An Economic Approach to Machine Learning in Health Policy

N. Meltem Daysal Sendhil Mullainathan Ziad Obermeyer
Suproteem K. Sarkar Mircea Trandafir

January 31, 2024

We consider the health effects of "precision" screening policies for cancer guided by algorithms. We show that machine learning models that predict breast cancer from health claims data outperform models based on just age and established risk factors. We estimate that screening women with high predicted risk of invasive tumors would reduce the long-run incidence of later-stage tumors by 40%. Screening high-risk women would also lead to half the rate of cancer overdiagnosis that screening low-risk women would. We show that these results depend crucially on the machine learning model's prediction target. A model trained to predict positive mammography results leads to policies with weaker health effects and higher rates of overdiagnosis than a model trained to predict invasive tumors.

For valuable comments, we thank Jiafeng Chen, Amy Finkelstein, Salome Aguilar Llanes, Dominic Russel, Andrei Shleifer, Jennifer Walsh, and seminar participants at Harvard University, the NBER Machine Learning in Healthcare conference and the Responsible Machine Learning in Healthcare conference. Daysal gratefully acknowledges financial support from the Danish Council for Independent Research (grant number 4182-00214). Mullainathan gratefully acknowledges support from the Center for Applied Artificial Intelligence at the University of Chicago. Sarkar gratefully acknowledges support from a National Science Foundation Graduate Research Fellowship under grant number DGE-2140743. Daysal: University of Copenhagen, CEBI, CESIfo, IZA (meltem.daysal@econ.ku.dk); Mullainathan: University of Chicago Booth School of Business (sendhil.mullainathan@chicagobooth.edu); Obermeyer: University of California, Berkeley (zobermeyer@berkeley.edu); Sarkar: Harvard University (suproteemsarkar@g.harvard.edu); Trandafir: The Rockwool Foundation Research Unit (mt@rff.dk).

1. Introduction

Preventive screening is a potentially powerful medical intervention against cancer. A timely mammogram, for example, could catch a deadly tumor before it grows—possibly saving a life and preventing more expensive later-stage treatments. The same mammogram, however, could identify a tumor that would have never grown enough to cause health issues. This could lead to personal trauma, needless biopsies, and potentially harsh treatments that could have been avoided. Weighing the costs and benefits of these outcomes brings us to central question of preventive intervention: Who should be screened? The most common approach to this question is to target screening through guidelines.[1] Several countries have developed age-based cancer screening programs, as age is often the best predictor of any cancer. Empirical data on cancer incidence is often used to determine these age-based criteria.[2]

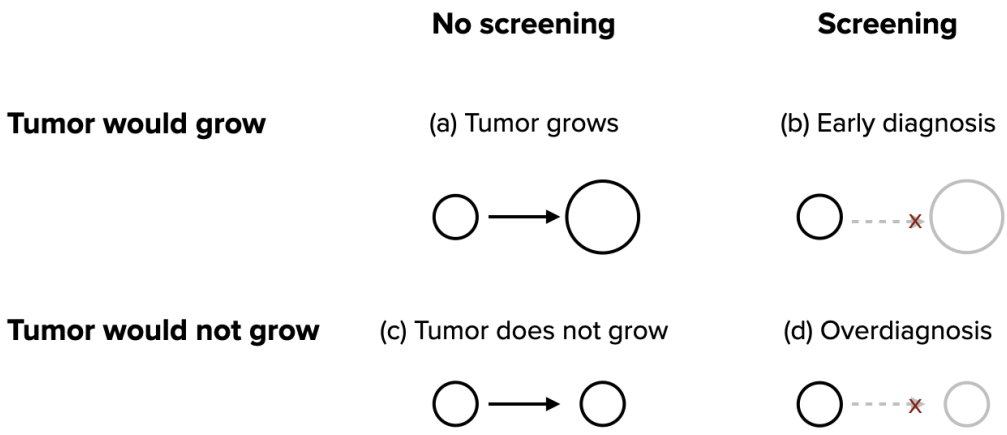
The same cancer incidence data used to set age-based criteria, though, can now be leveraged differently. The question of who to screen is a prediction policy problem (Kleinberg et al., 2015). We would like to screen those at high risk of cancer. As a result, we can potentially construct "precision" screening policies (e.g. Marcus et al., 2016; Conner et al., 2022) by building richer models of cancer risk. In this paper, we conduct such an exercise on Danish administrative data and evaluate its potential health effects.

The health effects of a cancer screening policy set through machine learning depend on the policy's ability to target women whose health would improve from a screen. Each cancer screen can improve health by detecting a tumor early, when

[1] Smith et al. (2019) reviews cancer screening guidelines and evidence on the effects of screening across cancer types. Kowalski (2021) reviews and analyzes recent evidence on breast cancer screening.

[2] The core empirical challenge in this literature is to estimate the effects of screening guidelines when counterfactual outcomes are not observed. In recent work, Einav et al. (2020) use a clinical oncology model of cancer growth to estimate the effects of changing screening recommendation ages. Kowalski (2023) analyzes data from a clinical trial that randomized screening to estimate health effects of mammography. Both these papers also discuss how cancer screening guidelines are more likely to be followed by people at lower risk of deadly cancer.

Figure 1: Health effects of screening depend on a tumor's potential to grow.



Notes – Screening policies aim to diagnose tumors at earlier stages, while keeping overdiagnosis low. For a tumor that would otherwise grow, early detection through screening leads to early diagnosis before the tumor grows and may cause health problems. For a tumor that would otherwise not grow, early detection through screening leads to overdiagnosis, as the tumor would never grow and cause health problems. Therefore, the health effects of using a machine learning model to set screening policy depend both on the model's predictability (which leads precision screening policies to catch more tumors) and its ability to target tumors that will grow (which makes catching those tumors valuable).

it may be easier to treat. A tumor that is not caught early through a screen has the potential to grow until it causes symptoms. If a tumor is diagnosed after causing symptoms, it may have spread beyond its original site, require more aggressive treatment, and result in reduced survival. Figure 1 describes the potential outcomes of a screen-detectable tumor that is either screened or left unscreened. An effective screening program increases early diagnosis of cancer—a machine learning model that can identify women at high risk of cancer can target screening to women who are likely to receive cancer diagnoses. However, not every cancer diagnosis leads to health benefits, and treatment of a cancer that would not otherwise grow may cause harm. Because not every positive screen is valuable, the goal of an effective cancer screening program is not just to maximize early diagnosis, but also to avoid high

rates of of overdiagnosis. If we were to target screening using a machine learning model that predicts screen-detectable cancer, the model's predictive performance alone would not determine all of the health effects of using it to guide screening. Therefore, in this paper we both train machine learning models to predict breast cancer and estimate the health effects of using these models to determine breast cancer screening policies.

We begin by training models to predict cancer-related outcomes using Danish demographic and healthcare claims data. Traditionally, age is the primary characteristic used to set screening criteria as it predicts cancer incidence and is readily observed. Some of the additional characteristics that could help to assess medical risks may not be as readily observed and may involve collecting new information through genetic testing or other costly procedures (e.g. Conner et al., 2022). However, the rich set of characteristics in health claims data is already available—it can be used to inform screening recommendations without having to collect costly new information. These claims data, paired with the existing risk factors, may help to build better models of cancer risk. For each woman in our dataset, we observe the breast cancer risk factors of age, age of first birth, and family history, as well as a large set of medical claims and prescriptions history. For our main analyses, we train a model using this data to predict an invasive tumor caught through a screen—not just an abnormal mammogram result—a distinction we will revisit later. The model is predictive, with an out-of-sample AUC-ROC score of 0.629. The added flexibility of our machine learning model leads to a large improvement in predictive performance. Our model improves predictive performance by 55% over using just age to predict invasive tumors, and by 45% over a model that considers age along with established demographic and medical risk factors.[3]

How should we interpret this gain in predictive performance? We can estimate

[3]Conner et al. (2022) estimate how using risk to target screening guidelines could increase the effectiveness of screening for chromosomal abnormalities. The authors find that risk scores obtained from a non-invasive procedure can predict subsequent positive test results from an invasive procedure. Huang et al. (2021) estimate how much machine learning models trained on healthcare claims data to predict bacterial test results can improve performance over models trained on just demographic data.

how effectively policies informed by our risk model catch tumors compared to policies informed by age. In the sample used to train the algorithm, Danish women aged 50-69 were targeted for mammography through a national screening program. A counterfactual policy that would raise the eligibility age for the program would both reduce overall screening and reduce the number of invasive tumors caught through the program. We use data on tumors caught through the program and screens conducted through the program to estimate effects of counterfactual policies with different eligibility criteria. We show that when compared to policies that determine program eligibility using age thresholds, policies that determine program eligibility using risk thresholds could both increase the number of invasive tumors caught and reduce overall screening. Compared to a policy that would raise the eligibility age for the program by three years, a policy that would target mammograms using the risk model would miss 31% fewer invasive tumors while keeping the overall number of screens constant. A different risk-based policy could keep the overall number of invasive tumors found constant, but would lead to an additional 31% reduction in total screens.[4] In addition to improving the efficiency of screening, we show policies that are targeted through risk models would screen women with larger tumors, which the medical literature has shown are more likely to become deadly.

The previous results demonstrate that algorithms can more effectively target screens to women for whom these screens would catch invasive tumors. However, these results do not yet speak to whether precision screening policies set by algorithms would lead to long-run health benefits. Catching some tumors early through screening might not lead to any health benefits. For example, some screen-caught tumors might never become symptomatic or deadly—a policy that catches these kinds of tumors at screen-detectable stages would not improve health outcomes, and would result in overdiagnosis of cancer. To make claims about

[4]These results have an analogy in clinical testing—the optimal level of testing depends on diagnostic skill. Abaluck et al. (2016) and Currie and MacLeod (2017) argue that the key question is more than just how much to test, because clinicians vary in diagnostic skill. In the context of screening, we are comparing the diagnostic skill of different risk models. Differences in skill imply that policies with the same welfare effects need not hold fixed the total level of screening.

long-run health effects, we must estimate how observed outcomes would differ from counterfactual outcomes if women were not targeted for screening. As we do not observe these counterfactuals in the data, these estimates are subject to a causal inference challenge. For a woman whose cancer was caught through screening, we do not know how the cancer would have developed if she had not received early treatment. For a woman who was not covered by the screening program but developed a symptomatic cancer, we do not know whether early screening would have improved her health outcomes. We require some other source of variation in the data to produce credible counterfactual estimates.

To construct these counterfactuals and estimate the health effects of screening, we use a natural experiment produced by the widespread introduction of breast cancer screening in Denmark from 2007-2010. Over this period, most Danish women aged 50-69 were introduced to a population-level screening mammography program. If a counterfactual population of women who were not targeted by the screening program would have the same cancer incidence patterns in the years following the policy change as in the years before the policy change, this natural experiment allows us to evaluate the health effects of screening by comparing post-policy health outcomes to pre-policy health outcomes. The clinical literature suggests we estimate these effects on a natural set of health outcomes related to long-run cancer incidence and early detection rates.

Welch et al. (2016) argue that an effective screening policy reduces the long-run incidence of large ($\geq$ 2cm diameter) tumors, which are more likely to cause health problems. The authors also argue that an effective screening policy should not increase the long-run detection of small ($<$ 2cm diameter) tumors too much relative to the decrease in large tumors, as this suggests that some of these small tumors would have never become large, and were therefore overdiagnosed. We first estimate effects of the existing Danish universal screening policy on these outcomes, and then estimate how these effects differ across high-risk and low-risk women. Across the full population covered by the Danish screening policy, we estimate that large tumor incidence decreased by 41 per 100,000 person-years while small tumor detection increased by 99 per 100,000 person years. The universal screening

policy decreased large tumor incidence by 31% and led to an overdiagnosis rate of 59%.

While these estimates correspond to health outcomes of the entire population of women covered by the screening policy, women at higher risk might benefit more from screening than women at lower risk. We find for women with high risk, large tumor incidence would decrease by an additional 61 per 100,000 person-years. In addition, small tumor detection would increase by 53 per 100,000 person-years. We estimate that for women at high risk, the health effects of screening for reducing large tumor incidence are 4.4 times larger than for women at low risk. The overdiagnosis rate of these women is also much lower, at 32% compared to 72%. Moreover, we find that women at high risk would have larger decreases in cancer mortality from the policy—women with high invasive tumor risk would have 32 fewer cancer deaths per 100,000 person-years.

While the previous estimates correspond to the effects of policies that differentially target women aged 50-69 for screening, it may also be of high value to target younger women for screening if they are at sufficiently high risk of cancer. Estimating the effects of such policies requires additional structure beyond our current environment, as there is no historical variation in screening eligibility for women younger than 50 in Denmark. We turn to a clinical oncology model of breast cancer development (Tan et al. 2006, as calibrated in Einav et al. 2020) to produce these estimates. We estimate that a policy that screens high-risk women starting at 40, and all women starting at 50, would reduce the share of tumors that become large by 15%.

In isolation, these results demonstrate that machine learning models can effectively predict cancer-related outcomes and can lead to large benefits when used to guide health policy. However, we also present a cautionary tale that illustrates that predictability alone is not enough to determine the health benefits of a policy guided by machine learning. One design choice was crucial for our results: We trained our main model to predict invasive tumors. A perhaps more natural choice would have been to train the model to predict abnormal mammography results. "Positive," abnormal mammograms are a direct product of screening—they

are identified before any additional invasive procedures, and finding a screen-identified abnormality could suggest that the mammogram was of high value. A model trained on abnormal mammograms also has similar predictability as our main model, with an AUC-ROC of 0.625. These results alone may suggest that targeting screening using abnormal mammogram risk could be as effective as targeting screening using other cancer-related risks.

However, using the same estimation strategies as we use to evaluate our main invasive tumor model, we find that policies set using the abnormal mammogram model would have much smaller health benefits. We estimate women at high risk of abnormal mammograms experience a 16% reduction in large tumor incidence, compared to 31% among women at high risk of invasive tumors. We estimate the cancer overdiagnosis rate of women at high risk of abnormal mammograms is 61%, compared to 32% for women at high risk of invasive tumors. To understand why policies set using abnormal mammogram risk would produce poorer outcomes, consider that common outcomes from screening mammography are false positive and earlier-stage cancers. A model trained to identify women at higher risk of an abnormality may also identify women at higher risk of false positives or cancers that may not grow.

This analysis demonstrates that the alignment between the prediction target and welfare objective is crucial in prediction policy problems. The fact that a model is predictive does not determine whether its prediction target is welfare-aligned. In this paper we show that a seemingly-intuitive prediction target—abnormal mammography result—may lead an algorithm to target screening to women whose cancers may never grow. A model trained on a prediction target that reflects cancer at later stages—invasive tumor—leads to more effective policy. Making decisions based on machine learning models whose prediction targets are misaligned can lead to inefficiency (Mullainathan and Obermeyer, 2017) and inequity (Pierson et al., 2021), and changing the prediction target to better match the welfare-relevant objective can lead to large improvements in health outcomes (Obermeyer et al., 2019). Finding the right prediction target requires domain knowledge and the ability to evaluate the decisions made using the model's outputs.

2. Background and Data

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.[5]

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, the standard screening method currently 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.[6]

Virtually all developed countries have policies concerning coverage of mammography screening. We compare screening programs across OECD countries in Table 1. In Europe, mammography screening is typically provided through organized screening programs (Altobelli and Lattanzi, 2014; Guthmuller et al., 2023). During the 1980s and 1990s, mammography screening was offered by a handful of

[5]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. More information is available at this link.

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

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 50–69. 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 aged 50–69 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 US Preventive Services Task Force recently issued new draft recommendations for biennial screening mammography for women aged 40 to 74, which increased the range from its previous recommendation to stop screening for women aged 40–49 in 2009.[7]

These widespread screening recommendations have recently come under scrutiny

[7]More information is available at this link. Churchill and Lawler (2023) document the effects of the task force's previous recommendation for women aged 50–74.

because of the potential harms that may outweigh the benefits. In particular, mammography screening can result in high false-positive rates and overdiagnosis—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. Current estimates come from randomized control trials (Miller et al., 2014; Kowalski, 2021) or natural experiments (Welch et al., 2006), and find overdiagnosis rates ranging from 14-81%.

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 estimates the effects of 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). Denmark has increased the population coverage of its breast cancer screening programs, with a nationwide push beginning in 2007. We discuss the Danish breast cancer screening context further in Appendix A.1.

2.2. Data

We use national administrative data from Denmark over the period 2004–2019. Using individual identification numbers, we are able to add information from several other sources of administrative data. First, we use the Danish Breast Cancer Cooperative Group data set, which covers the near-universe of invasive breast cancer diagnoses over our sample period, with detailed data on clinical characteristics

(including tumor size), and date and type of surgery. Information on screening comes from the Danish Quality Database of Mammography Screening database, which covers the nationwide screening program from 2008–2016 (Langagergaard et al., 2013; Mikkelsen et al., 2016). We obtain medical history and demographic information from other registers, which we describe further in Appendix A.2

Our sample is constructed as follows. Over the period 2004–2019, we identify in each year women with no invasive breast cancer history aged 50-69—the age range targeted by the nationwide program. We restrict our analysis to women living in regions of Denmark that did not have existing screening programs, which account for 75% of women in the target age range across the country. Our data frame covers a total of 8,580,783 person-years and 973,310 women. Out of these observations, 4,945,009 occurred starting in 2011, after the mammography screening introduction had concluded. Based on the invasive cancer data from the Danish Breast Cancer Cooperative Group (Christiansen et al., 2016), we also identify cancer characteristics for every woman diagnosed with invasive cancer in our sample, and whether the cancer was found through the screening program. The incidence of invasive cancers found across our entire sample was 307 per 100,000 person-years, of which 99 per 100,000 person-years were large tumors with diameter greater than or equal to 2cm. We use the Danish Quality Database of Mammography Screening to identify abnormal mammography results identified through the screening program. Table 2 describes the composition of our data across these categories.

2.3. Model training and comparison

We randomly split our main sample into a 50% training sample and 50% evaluation sample. We split the data by individual so each woman appears solely in either the train or evaluation sample. We develop and evaluate the performance of our cancer risk models on a subsample of 441,515 women aged 50-69 who were screened through the population program in 2009 and 2010.[8] We train gradient-boosted

[8]Although the national program was announced in 2007, the widespread introduction of population-level screening mammography did not begin until the middle of 2008. Our dataset has

trees using age, family history, nulliparity, age at first birth, and 581 variables that correspond to previous medical and prescription claims. We include more information about our training procedure in Appendix A.2.

For our main model, we predict whether the screen led to an invasive tumor diagnosis. As cancer is a rare event, to evaluate our model's predictive power we consider the area under the curve of the receiver operating characteristic curve (AUC-ROC).[9] The AUC-ROC corresponds to the probability that the model correctly ranks a randomly-selected positive class–negative class pair. Our full model yields an AUC-ROC of 0.629. For comparison, using just age to predict invasive tumor has an AUC-ROC of 0.583. A gradient boosted-tree model model that includes age, nulliparity, age of first birth, and family history increases the AUC-ROC to 0.585, and adding history of progestogens, estrogens and angiotensin receptor blockers—drugs often found to be correlated with breast cancer—increases the AUC-ROC to 0.589. Compared to a random classifier, which would have an AUC-ROC or 0.5, our model increases predictive performance by $(0.629 - 0.5)/(0.583 - 0.5) \approx 55\%$ over using just age to predict invasive tumor incidence. Our model also increases predictive performance by $(0.629 - 0.5)/(0.589 - 0.5) \approx 45\%$ over using a model trained on established risk factors.

We additionally train a model to predict abnormal screening mammography results among the women who were screened during this period. This predictor has similar performance as the invasive tumor predictor, with an AUC-ROC of 0.625. It may be intuitive to target a screening policy to women who are likely to have an abnormal screen result. We will revisit this question in Section 5.

3. Turning Predictions into Screening Policies

How can we interpret the gain in predictive performance from this machine learning model? In this section, we compare how effectively the invasive tumor

comprehensive labels for cancers caught through the program in 2009 and 2010.

[9] Raw accuracy scores are less appropriate for predicting a rare event than the AUC-ROC metric. A classifier that predicts no cancer for all women would have an accuracy higher than 99 percent, but would not provide useful rankings of women by breast cancer risk.

predictor can predict invasive tumors compared to using just age as a predictor. We show that using risk to guide cancer screening policies can catch more cancers and reduce overall screening compared to using age to guide cancer screening policies. Furthermore, we show that risk-based policies target screening to women with tumors with larger diameter, which may be more problematic. All our calculations are performed on the 50% evaluation subsample consisting of women whose health data were not used to train our models.

3.1. Algorithms can target screening more effectively

In our sample we observe if an invasive tumor was caught through the screening program. This presents a clear counterfactual test: If a subset of those women had not been screened by the program, how many program-caught tumors would be missed? This empirical exercise nests the existing policy discussions around the effects of raising minimum screening ages: For example, how many screen-caught tumors would be missed if the minimum age were raised from 50 to 53? We can conduct a similar counterfactual analysis by restricting the screened subset according to risk scores: How many screen-caught tumors would be missed the 20% lowest-risk women were not screened?[10]

Figure 2 presents tradeoffs between reduced screening and missed cancers according to age- and risk-based policies that sequentially restrict screening criteria. Across every reduction in total screens arising from raising the eligibility age, there exists a risk-based policy that misses fewer tumors. For example, compared to raising the eligibility age by three years, reallocating screens using the risk model's ranking could miss 31% fewer invasive cancers. Alternatively, total screening could be reduced 31% further under the model's ranking while keeping the total number of tumors missed constant. Table 3a and Table 3b present these counterfactual

[10]These policy counterfactuals considering reducing screening program coverage across a subpopulation that accepted the invitation to pursue mammography. As opportunistic mammography is rare in Denmark (Jensen et al., 2005), and our outcome variable is cancer caught through the screening program, we are less likely to be overestimating the number of women who comply with mammography recommendations—a policy evaluation challenge discussed in Einav et al. (2020) and Kowalski (2023).

estimates for algorithmic policies and policies that consider a range of potential minimum screening age increases.

Across the board, targeting screening through risk improves the efficiency of screening. For any policy change that raises screening eligibility criteria using age, there exist risk-based policies that reduce total screening or catch more invasive tumors.

3.2. Algorithms do not miss more dangerous-looking tumors

So far we have shown that precision screening policies catch more invasive tumors per screen than age-based policies do. However, it is also helpful to test that reprioritizing screens using a risk model does not miss women with cancers that appear more likely to be dangerous. One way to test for this concern is to compare the characteristics of tumors cut by narrowing screening according to age versus the algorithm: If an algorithmic policy cuts screens that find dangerous-looking tumors more often than an age-based policy does, the algorithmic policy may be missing more problematic cancers. We conduct one such exercise in Figure 3, where we compare the tumors cut by either policy regime across their diameters, a proxy for cancer severity. Narrowing screening criteria according to risk misses tumors that are no larger, on average, than those missed by narrowing using age thresholds. In fact, narrowing screening criteria using risk-based policies would target screening to women who have larger tumors on average.

Our comparisons are on the characteristics of the cancer at the time it is caught. As the cancers caught in this subsample were treated, we cannot draw comparisons between the effects of screening on eventual health outcomes across the risk distribution. To do this, we require additional variation in the sets of people who are screened, which we will find in the natural experiment that introduced widespread nationwide screening to Denmark.

4. Health Effects of Screening Policies

While we have demonstrated machine learning models can predict various types of cancer in a hold-out set, these results are not sufficient to make claims about the health effects of algorithmic screening policies. In this section we present a series of counterfactual policy estimates using policy evaluation methods to show that our main risk model can indeed be used to target screens to improve health outcomes. All our calculations are performed on the 50% evaluation subsample consisting of women whose health data were not used to train our models.

4.1. Longer-term impacts of precision screening policies

So far, our counterfactual analyses have considered the effects of risk-based policies on tumors caught through screening. However, once cancers become symptomatic or harmful enough, they may be caught eventually without a screen. How can we be sure catching these tumors early is valuable? Making claims about the longer-term impacts of precision screening therefore requires estimates of the eventual health effects of screening programs. We consider one such estimation framework using the natural experiment created by the Danish implementation of population-level breast cancer screening.

From 2007-2010, most Danish municipalities were introduced to widespread breast cancer screening for women aged 50-69. We show in Figure 4 that the screening program increased the detection of small tumors. Moreover, the longer-run incidence of large tumors decreased, suggesting screening reduced the eventual incidence of cancers at more difficult-to-treat stages. If we take large tumor reduction as a policy goal, as is considered elsewhere in the clinical literature (e.g. Welch et al., 2011), the results suggest screening had positive health impacts on this population.[11]

[11] Our results consider health outcomes before and after the Danish universal screening program was introduced, and our effects should be interpreted as applying to the population of Danish women who comply with the program. Einav et al. (2020) discuss the importance of estimating health effects of screening policies on compliers. In the Danish context compliers are a relatively larger portion of the population: Opportunistic mammography (outside of a screening program) is

Who were the women for which this decrease in large tumor incidence was more pronounced? In Figure 5 we show changes in the number of large tumors caught after the policy was introduced, separating the population into above- and below-median risk. The figure shows that the reduction in large tumors was more pronounced among the high-risk population. This result implies that the cancer risk model targets women who benefited more from screening.

We estimate the effects of the screening policy using the following regression specification

$$\text{LargeTumor}_{i,t} = \beta_0^{\text{large}} + \beta_1^{\text{large}} \cdot \text{Post}_t + \varepsilon_{i,t} \tag{1}$$

where $\text{LargeTumor}_{i,t}$ indicates large tumor incidence per 100,000 person-years and Post_t indicates whether the observation occurs after 2007.

We further estimate the effects of the screening policy by risk using the following regression specification

$$\begin{aligned}\text{LargeTumor}_{i,t} =& \gamma_0^{\text{large}} + \gamma_1^{\text{large}} \cdot \text{Post}_t + \gamma_2^{\text{large}} \cdot \text{HighRisk}_{i,t} \\ &+ \gamma_3^{\text{large}} \cdot \text{Post}_t \cdot \text{HighRisk}_{i,t} + \varepsilon_{i,t}\end{aligned} \tag{2}$$

where $\text{LargeTumor}_{i,t}$ indicates large tumor incidence per 100,000 person-years, Post_t indicates whether the observation occurs after 2007, and $\text{HighRisk}_{i,t}$ is an indicator for above-median risk. We report the results of these regressions in Table 4, and consider alternative specifications in Appendix B. High risk predicts 61 fewer large tumors per 100,000 person-years after the introduction of screening. For women at high risk, this corresponds to a $-\gamma_3^{\text{large}}/(\gamma_0^{\text{large}} + \gamma_2^{\text{large}}) = 61/(79+118) \approx 31\%$ decrease in large tumor incidence.

These results show that screening reduces large tumor incidence, and that higher-risk women especially benefit on this health outcome. How does the even-

rare in Denmark (Jensen et al., 2005), while compliance with the screening program is close to 80% (Lynge et al., 2017). While our results do not speak to the effects of screening onn women who do not comply with screening programs, strategies to encourage participation among this population are an interesting topic for future research.

tual reduction in large tumors trade off with the increase in small tumors caught through screening? Table 5 reports estimates of the effects of screening on small tumor detection, using the specifications

$$\text{SmallTumor}_{i,t} = \beta_0^{\text{small}} + \beta_1^{\text{small}} \cdot \text{Post}_t + \varepsilon_{i,t} \tag{3}$$

$$\begin{aligned}\text{SmallTumor}_{i,t} =& \gamma_0^{\text{small}} + \gamma_1^{\text{small}} \cdot \text{Post}_t + \gamma_2^{\text{small}} \cdot \text{HighRisk}_{i,t} \\ &+ \gamma_3^{\text{small}} \cdot \text{Post}_t \cdot \text{HighRisk}_{i,t} + \varepsilon_{i,t}\end{aligned} \tag{4}$$

In the full sample, large tumor incidence decreased by 41 per 100,000 person-years and small tumor incidence increased by 99 per 100,000 person-years. These results suggest that the universal screening policy led to an overdiagnosis rate of $(\beta_1^{\text{small}} - \beta_1^{\text{large}})/\beta_1^{\text{small}} = (99-41)/99 \approx 59\%$. However Table 5 shows that for women at high risk, the relative decrease in large tumors from the policy is larger than the relative increase in small tumors. We estimate that the overdiagnosis rate for screen-caught tumors for women with low risk is $(\gamma_1^{\text{small}} - \gamma_1^{\text{large}})/\gamma_1^{\text{small}} = (64-18)/64 \approx 72\%$. We estimate that the overdiagnosis rate for screen-caught tumors for women with high risk is $((\gamma_1^{\text{small}} + \gamma_3^{\text{small}}) - (\gamma_1^{\text{large}} + \gamma_3^{\text{large}}))/(\gamma_1^{\text{small}} + \gamma_3^{\text{small}}) = ((64+53)-(18+61))/(64+53) \approx 32\%$. The overdiagnosis rate for women at high risk is less than half the overdiagnosis rate for women at low risk. In Appendix B, we consider alternative specifications that omit the screening policy introduction years and add controls for age, and find similar results.

4.2. Precision screening policies and cancer mortality

In addition to effects on reductions in large tumors, we also estimate the effects of screening on cancer mortality. We estimate these effects using the following

regression specification

$$\begin{aligned}\text{CancerMortality}_{i,t} =& \gamma_0^{\text{mortality}} + \gamma_1^{\text{mortality}} \cdot \text{Post}_t + \gamma_2^{\text{mortality}} \cdot \text{HighRisk}_{i,t} \\ &+ \gamma_3^{\text{mortality}} \cdot \text{Post}_t \cdot \text{HighRisk}_{i,t} + \varepsilon_{i,t} \end{aligned} \tag{5}$$

where $\text{CancerMortality}_{i,t}$ indicates cancer mortality per 100,000 person-years, Post_t indicates whether the observation occurs after 2007, and $\text{HighRisk}_{i,t}$ is an indicator for above-median risk. We report the results of this regression in Table 6. High risk predicts 32 fewer cancer deaths per 100,000 person-years after the introduction of screening. For women at high risk, this corresponds to a $-\gamma_3^{\text{mortality}}/(\gamma_0^{\text{mortality}} + \gamma_2^{\text{mortality}}) = 32/(23+41) \approx 50\%$ decrease in cancer mortality. These results indicate that high risk women also experienced large decreases in cancer mortality after the screening program was introduced.

However, given the contemporaneous improvements in breast cancer treatment technology over this period, we cannot rule out that the Danish medical system differentially improved its ability to treat high-risk women. For example, we have previously shown that high-risk women have larger diameter tumors on average than low-risk women. If treatment technology improved more for larger diameter tumors or tumors with more complications over this period, some of our estimated effect on cancer mortality could be driven by this differential improvement in treatment technology. This is why our main empirical design uses size of caught tumors as the outcome variable, not cancer mortality. This follows from the observation in Welch et al. (2016) that compared to estimating mortality effects, estimating tumor size-related effects requires the weaker assumption that the underlying disease burden does not change differently across groups over time.

4.3. Effects of more flexible precision screening policies

So far, we have only considered policies that narrow the set of women to screen. These counterfactuals speak to the existing policy discussions on raising recommended screening ages. However, these counterfactuals may not fully describe

the benefits of using machine learning to guide screening policy. Specifically, expanding screening to higher-risk, younger women might catch tumors at earlier stages and prevent the eventual incidence of deadly cancer. To estimate the effects of such policies, we require additional structure beyond that provided for by policy environment and natural experiments. We calibrate the Erasmus clinical oncology model of cancer development (Tan et al., 2006) to the Danish context to cancer incidence rates across our sample.

We summarize its key aspects here and include a more detailed discussion in Appendix C. The model considers a population of 20-year-old cancer-free women who, each year, may develop a tumor. In the model, probability of developing a tumor is increasing in age, and tumors develop at variable rates. Tumors may either be invasive—which become detectable symptomatically if they grow enough, or ductal carcinoma in situ (DCIS)—which are only detectable through screening and may never become invasive. We largely follow the parameterization of the model in Einav et al. (2020), while allowing for tumor incidence to be influenced by risk. We calibrate the full model on Danish cancer incidence data, and augment our main sample with the cancer history of women aged 40-49.

With the model, we can estimate health outcomes under counterfactual policies that take risk into account. We consider the health effects of policies that lower the screening age to 45 or 40. We also consider policies that only screen younger women at high risk: These policies only screen those women with risk greater than the median risk of a 50-year-old woman. In addition, we consider the effects of screening higher-risk women starting at 40 and all women starting at 55. Table 7 reports the estimated incidence of screens and tumors after the introduction of screening on each of these counterfactual populations. We estimate that setting the screening threshold to 40 for high-risk women, and continuing to cover all women aged 50 and older for screening would reduce large tumor incidence for women aged 40-69 by 15%.

5. Alignment

Our analysis has shown that targeting screening based on invasive tumor risk could lead to large health benefits. However, we will show that an crucial decision for training this model was the choice of training label. Even if models are similarly predictive, they may lead to policies with very different health outcomes. In this section we discuss how an intuitive alternative strategy—determining risk by predicting abnormal mammograms—could lead to worse health outcomes than our approach would.

5.1. Which "cancer" to predict?

Throughout our analyses, we have focused on a model trained to predict invasive tumors. However, we could have also used a number of other models trained to predict different cancer-related outcomes. In particular, we could have followed a seemingly more natural approach—to decide who to invite for screening mammography, we could have targeted those women who are likely to receive an abnormal, "positive" mammogram result. These abnormal results are produced by the screening procedure, and are observed before any further procedures. If the goal of a screening policy was to target screening to those who would have abnormal mammogram results, using such a predictor may be an intuitive one. The ultimate health effects of using such a predictor to guide policy, however, hinge on its ability identify women with tumors that benefit from early catching. The diagnostic skill of an algorithm that predicts abnormalities, therefore, depends on the diagnostic skill of the machines and clinicians who produce the abnormal label. If those women whose mammograms are likely to be flagged for abnormalities are not those women who are at highest risk for potentially-dangerous cancers, such a predictor may be misaligned.

5.2. Effects of screening by abnormal mammography risk

To test the health effects of targeting policy using such a predictor, we conduct our main counterfactual policy analysis using a model trained to predict abnormal mammography results. We consider the long-run effects of screening on large and small tumor incidence among those women who are labeled as high-risk and low-risk by the abnormal mammogram model. Column (3) of Table 4 re-estimates Equation (2) and Column (3) of Table 5 re-estimates Equation (4), by replacing $\text{HighRisk}_{i,t}$ with above-median abnormal mammogram risk.

We find that women labeled as "high risk" by the abnormal mammogram predictor have both lower reductions in long-run large tumor incidence and higher rates of cancer diagnosis from screening, compared to women labeled as "high risk" by the invasive tumor predictor. We find that high abnormal mammogram risk predicts 26 fewer large tumors per 100,000 person-years after the introduction of screening. For women at high abnormal mammogram risk, this corresponds to a $-\gamma_3^{\text{large}}/(\gamma_0^{\text{large}}+\gamma_2^{\text{large}}) = 26/(104+54) \approx 16\%$ decrease in large tumor incidence. Note that for our main invasive tumor risk model, the corresponding decrease in large tumors was 31%. In addition, we estimate that the overdiagnosis rate for screen-caught tumors for women with high abnormal mammogram risk is $((\gamma_1^{\text{small}}+\gamma_3^{\text{small}})-(\gamma_1^{\text{large}}+\gamma_3^{\text{large}}))/(\gamma_1^{\text{small}}+\gamma_3^{\text{small}}) = ((60+77)-(28+26))/(60+77) \approx 61\%$. The corresponding overdiagnosis rate was 32% for women with high invasive tumor risk.

As the Erasmus clinical oncology model implies, women who have invasive tumors are more likely to develop deadly cancers than women who have abnormal mammography results. The abnormal mammogram predictor may oversample women who have abnormalities that are unlikely to become problematic relative to the invasive tumor predictor. As false positives and cancers that may never grow are common outcomes of screening mammography, targeting screening based on abnormal mammogram risk may be more likely to target women at risk of false positives and cancers that may never grow than targeting screening based on invasive tumor risk. Our analysis in Table 4 reveals that although the abnormal

mammogram model is predictive, it is not as welfare-aligned—using it to guide policy would improve health outcomes less than an invasive tumor model would.

6. Discussion

We consider the health impacts of using machine learning to form precision screening policies for cancer. We first demonstrate that more complex machine models trained on health data can predict cancer better than simpler models that use established risk factors. Although these results show that machine learning models can be predictive, they do yet demonstrate that such models can be used to form beneficial screening policies. We do this using a series of policy evaluation methods that find such screening policies could lead to large health benefits. Moreover, we demonstrate the choice of prediction target is key—not all models with similar predictive performance can be used to construct policies with similar health effects. Our results demonstrate that a key consideration when using machine learning models to make policy decisions is the model's prediction target.

References

Abaluck, Jason, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh, "The determinants of productivity in medical testing: Intensity and allocation of care," *American Economic Review*, 2016, *106* (12), 3730–3764. 4

Altobelli, Emma and Annalisa Lattanzi, "Breast cancer in European Union: an update of screening programmes as of March 2014," *International journal of oncology*, 2014, *45* (5), 1785–1792. 8

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. 10

Bitler, Marianne P and Christopher S Carpenter, "Health insurance mandates, mammography, and breast cancer diagnoses," *American Economic Journal: Economic Policy*, 2016, *8* (3), 39–68. 9

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. 10

Chen, Tianqi and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in "Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining" 2016, pp. 785–794. 43

Christiansen, Peer, Bent Ejlertsen, Maj-Britt Jensen, and Henning Mouridsen, "Danish breast cancer cooperative group," *Clinical epidemiology*, 2016, *8*, 445. 11

Churchill, Brandyn F and Emily Lawler, "Government Recommendations and Health Behaviors: Evidence from Breast Cancer Screening Guidelines," *Available online*, 2023. 9

Conner, Peter, Liran Einav, Amy Finkelstein, Petra Persson, and Heidi L Williams, "Targeting Precision Medicine: Evidence from Prenatal Screening," Technical Report, National Bureau of Economic Research 2022. 1, 3

Currie, Janet and W Bentley MacLeod, "Diagnosing expertise: Human capital, decision making, and performance among physicians," *Journal of labor economics*, 2017, *35* (1), 1–43. 4

Domingo, Laia, Katja Kemp Jacobsen, My von Euler-Chelpin, Ilse Vejborg, Walter Schwartz, Maria Sala, and Elsebeth Lynge, "Seventeen-years overview of breast cancer inside and outside screening in Denmark," *Acta Oncologica*, 2013, *52* (1), 48–56. 40

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. 1, 6, 10, 13, 15, 19, 54

Guthmuller, Sophie, Vincenzo Carrieri, and Ansgar Wübker, "Effects of organized screening programs on breast cancer screening, incidence, and mortality in Europe," *Journal of Health Economics*, 2023, *92*, 102803. 8

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. 10

Huang, Shan, Michael A Ribers, and Hannes Ullrich, "The value of data for prediction policy problems: Evidence from antibiotic prescribing," 2021. 3

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. 10

Jacobsen, Katja Kemp, My von Euler Chelpin, Ilse Vejborg, and Elsebeth Lynge, "Impact of Invitation Schemes on Breast Cancer Screening Coverage: A Cohort Study from Copenhagen, Denmark," *Journal of Medical Screening*, March 2017, *24* (1), 20–26. 40, 41

Jensen, Allan, Anne Helene Olsen, My von Euler-Chelpin, Sisse Helle Njor, Ilse Vejborg, and Elsebeth Lynge, "Do nonattenders in mammography screening programmes seek mammography elsewhere?," *International journal of cancer*, 2005, *113* (3), 464–470. 13, 16, 42, 54

Jørgensen, Karsten Juhl and Peter C. Gøtzsche, "Overdiagnosis in Publicly Organised Mammography Screening Programmes: Systematic Review of Incidence Trends," *BMJ*, July 2009, *339*, b2587. 42

Jørgensen, Karsten Juhl MD, Peter C. MD Gøtzsche, Mette Kalager, and Per-Henrik MD Zahl, "Breast Cancer Screening in Denmark: A Cohort Study of Tumor Size and Overdiagnosis," *Annals of Internal Medicine*, March 2017, *166* (5), 313–323. 42

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, "Prediction policy problems," *American Economic Review*, 2015, *105* (5), 491–95. 1

Kowalski, Amanda E, "Mammograms and Mortality: How Has the Evidence Evolved?," *Journal of Economic Perspectives*, 2021, *35* (2), 119–140. 1, 10

_ , "Behaviour within a Clinical Trial and Implications for Mammography Guidelines," *The Review of Economic Studies*, 2023, *90* (1), 432–462. 1, 13

Langagergaard, Vivian, Jens P Garne, Ilse Vejborg, Walter Schwartz, Martin Bak, Anders Lernevall, Nikolaj B Mogensen, Heidi Larsson, Berit Andersen, and Ellen M Mikkelsen, "Existing data sources for clinical epidemiology: the Danish Quality Database of Mammography Screening," *Clinical epidemiology*, 2013, pp. 81–88. 11

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. 16, 40, 41, 42, 55

Løberg, Magnus, Mette Lise Lousdal, Michael Bretthauer, and Mette Kalager, "Benefits and Harms of Mammography Screening," *Breast Cancer Research*, May 2015, *17* (1), 63. 10

Marcus, Pamela M, Nora Pashayan, Timothy R Church, V Paul Doria-Rose, Michael K Gould, Rebecca A Hubbard, Michael Marrone, Diana L Miglioretti, Paul D Pharoah, Paul F Pinsky et al., "Population-Based Precision Cancer Screening: A Symposium on Evidence, Epidemiology, and Next StepsPopulation-Based Precision Cancer Screening," *Cancer Epidemiology, Biomarkers & Prevention*, 2016, *25* (11), 1449–1455. 1

Mikkelsen, Ellen M, Sisse H Njor, and Ilse Vejborg, "Danish quality database for mammography screening," *Clinical epidemiology*, 2016, *8*, 661. 11

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. 10

Mullainathan, Sendhil and Ziad Obermeyer, "Does machine learning automate moral hazard and error?," *American Economic Review*, 2017, *107* (5), 476–480. 7

Njor, Sisse Helle, Anne Helene Olsen, Mogens Blichert-Toft, Walter Schwartz, Ilse Vejborg, and Elsebeth Lynge, "Overdiagnosis in Screening Mammography in Denmark: Population Based Cohort Study," *BMJ*, February 2013, *346*, f1064. 42

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, 2019, *366* (6464), 447–453. 7

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. 40

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. 41

Pierson, Emma, David M Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer, "An algorithmic approach to reducing unexplained pain disparities in underserved populations," *Nature Medicine*, 2021, *27* (1), 136–140. 7

Smith, Robert A, Kimberly S Andrews, Durado Brooks, Stacey A Fedewa, Deana Manassaram-Baptiste, Debbie Saslow, and Richard C Wender, "Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening," *CA: a cancer journal for clinicians*, 2019, *69* (3), 184–210. 1

Tan, Sita YGL, Gerrit J Van Oortmarssen, Harry J De Koning, Rob Boer, and J Dik F Habbema, "Chapter 9: the MISCAN-Fadia continuous tumor growth model for breast cancer," *JNCI Monographs*, 2006, *2006* (36), 56–65. 6, 19, 54

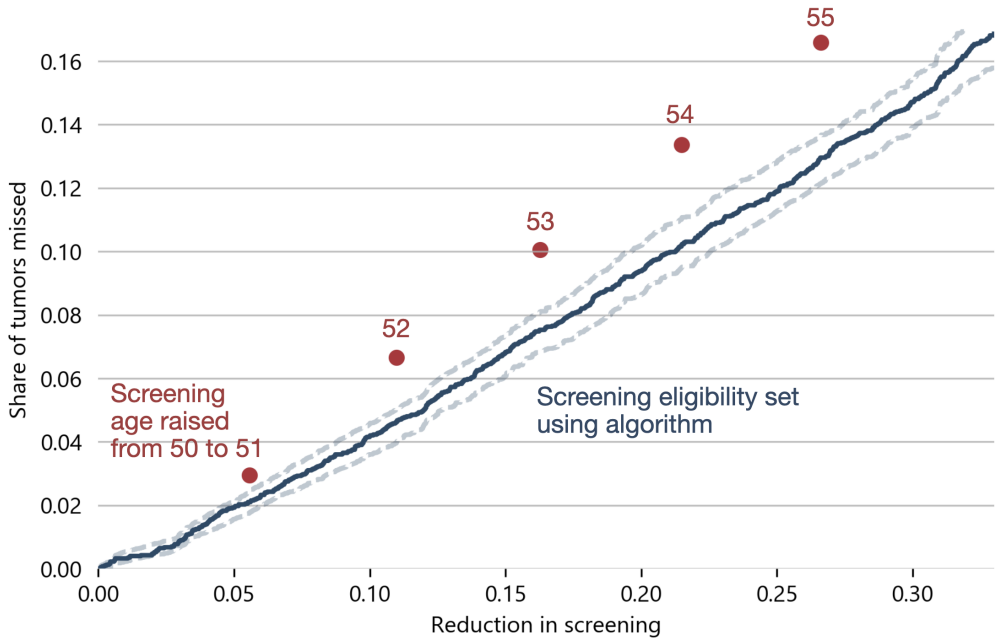
von Euler-Chelpin, My, Anne Helene Olsen, Sisse Njor, Ilse Vejborg, Walter Schwartz, and Elsebeth Lynge, "Socio-demographic determinants of participation in mammography screening," *International journal of cancer*, 2008, *122* (2), 418–423. 41

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. 10

_ **, Lisa Schwartz, and Steve Woloshin**, *Overdiagnosed: making people sick in the pursuit of health*, Beacon Press, 2011. 15

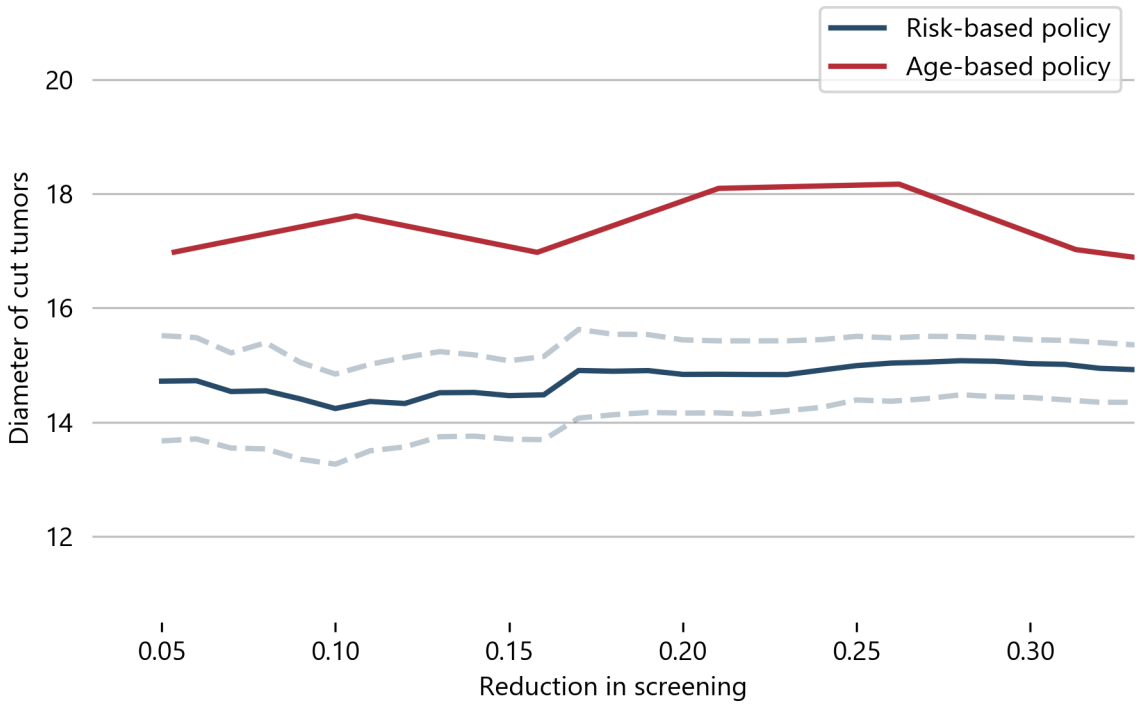
_ , **Philip C Prorok, A James O'Malley, and Barnett S Kramer**, "Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness," *New England Journal of Medicine*, 2016, *375* (15), 1438–1447. 5, 10, 18

Figure 2: Precision screening policies improve the yield of screening over age-based policies.



Notes – This figure shows the share of invasive tumors caught through screening that would be missed if the eligibility criteria were to be narrowed using age-based or risk-based criteria. The red dots consider the effects of raising the minimum age from 50. The blue curve considers the effects of raising the risk threshold for screening. We include 95% confidence intervals that correspond to re-estimates of the risk curve by randomly re-sampling the dataset with replacement. For every counterfactual policy that increases the minimum age for screening, there is an algorithmic policy that misses fewer tumors.

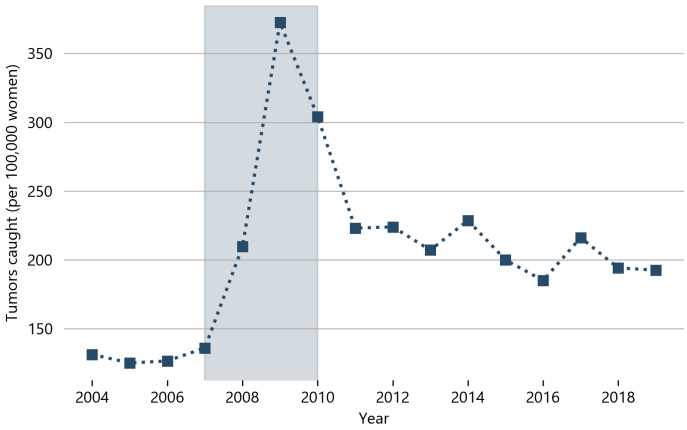
Figure 3: Narrowing screening criteria through age would miss larger tumors than the tumors that would be missed by narrowing screening criteria using risk.



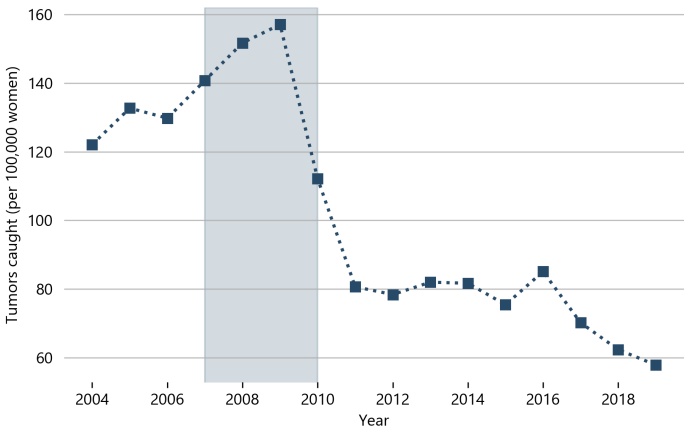
Notes – This figure shows the cumulative mean diameter of screen-caught tumors that would be missed if the eligibility critiera were to be narrowed using age-based or algorithmic rules. The red curve considers the cumulative mean diameter of tumors missed by narrowing the criteria via age-based rules, and the blue curve considers the cumulative mean diameter of tumors missed by narrowing the criteria via the algorithm's risk rankings. We include 95% confidence intervals that correspond to re-estimates of the risk curve by randomly re-sampling the dataset with replacement. Narrowing screening eligibility using the algorithm would miss smaller tumors than those missed by raising eligibility ages.

Figure 4: Population-level screening increased small tumor detection and reduced large tumor incidence.

(a) Small tumors (< 2cm)

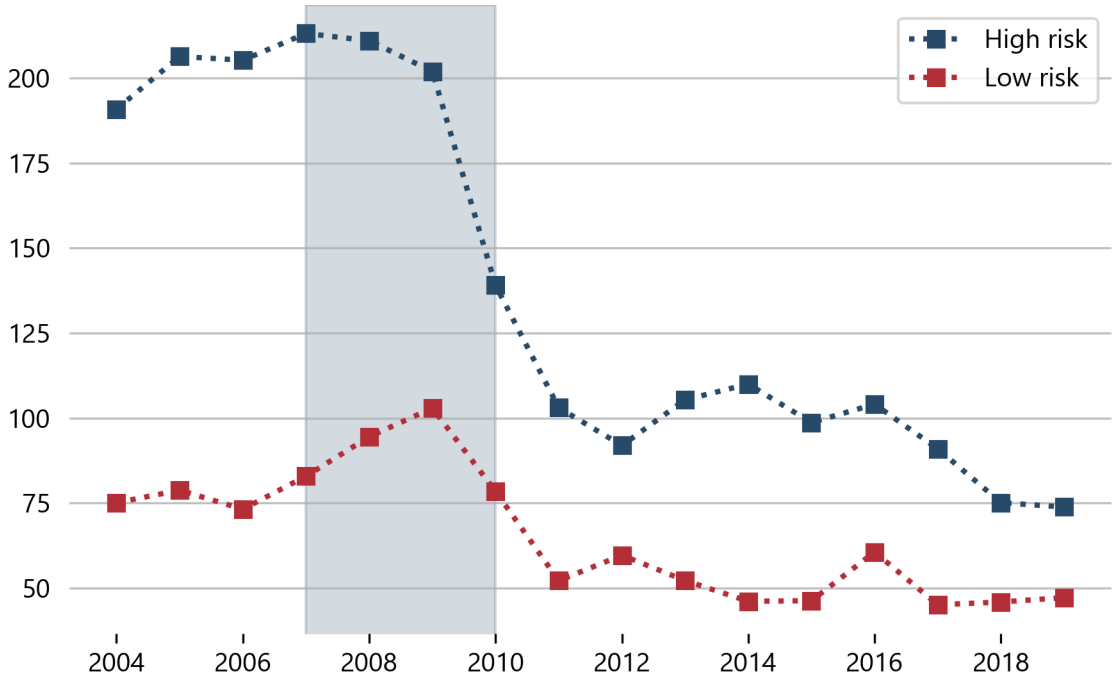


(b) Large tumors (≥ 2cm)



Notes – This figure plots the incidence of small and large tumors per 100,000 person-years from 2004-2019 across the main sample. The introduction of screening from 2007-2010 is shaded in blue. The screening policy increased the detection of small tumors and decreased the long-run incidence of large tumors.

Figure 5: High-risk women targeted by precision screening policies have stronger decreases in large tumor incidence.



Notes – This figure plots the incidence of large tumors per 100,000 person-years from 2004-2019 across the main sample, split by invasive tumor risk. High risk, plotted in blue, corresponds to risk above median. Low risk, plotted in red, corresponds to risk below median. The introduction of screening from 2007-2010 is shaded in blue.

Table 1: Breast cancer screening programs across OECD countries.

Country	Start date (regional / national)	Ages screened	Invitation interval (years)
Australia	1991 / 1995	50–74	2
Austria	2014	45–69	2
Belgium	2001	50–69	2
Canada[1]	1988	50–74[3]	1 / 2
Chile	—		
Colombia	—		
Costa Rica	—		
Czech Republic	2002	45+	2
Denmark	1991 / 2008	50–69	2
Estonia	2003	50–69	2
Finland	1987	50–69	2
France	1989 / 2004	50–74	2
Germany	2001 / 2005	50–69	2
Greece	—		
Hungary	1995 / 2001	45–65	2
Iceland	1987	40–74	2 / 3[4]
Ireland	2000	50–69	2
Israel	1992 / 1998	50–74	2
Italy[2]	1990	50–69[5]	2 / 1[6]
Japan	2004	40+	2
Korea	2002	40+	2
Latvia	2009	50–68	2
Lithuania	2005	50–69	2
Luxembourg	1992	50–69	2
Mexico	—		
Netherlands	1989	50–74	2
New Zealand	1999	45–69	2
Norway	1996 / 2005	50–69	2
Poland	2006	50–69	2
Portugal[2]	1990 / 2009	45–69[7]	2
Slovak Republic	2019	50–69	Ongoing
Slovenia	2008 / 2018	50–69	2
Spain[2]	1990 / 2001	50–69[8]	2
Sweden	1986	40–74	2
Switzerland[2]	1999	50–69[9]	2
Turkey	2004 / 2007	40–69	2
United Kingdom[2]	1988	50–70[10]	3
United States	—		

Notes – [1] Programs at the province and territory level do not cover the entire population. [2] Programs at the regional level, covering the entire population. [3] 40–74 in some regions. [4] Invitations are sent every 2 years to women 40–69 years old and every 3 years to women 70–74 years old. [5] 45–74 in some regions. [6] Invitations are sent every 2 years to women 50–69 years old and every year to women 45–49 years old. [7] 45–74 or 50–69 in some regions. [8] 45–69 in some regions. [9] 50–74 in some regions. [10] Some regions send invitations to women as young as 47 or as old as 73.

Table 2: Summary statistics for the main sample.

	Main Sample (2004-2019)	Post Policy Sample (2011-2019)
n observations	8,580,783	4,945,009
n women	973,310	804,094
Age	59.1	59.2
Invasive Tumor	328.6	310.9
Large Tumor	99.1	73.8
Abnormal Mammogram Result		714.4

Notes – This table reports population counts and means for key outcomes. The cancer-related outcome variables are scaled by 100,000 so that they correspond to outcome per 100,000 observations. The main sample ranges from 2004-2019, and consists of women aged 50-69 who lived in a region that adopted universal screening for the first time during the nationwide Danish screening program introduction in the late 2000s. The second panel reports statistics over the sample from 2011-2019, after the introduction of the universal screening policy had concluded.

Table 3: Precision screening policies can miss fewer cancers and lead to fewer overall screens.

(a) Number of screens held constant:

Years Screening Age Raised	Cancers Missed	Cancers Missed (Risk Ranking)
1	3.0%	2.1%
2	6.7%	4.6%
3	10.1%	7.5%
4	13.4%	10.2%
5	16.6%	13.0%

Notes – Screen-caught invasive tumors missed by an age-based policy that raises the minimum screening age from 50, and an algorithmic policy that raises the risk threshold to conduct the same number of screens as as the corresponding age-based policy.

(b) Cancers caught held constant:

Years Screening Age Raised	Screens Reduced	Screens Reduced (Risk Ranking)
1	5.5%	7.4%
2	10.9%	14.7%
3	16.3%	21.3%
4	21.5%	27.2%
5	26.6%	32.5%

Notes – Share of screens reduced by an age-based policy that raises the minimum screening age from 50, and an algorithmic policy that raises its risk threshold to miss the same number of screen-caught tumors as the corresponding age-based policy.

Table 4: Screening reduces large tumor incidence, particularly among women with high invasive tumor risk.

	Large tumor (per 100,000 person-years)		
	(1)	(2)	(3)
Post	-40.6***	-17.9***	-28.0***
	(3.1)	(4.2)	(5.0)
Post × High invasive tumor risk		-61.1***	
		(8.3)	
Post × High abnormal mammogram risk			-26.2***
			(7.1)
High invasive tumor risk		118.1***	
		(7.6)	
High abnormal mammogram risk			53.8***
			(7.2)
(Intercept)	131.4***	79.1***	104.1***
	(3.6)	(3.7)	(4.5)
Observations	4,208,756	4,208,756	4,208,756
R^2	2.99×10^{-5}	0.0001748	6.16×10^{-5}

Notes – This table reports estimates of the effect of screening policies on large tumor incidence. Each "high risk" variable corresponds to risk score above median. Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 5: Screening increases detection of small tumors.

	Small tumor (per 100,000 person-years)		
	(1)	(2)	(3)
Post	99.4***	64.1***	60.4***
	(4.5)	(4.1)	(5.7)
Post × High invasive tumor risk		52.6***	
		(9.2)	
Post × High abnormal mammogram risk			76.8***
			(8.9)
High invasive tumor risk		102.2***	
		(7.5)	
High abnormal mammogram risk			36.3***
			(7.2)
(Intercept)	129.8***	84.9***	111.1***
	(3.6)	(3.9)	(4.7)
Observations	4,208,756	4,208,756	4,208,756
R^2	8.81×10^{-5}	0.0003398	0.0002106

Notes – This table reports estimates of the effect of screening policies on small tumor incidence. Each "high risk" variable corresponds to risk score above median. Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 6: Screening reduces cancer mortality, particularly among women with high invasive tumor risk.

	Cancer mortality (per 100,000 person-years)		
	(1)	(2)	(3)
Post	-15.0***	-1.7	-3.1
	(2.2)	(2.3)	(2.6)
Post × High invasive tumor risk		-31.8***	
		(4.6)	
Post × High abnormal mammogram risk			-24.2***
			(4.4)
High invasive tumor risk		40.6***	
		(4.3)	
High abnormal mammogram risk			25.7***
			(4.0)
(Intercept)	40.4***	22.6***	27.8***
	(1.1)	(1.1)	(2.3)
Observations	4,208,756	4,208,756	4,208,756
R^2	1.42×10^{-5}	5.28×10^{-5}	2.8×10^{-5}

Notes – This table reports estimates of the effect of screening policies on cancer mortality. Each "high risk" variable corresponds to risk score above median. Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table 7: Estimates of total screens and tumor incidence for additional counterfactual screening policies.

	Outcome per 100,000 person-years	
Policy	Large Tumor	Small Tumor
Everyone at 50	80	185
Everyone at 45	70	193
High-risk at 45, everyone at 50	73	191
Everyone at 40	67	198
High-risk at 40, everyone at 50	68	191
Everyone at 55	91	168
High-risk at 40, everyone at 55	83	175

Notes – This table reports estimates of total screens and tumor incidence per 100,000 person-years under counterfactual screening policies. Using the calibrated clinical oncology model, we estimate each of these quantities for women aged 40-69 from 2008-2019 in the simulated populations.

For Online Publication: Internet Appendix

A. Additional Details on Data and Context

A.1. Breast cancer screening in Denmark

Since the establishment of the National Cancer Register in 1943, breast cancer has remained the most common cancer among women in Denmark (Olsen et al., 2003). Denmark has been slower to introduce breast cancer screening programs than other European countries. 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.[12] 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.[13] 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

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

[13]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).

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, 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 introduction across Denmark was complete by the end of the decade. 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.[14] 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., 2008). 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., 2008; 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., 2008; Jacobsen et al., 2017). During the first four rounds of the nationwide screening program, national

[14]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.

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. Estimates of overdiagnosis, on the other hand, largely depend on 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 and Gøtzsche, 2009; Jørgensen et al., 2017)

A.2. Danish data registers and model training

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.[15] 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. The Danish Population Register provides a snapshot of the population on January 1, starting from 1980, detailing for each person their age and municipality of residence.

We collect medical features from the National Patient Register, National Health Service Register, and National Prescription Drug Register. The dependent variables for each woman in each year are the number of each type of claim filed for in the previous five years. The family history variable is an indicator for whether the woman's mother had an invasive cancer in prior years. We train models using the XGBoost library (Chen and Guestrin, 2016). We split the sample into a 50% training sample and 50% test sample, where the set of women in each sample is disjoint. We select hyperparameters for each model across each outcome variable and each set of features using five-fold cross-validation on the training sample. We conduct a grid search over the maximum tree depth, which we vary from two to five, and the learning rate, which we vary from 0.001 to 0.1 across a logarithmic five-parameter grid. During training, we set the XGBoost parameter for positive weight scaling so that the positive and negative labels are balanced in the training set. The main text reports AUC-ROC scores for each model on the test sample.

B. Additional Results on Long-Run Effects of Screening

In this section, we report additional results that follow the specification of Equation (2).

[15] 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.

Exclude policy introduction period. Table A1, Table A2, and Table A3 re-estimate Equation (2) while excluding observations from 2007-2010 that occurred during the policy introduction period. Invasive tumor risk continues to correspond with reduced large tumor incidence and cancer mortality after policy introduction.

Age controls. Table A4, Table A5, and Table A6 add controls for $\text{Age}_{i,t}$ and $\text{Age}_{i,t} \cdot \text{Post}_t$ to Equation (2). These specifications allow us to estimate differences in effects of screening that are not driven by differences in age. We find similar results in these specifications.

Effects of screening by continuous risk measure. Table A7, Table A8, and Table A9 re-estimate Equation (2) using a continuous measure of risk. Risk scores are standardized to have zero mean and unit variance. We find similar results using this alternative specification of risk.

Table A1: Screening and large tumor incidence (excluding policy introduction years).

	Large tumor (per 100,000 person-years)		
	(1)	(2)	(3)
Post	-34.6***	-14.2***	-17.8***
	(4.1)	(4.5)	(5.3)
Post × High invasive tumor risk		-49.3***	
		(8.3)	
Post × High abnormal mammogram risk			-32.6***
			(8.1)
High abnormal mammogram risk			41.7***
			(7.3)
High invasive tumor risk		93.0***	
		(7.6)	
(Intercept)	114.1***	71.3***	92.8***
	(3.6)	(3.9)	(4.7)
Observations	3,427,656	3,427,656	3,427,656
R^2	2.55×10^{-5}	0.0001271	3.95×10^{-5}

Notes – This table reports estimates of the effect of screening policies on large tumor incidence. The table excludes observations from 2007-2010, which corresponded to the period of policy introduction. Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table A2: Screening and small tumor incidence (excluding policy introduction years).

	Small tumor (per 100,000 person-years)		
	(1)	(2)	(3)
Post	80.8***	62.0***	56.1***
	(4.6)	(5.3)	(6.2)
Post × High invasive tumor risk		28.3***	
		(9.3)	
Post × High abnormal mammogram risk			50.1***
			(9.1)
High invasive tumor risk		79.9***	
		(7.5)	
High abnormal mammogram risk			14.1**
			(7.3)
(Intercept)	114.4***	77.5***	106.7***
	(3.7)	(4.1)	(5.0)
Observations	3,427,656	3,427,656	3,427,656
R^2	7×10^{-5}	0.0002184	0.0001163

Notes – This table reports estimates of the effect of screening policies on small tumor incidence. The table excludes observations from 2007-2010, which corresponded to the period of policy introduction. Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table A3: Screening and cancer mortality (excluding policy introduction years).

	Cancer mortality (per 100,000 person-years)		
	(1)	(2)	(3)
Post	-8.1***	0.7	-2.8
	(2.2)	(2.5)	(2.8)
Post × High invasive tumor risk		-21.6***	
		(4.5)	
Post × High abnormal mammogram risk			-11.1***
			(4.3)
High invasive tumor risk		26.5***	
		(4.0)	
High abnormal mammogram risk			14.1***
			(3.9)
(Intercept)	32.6***	20.4***	25.4***
	(1.1)	(2.1)	(2.5)
Observations	3,427,656	3,427,656	3,427,656
R^2	5.86×10^{-6}	2.45×10^{-5}	1.1×10^{-5}

Notes – This table reports estimates of the effect of screening policies on cancer mortality. The table excludes observations from 2007-2010, which corresponded to the period of policy introduction. Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table A4: Screening and large tumor incidence (age controls).

	Large tumor (per 100,000 person-years)		
	(1)	(2)	(3)
Post	-41.9***	-16.4***	-54.1***
	(4.0)	(5.9)	(8.8)
Post × High invasive tumor risk		-65.9***	
		(12.3)	
Post × High abnormal mammogram risk			24.2
			(15.4)
High invasive tumor risk		130.5***	
		(11.4)	
High abnormal mammogram risk			-14.6
			(14.0)
(Intercept)	132.6***	73.7***	139.1***
	(3.6)	(5.3)	(8.1)
Controls	✓	✓	✓
Observations	4,208,756	4,208,756	4,208,756
R^2	7.51×10^{-5}	0.0001773	7.6×10^{-5}

Notes – This table reports estimates of the effects of screening policies on large tumor incidence. Each "high risk" variable corresponds to risk score above median. This table adds controls for $\text{Age}_{i,t}$ and $\text{Post}_t \times \text{Age}_{i,t}$ to Equation (2). Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table A5: Screening and small tumor detection (age controls).

	Small tumor (per 100,000 person-years)		
	(1)	(2)	(3)
Post	98.1***	95.4***	93.7***
	(4.5)	(6.7)	(9.6)
Post × High invasive tumor risk		-11.7	
		(13.6)	
Post × High abnormal mammogram risk			8.9
			(16.1)
High invasive tumor risk		134.9***	
		(11.7)	
High abnormal mammogram risk			7.1
			(13.5)
(Intercept)	130.5***	69.7***	126.9***
	(3.6)	(5.4)	(7.7)
Controls	✓	✓	✓
Observations	4,208,756	4,208,756	4,208,756
R^2	0.0002446	0.0003570	0.0002453

Notes – This table reports estimates of the effects of screening policies on small tumor incidence. Each "high risk" variable corresponds to risk score above median. This table adds controls for $\text{Age}_{i,t}$ and $\text{Post}_t \times \text{Age}_{i,t}$ to Equation (2). Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table A6: Screening and cancer mortality (age controls).

	Cancer mortality (per 100,000 person-years)		
	(1)	(2)	(3)
Post	-15.5***	-3.5	-8.9**
	(2.2)	(3.2)	(4.4)
Post × High invasive tumor risk		-28.2***	
		(6.5)	
Post × High abnormal mammogram risk			-13.2*
			(7.7)
High invasive tumor risk		38.4***	
		(6.1)	
High abnormal mammogram risk			5.6
			(6.9)
(Intercept)	40.9***	23.6***	38.1***
	(2.0)	(2.9)	(3.1)
Controls	✓	✓	✓
Observations	4,208,756	4,208,756	4,208,756
R^2	3.24×10^{-5}	5.31×10^{-5}	3.37×10^{-5}

Notes – This table reports estimates of the effects of screening policies on cancer mortality. Each "high risk" variable corresponds to risk score above median. This table adds controls for $\text{Age}_{i,t}$ and $\text{Post}_t \times \text{Age}_{i,t}$ to Equation (2). Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table A7: Screening and large tumor incidence (standardized risk).

	Large tumor (per 100,000 person-years)		
	(1)	(2)	(3)
Post	-40.6***	-54.4***	-41.5***
	(3.1)	(4.3)	(3.1)
Post × Invasive tumor risk (standardized)		-56.1***	
		(5.4)	
Post × Abnormal mammogram risk (standardized)			-13.4***
			(4.1)
Invasive tumor risk (standardized)		91.8***	
		(5.0)	
Abnormal mammogram risk (standardized)			27.8***
			(3.8)
(Intercept)	131.4***	143.7***	132.2***
	(3.6)	(3.1)	(3.6)
Observations	4,208,756	4,208,756	4,208,756
R^2	2.99×10^{-5}	0.0003047	6.37×10^{-5}

Notes – This table reports estimates of the effects of screening policies on large tumor incidence. Each risk score is standardized. Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table A8: Screening and small tumor incidence (standardized risk).

	Small tumor (per 100,000 person-years)		
	(1)	(2)	(3)
Post	99.4***	84.2***	98.5***
	(4.5)	(4.7)	(4.5)
Post × Invasive tumor risk (standardized)		15.5***	
		(5.8)	
Post × Abnormal mammogram risk (standardized)			42.9***
			(4.7)
Invasive tumor risk (standardized)		83.1***	
		(4.9)	
Abnormal mammogram risk (standardized)			16.1***
			(3.8)
(Intercept)	129.8***	140.9***	130.2***
	(3.6)	(3.9)	(3.6)
Observations	4,208,756	4,208,756	4,208,756
R^2	8.81×10^{-5}	0.0005314	0.0002221

Notes – This table reports estimates of the effects of screening policies on small tumor incidence. Each risk score is standardized. Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

Table A9: Screening and cancer mortality (standardized risk).

	Cancer mortality (per 100,000 person-years)		
	(1)	(2)	(3)
Post	-15.0***	-18.1***	-15.4***
	(2.2)	(2.4)	(2.2)
Post × Invasive tumor risk (standardized)		-23.6***	
		(2.9)	
Post × Abnormal mammogram risk (standardized)			-13.1***
			(2.4)
Invasive tumor risk (standardized)		28.3***	
		(2.8)	
Abnormal mammogram risk (standardized)			14.6***
			(2.2)
(Intercept)	40.4***	44.2***	40.8***
	(1.1)	(2.2)	(2.0)
Observations	4,208,756	4,208,756	4,208,756
R^2	1.42×10^{-5}	7.76×10^{-5}	3.16×10^{-5}

Notes – This table reports estimates of the effects of screening policies on cancer mortality. Each risk score is standardized. Standard errors are clustered by person, and *,**, and *** denote statistical significance at the 10%, 5%, and 1% level, respectively.

C. Clinical Oncology Model of Breast Cancer Development

The Erasmus model (Tan et al., 2006) is a clinical model of cancer development. We use the model to simulate the health effects of counterfactual screening policies. The model considers a population of women, starting at age 20, who may develop cancer. At onset, cancer is either invasive or non-invasive. Non-invasive cancers are also called ductal carcinoma in situ (DCIS). DCIS cancers are either DCIS-regressive (they do not become harmful), DCIS-invasive (they eventually become invasive), or DCIS-clinical (they do not become harmful, but are eventually detected clinically). Screening can detect all types of tumors before they become clinically detected. For the underlying clinical model, we largely follow the parameterization of Einav et al. (2020), with three points of departure. First, we calibrate probability of death of other causes to Danish mortality statistics released by Statistics Denmark. Second, we do not assume that mammography demand depends on underlying breast cancer stage, as opportunistic screening is rare and compliance is high in Denmark (Jensen et al., 2005). Third, we calibrate probability of cancer onset as a function of risk scores.

We modify the parameterization of cancer onset to vary with risk scores. First, we match the empirical evolution of risk scores across our population. Second, we calibrate tumor incidence in each year in the model to be a function of risk score. We extrapolate our empirical measure of cancer risk, which is computed for women aged 50-69 for whom we have medical information, to women aged 20 and older using a linear function. We linearly fit the mean risk of women aged 50-69 in our sample as a function of age. We assume the standard deviation of risk is the same for each age, and parameterize it using the standard deviation of risk scores in our sample. Using this linear calibration, we then extrapolate the distribution of risk scores to women aged 20 and assume each woman is endowed with a risk score at age 20. While our empirical risk score is for invasive cancer, which is observable, and not for tumor onset, which is unobservable, the extrapolation follows the assumption in the Erasmus model that the risk of tumor onset is proportional to the risk of invasive tumor incidence. Next, we calibrate tumor incidence as a

function of risk score. We assume the probability of onset is a linear function of risk score, so that $P(\text{onset}_{i,t}) = a + b \times \text{risk}_{i,t}$. The two parameters to calibrate are a and b, which we search over the grid formed by $a \in [0, 0.001]$ and $b \in [0, 0.01]$, with 21 candidates for each parameter. Our calibrated model has the parameters $a = 0.00025$ and $b = 0.0080$. We assume that the screening policy starts in 2008. We match screening compliance after the policy introduction to reported Danish compliance rates (Lynge et al., 2017). Each parameterization is simulated over a panel of 10 million women. We calibrate the model to match 48 moments: The total incidence of small tumors from 2004-2019, the incidence of large tumors from 2004-2019 for women with above-median risk, and the incidence of large tumors from 2004-2019 for women with below-median risk.

For counterfactual policy estimates, we simulate panels of 10 million women under varying policy regimes. The minimum universal screening age is set to 40, 45, 50, or 55. For policies that combine both minimum universal screening ages and risk-based screening ages, we set the risk cutoff for younger women to be the median risk of women at the minimum universal screening age cutoff. The main text reports total screening and tumor incidence per 100,000 person-years in the simulated panel from 2008-2019 across women aged 40-69.