wronecki, and Ilse Vejborg, "Outcome of breast cancer screening in Denmark," *BMC cancer*, 2017, *17* (1), 897.

Løberg, Magnus, Mette Lise Lousdal, Michael Bretthauer, and Mette Kalager, "Benefits and Harms of Mammography Screening," *Breast Cancer Research*, May 2015, *17* (1), 63.

Maciosek, Michael V, Ashley B Coffield, Thomas J Flottemesch, Nichol M Edwards, and Leif I Solberg, "Greater use of preventive services in US health care could save lives at little or no cost," *Health Affairs*, 2010, *29* (9), 1656–1660.

Mikkelsen, Ellen M, Sisse H Njor, and Ilse Vejborg, "Danish quality database for mammography screening," *Clinical epidemiology*, 2016, *8*, 661.

Miller, Anthony B., Claus Wall, Cornelia J. Baines, Ping Sun, Teresa To, and Steven A. Narod, "Twenty Five Year Follow-up for Breast Cancer Incidence and Mortality of the Canadian National Breast Screening Study: Randomised Screening Trial," *BMJ*, February 2014, *348*, g366.

Mullainathan, Sendhil and Ziad Obermeyer, "A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions," *NBER Working Paper*, 2019, *26168.*

NCHS, "Table 70. use of mammography among women aged 40 and over, by selected characteristics: United States, selected years 1987–2015," 2019.

Nelson, Heidi D, Kari Tyne, Arpana Naik, Christina Bougatsos, Benjamin K Chan, and Linda Humphrey, "Screening for breast cancer: an update for the US Preventive Services Task Force," *Annals of internal medicine*, 2009, *151* (10), 727–737.

Novgorodsky, David and Bradley Setzler, "Practical Guide to Event Studies," 2019.

Olsen, AH, A Jensen, Sisse Helle Njor, E Villadsen, W Schwartz, I Vejborg, and Elsebeth Lynge, "Breast cancer incidence after the start of mammography screening in Denmark," *British Journal of Cancer*, 2003, *88* (3), 362–365.

Olsen, Anne Helene, Sisse H Njor, Ilse Vejborg, Walter Schwartz, Peter Dalgaard, Maj-Britt Jensen, Ulla Brix Tange, Mogens Blichert-Toft, Fritz Rank, Henning Mouridsen et al., "Breast cancer mortality in Copenhagen after introduction of mammography screening: cohort study," *Bmj*, 2005, *330* (7485), 220.

Penston, James, "Should we use total mortality rather than cancer specific mortality to judge cancer screening programmes? Yes," *Bmj*, 2011, *343.*

Siu, Albert L, "Screening for breast cancer: US Preventive Services Task Force recommendation statement," *Annals of internal medicine*, 2016, *164* (4), 279–296.

Steele, Robert JC and David H Brewster, "Should we use total mortality rather than cancer specific mortality to judge cancer screening programmes? No," *Bmj*, 2011, *343*, d6397.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *arXiv e-prints*, 2018, pp. arXiv–1804.

Welch, H. Gilbert, Lisa M. Schwartz, and Steven Woloshin, "Ramifications of Screening for Breast Cancer: 1 in 4 Cancers Detected by Mammography Are Pseudocancers," *BMJ*, March 2006, *332* (7543), 727.

Welch, H Gilbert, Lisa Schwartz, and Steve Woloshin, *Overdiagnosed: making people sick in the pursuit of health*, Beacon Press, 2011.

_ , Philip C Prorok, A James O'Malley, and Barnett S Kramer, "Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness," *New England Journal of Medicine*, 2016, *375* (15), 1438–1447.

Zackrisson, Sophia, Ingvar Andersson, Lars Janzon, Jonas Manjer, and Jens Peter Garne, "Rate of Over-Diagnosis of Breast Cancer 15 Years after End of Malmö Mammographic Screening Trial: Follow-up Study," *BMJ*, March 2006, *332* (7543), 689–692.

Table 1: Summary statistics for our full dataset. Observations from 1986-2007 are reported in the "Policy Rollout Sample," and further broken down by whether the municipality had an invitation mammography policy or an elective screening policy. Observations from 2010-2017 are reported in the "Universal Screening Sample," and further broken down by whether a mammography invitation was accepted by the invited woman.

	Full Sample	Policy Rollout Sample		Universal Screening Sample	
		Invited Screening	Elective Screening	Accepted Invitation	Declined Invitation
n observations	15,830,795	1,879,355	11,563,577	1,926,679	461,184
n women	1,679,450	310,533	1,222,507	756,317	294,817
Age	59.3 (6.0)	59.1 (6.1)	59.2 (6.0)	59.7 (5.9)	59.5 (6.1)
Invasive Cancer Found (per 100,000 obs)	288.6	352.8	254.3	399.0	426.0
Large Tumor Found (per 100,000 obs)	124.5	114.3	131.8	81.6	162.4
Lymph Node Spread (per 100,000 obs)	128.6	117.7	130.5	117.5	171.9

Figure 1: This figure reports the calibration performance for gradient boosting and penalized logistic regression models used to predict large (≥2cm) tumor incidence. The figure describes the rate of cancer found versus the population mean across risk quantiles for each model.

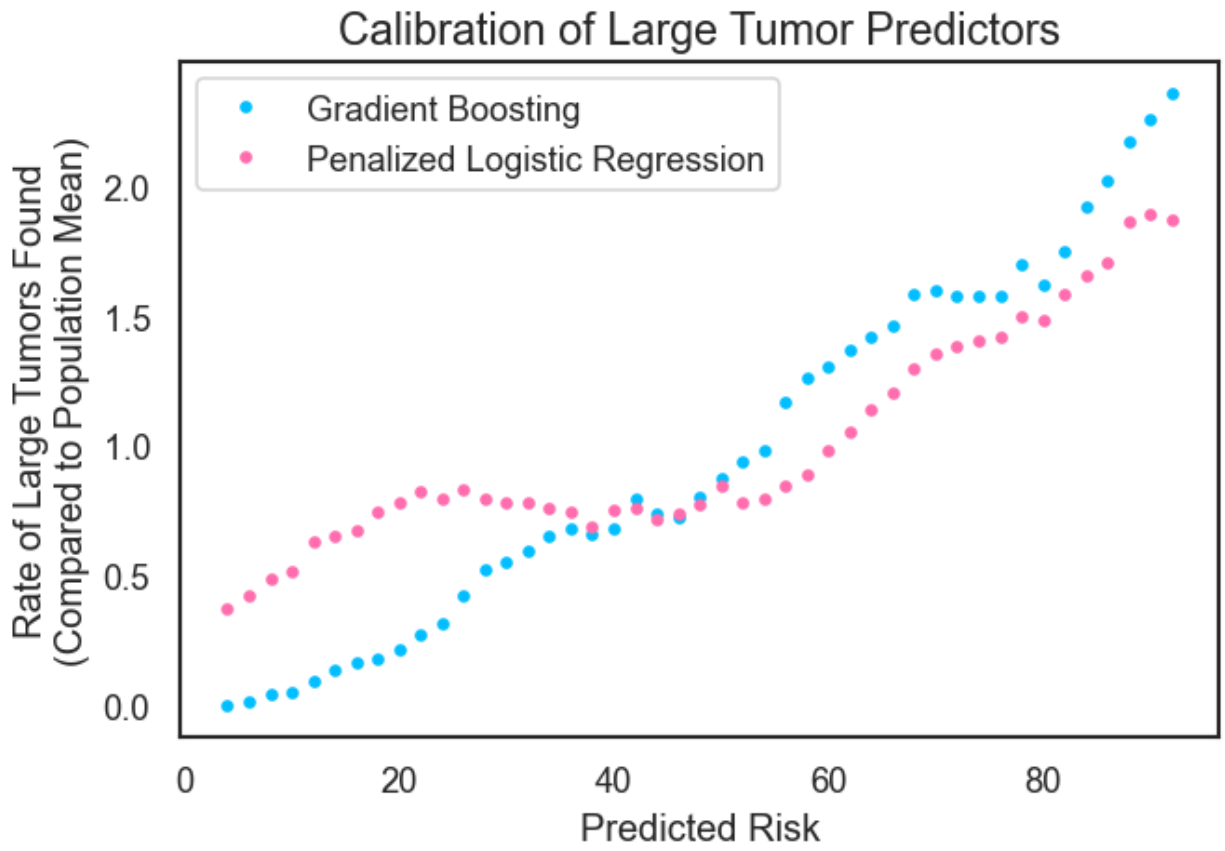


Figure 2: This figure reports the yield from screening for gradient boosting and age-based rankings of risk. The figure describes the number of large tumors caught in a hold-out sample after conducting mammography according to relative risk rankings. Costs are calculated as a fixed $100 per mammogram. The dashed green line reflects the yield of screening from a random allocation, with no segmentation by risk.

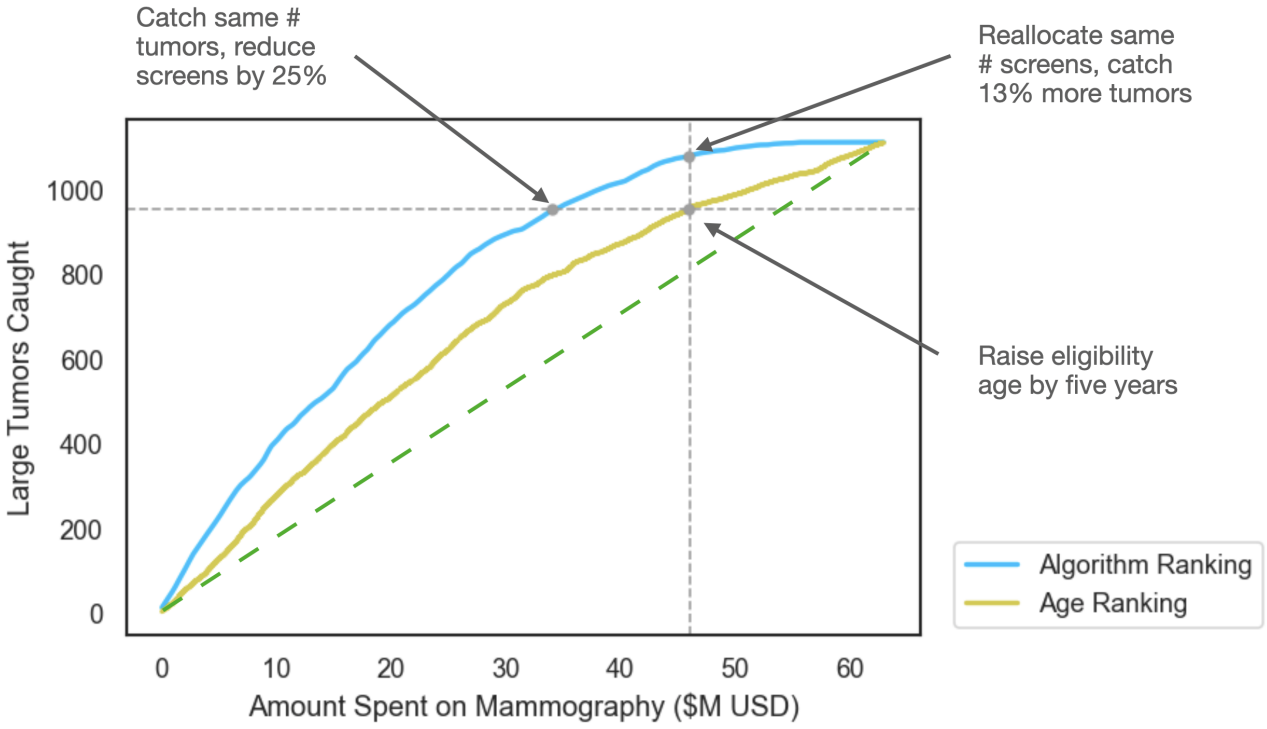
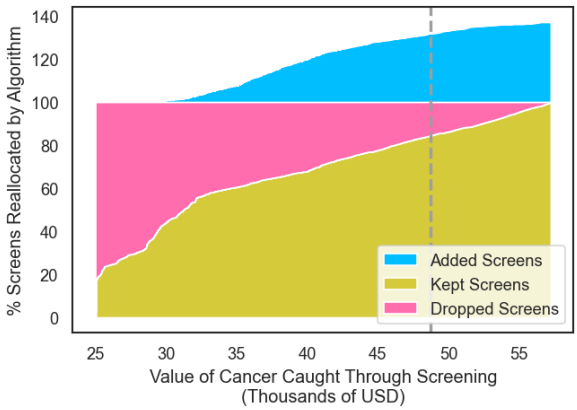


Figure 3: These figures report the reallocation of screens and the sources of value from model-based rankings, as compared to a counterfactual policy that would raise the minimum invited screening age by five years. The dashed gray line shows the implied value of catching a cancer through screening from raising the minimum eligibility age by five years.



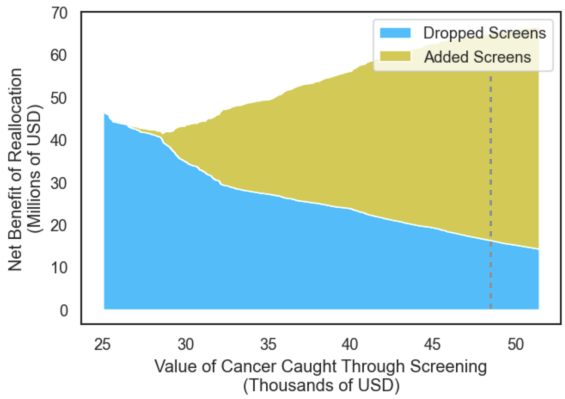
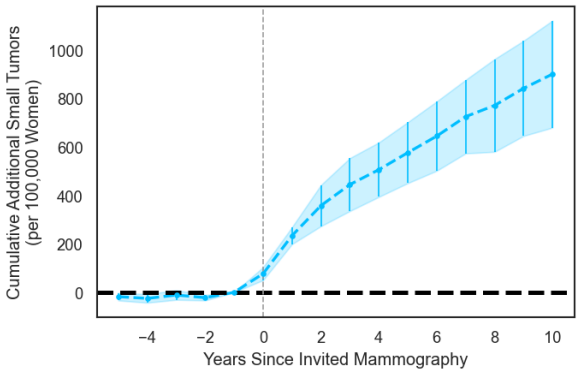


Figure 4: This figure reports cumulative change in tumors caught (per 100,000 women) in the years following invited screening in Danish municipalities, across large and small tumors. The increase in small tumors found coincided with the decrease in large tumors found. However, these two did not substitute on-for-one, suggesting overdiagnosis.

(a) Additional small tumors found after invited screening.



(b) Additional large tumors found after invited screening.

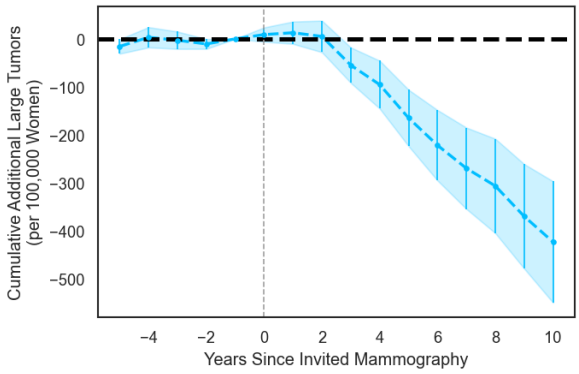
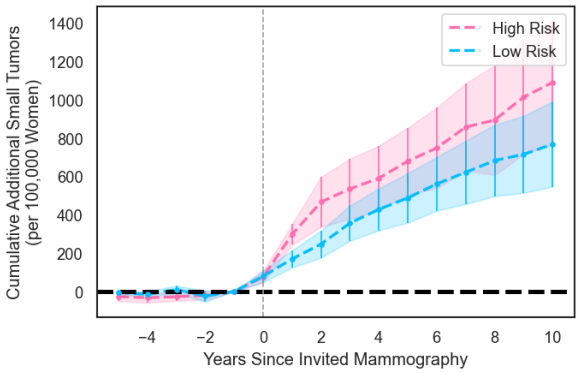


Table 2: This table reports differences-in-differences estimates of the effect of invited mammography policies on tumor incidence. All regressions use year and municipality fixed effects, and standard errors are clustered at the municipality level (the level of the policy rollout). (2) and (4) also include interaction terms between an invitation to screen and having a risk score higher than the median risk estimated from our models.

	(1)	(2)	(3)	(4)
	Large Tumors	Large Tumors	Small Tumors	Small Tumors
Invited Mammography	-36.0	-23.2	91.8	65.9
	(6.1)	(7.5)	(4.2)	(4.0)
High Risk		33.6		15.9
		(0.7)		(1.1)
Invited Mammography x High Risk		-23.1		54.1
		(5.2)		(3.9)
Intercept	128.3	109.8	63.9	54.6
	(3.1)	(2.8)	(1.4)	(1.3)
Adj. R^2	9.90×10^{-6}	3.02×10^{-5}	1.69×10^{-4}	1.87×10^{-4}
N	13,442,932	13,442,932	13,442,932	13,442,932

Figure 5: This figure reports the cumulative change in tumors caught (per 100,000 women) in the years following invited screening in Danish municipalities, segmented by the algorithm's prediction of high and low risk. While the additional small tumors found after invited screening do not substitute one-for-one into additional large tumors found across either risk category (indicating overdiagnosis), the drop-off in large tumors found is larger for the high risk group.

(a) Additional small tumors found after invited screening, by risk category.



(b) Additional large tumors found after invited screening, by risk category.

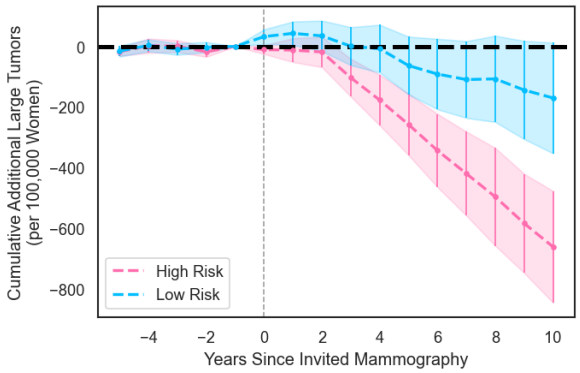


Table 3: This table reports differences-in-differences estimates of the effect of invited mammography policies on mortality. All regressions use year and municipality fixed effects, and standard errors are clustered at the municipality level (the level of the policy rollout). (2) and (4) also include interaction terms between an invitation to screen and having a risk score higher than the median risk estimated from our models.

	(5) All-Cause Mortality	(6) All-Cause Mortality	(7) Cancer Mortality	(8) Cancer Mortality
Invited Mammography	-54.7	0.4	-12.8	-3.8
	(57.4)	(105.1)	(7.6)	(8.2)
High Risk		658.0		-19.0
		(12.0)		(3.7)
Invited Mammography x High Risk		-57.3		-19.8
		(125.1)		(3.9)
Intercept	1087.7	707.8	103.4	114.4
	(17.5)	(16.5)	(2.7)	(3.1)
Adj. R^2	3.40×10^{-4}	1.51×10^{-3}	1.17×10^{-4}	1.43×10^{-4}
N	13,442,932	13,442,932	13,442,932	13,442,932