# Using Health Economics Tools to Enhance the Clinical Utility of Artificial Intelligence-Based Diagnostics: A Case Study in Breast Cancer Screening

Nathaniel Hendrix

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

David L. Veenstra, Chair

Aasthaa Bansal

A. Brett Hauber

Christoph I. Lee

Program Authorized to Offer Degree:

Pharmacy

University of Washington

**Abstract**

Using Health Economics Tools to Enhance the Clinical Utility of Artificial Intelligence-Based Diagnostics: A Case Study in Breast Cancer Screening

Nathaniel Hendrix

Chair of the Supervisory Committee:
David L. Veenstra
Professor
Department of Pharmacy

Researchers in artificial intelligence (AI) have recently produced several products for medical diagnosis that perform at the same level as human clinicians. Artificial intelligence products will also need to be developed that will be trusted by clinicians and are known to produce positive effects in patients. One important area where AI may be applied is breast cancer screening which, despite its benefits, currently harms many women through false positives and overdiagnosis. This dissertation involved the use of two tools from health economics – discrete choice experiments and outcomes modeling – to solve translational issues affecting AI, all in the setting of breast cancer screening. In the first aim, we assessed primary care providers' (PCPs') preferences for a hypothetical AI system for mammogram interpretation. We used qualitative interviewing to develop a discrete choice instrument, which we administered online to ninety-one PCPs from around the United States. While advances in improving AI's diagnostic accuracy were important to respondents, they also reported valuing the diversity of training data and understandability of AI decision-making. The surveyed PCPs were broadly accepting of using AI to "triage" likely negative screens, so that radiologists do not need to interpret every image. In the second aim, we used outcomes modeling to compare the performance of 28 AI algorithms that had been developed for breast cancer screening. We first performed receiver operating characteristic (ROC) curve analysis to get a conventional metric (area under the curve) for model comparison. We then used a model of breast cancer screening and outcomes to estimate the quality-adjusted life years (QALYs) associated with using each model at its optimal operating point. These outcomes were compared with the outcomes associated with using two other methods of operating point selection – Youden's index and decision curve analysis. Outcomes modeling ranked algorithms in the same order as area under ROC curve and did not produce substantially different outcomes at the QALY-optimizing operating point compared to the use of decision curve analysis. This suggests that outcomes modeling may be most useful in model comparison and operating point selection when detailed data including case heterogeneity is available.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

I would first like to express how grateful I am to all the members of my committee. The expertise that each of you brought to this project has allowed me to produce a dissertation that I can be very proud of and that I know will have an impact on patient care.

Many mentors have believed in me and helped me to get to where I am. Beth Devine has been a consistent supporter of my curiosity about methods and has given me more opportunities to put them into practice than I can count. Seeing the work that Deborah Atherly does at PATH was what sealed the deal on my joining the PhD program and her support has been invaluable ever since. My peers at CHOICE, too, have been an inspiration to always do my best and to keep finding new ways to challenge myself.

Finally, thanks to my partner, Mark, for his love, patience, and support.

# Chapter 1. INTRODUCTION

Artificial intelligence (AI) has the potential to solve many pressing problems in medicine. Its scalability means that high-quality medical care can be made more accessible at a lower cost.[1] It is also immune to fatigue and burnout, which can lead to medical errors.[2] The past decade's advances in neural networks have meant that image analysis is now a feasible task for AI.[3] These advances have, in turn, led to the United States Food & Drug Administration's approval of several AI-based technologies for diagnosis.[4] We designed this dissertation's two aims to solve unresolved translational issues in AI with the tools of health economics so that AI's benefits can be made available to patients as soon as possible.

The first aim is a discrete choice experiment (DCE) designed to test primary care providers' (PCPs') preferences for different attributes of hypothetical AI systems for use in breast cancer screening. We first used qualitative interviewing with four PCPs to determine which attributes should be included in the experiment. We then constructed an instrument, validated it and optimized its presentation with four user interviews, and administered it online to a national sample of PCPs from university-affiliated practices. Each respondent's choices were recorded for a series of fifteen tasks that comprised a choice between two AI products and screening by radiologist alone. The respondents' choices reveal the relative importance of each attribute of AI as well as how good an AI would need to be in order to recommended over radiologist alone. We analyzed these choices using two methods: first, we used a random parameters logit model to control for the heterogeneity of preferences among our sample; then, we used a latent class model to explain that heterogeneity by assigning each respondent into a number of classes that describe decision-making strategies.

One potential benefit of AI for breast cancer screening over radiologists is its flexibility. We can shift AI's "operating point" between positive and negative screens such that it optimizes patient outcomes, accommodates different risk preferences, or better suits different populations. Conventional methods of selecting this operating point are generally grounded in heuristics that do not consider all of the complexities that an outcomes model can. In this aim, we compared outcomes modeling to two methods of operating point selection and one method of model performance comparison in order to assess the potential effects on patient outcomes of 28 AI

mammogram interpretation models that were submitted to a contest. We compared the ranking of maximum quality-adjusted life years (QALYs) from each model to the ranking by area under the receiver operating characteristic curve for model comparison. We then compared the QALYs gained at the operating point selected by outcomes modeling versus Youden's index and decision curve analysis.[5,6]

# Chapter 2. ARTIFICIAL INTELLIGENCE IN BREAST CANCER SCREENING: PRIMARY CARE PROVIDER PREFERENCES

The capacity of breast cancer screening (BCS) to reduce cancer mortality has been well-proven.[7] Its high false positive rate, however, means that participating women have a substantial probability of undergoing unnecessary medical care.[8–10] Using artificial intelligence (AI) to interpret screening mammograms has been proposed as a way of lowering the false positive rate while maintaining the benefits of BCS to cancer mortality.[11,12] The US Food & Drug Administration's first approval of AI BCS took place in 1998 for a technology called computer-aided detection, which later studies revealed to be ineffective in clinical settings.[13–15] Recent advances in AI BCS such as convolutional neural networks and deep learning have led to the approval of a new generation of algorithms that promise to detect more cancers, reduce radiologist burden by triaging negative screens out of their workflow, and to predict future cancer risk.[4,16,17]

Despite the current environment of optimism around AI BCS, substantial challenges with its integration into the clinic remain unsolved. For example, some of the most advanced methods in AI produce results without any understandable interpretation of how it arrived at them, potentially lowering trust and exposing clinicians to liability for errors.[18,19] The quality of trials for clinical AI has been low, perhaps exacerbated by the fact that these trials are not required for FDA approval.[20] Finally, with many AI developers focusing on the availability of data rather than its representativeness, AI may exacerbate existing health disparities.[21,22] For these reasons and more, simply improving AI's accuracy to match or exceed that of human clinicians is only a first step in its implementation.[23] Gaining the trust of multiple stakeholders from clinicians to patients, payers, and regulators is essential before the promised benefits of AI BCS can reach patients.[24]

In order to understand how AI developers can best meet the needs of one key stakeholder, we sought to quantify the relative importance of different attributes of AI BCS to primary care providers (PCPs) who commonly order screening mammography. We conducted a discrete choice experiment (DCE) among PCPs using hypothetical AI BCS algorithms with a range of different performance characteristics. Discrete choice experiments are used to quantify the strength of stakeholders' preferences for different attributes of a technology, and their results can be applied

to cost-effectiveness analyses, research prioritization, and regulatory decisions.[25,26] We chose to focus on PCPs because patients often make breast cancer screening decisions with their PCPs. Our hypothesis was that sensitivity (ability to detect additional cancers) and the understandability of decisions made by AI would be the most important attributes to ordering PCPs.

## 2.1 METHODS

We used a DCE to quantify PCP's preferences for attributes of AI BCS by presenting them with a series of choices between two hypothetical AI products and screening by radiologist alone. We first used qualitative interviewing to determine what the most important attributes are, then constructed choice questions following an experimental design including those attributes, and distributed the DCE as part of a survey to a sample of PCPs from across the US. We analyzed responses in two ways: first, with a random parameters logit model to control for preference heterogeneity on an individual level, and then with a latent class model to account for heterogeneity by assigning respondents to classes of similar decision makers. The study was approved by the University of Washington Human Subjects Division.

### 2.1.1 *Instrument Development*

We conducted a literature search to identify candidate attributes and levels for the hypothetical products described in the DCE, which we then refined through a series of four qualitative interviews with PCPs. The final set of six attributes included sensitivity, specificity, radiologist involvement, understandability of AI decision-making, supporting evidence, and diversity of training data. These attributes are shown with their respective levels in Table 2.1. In order to provide a basis of comparison to participants, we included notes next to the sensitivity and specificity attributes indicating that the average U.S. radiologist misses approximately 15% of cancers and has an 11% false-positive rate.[27]

Table 2.1. Attributes and levels from the discrete choice experiment

| Attribute | Levels |
|---|---|
| Sensitivity | • Misses 6% of cancers<br>• Misses 11% of cancers<br>• Misses 15% of cancers |
| Specificity | • 6% of women without cancer receive a false positive<br>• 11% of women without cancer receive a false positive<br>• 15% of women without cancer receive a false positive |
| Radiologist involvement | • All images reviewed by radiologist<br>• 30% of images most likely to contain cancer reviewed by radiologist<br>• No images reviewed by radiologist |
| Transparency of artificial intelligence (AI) decision-making | • Decision-making rationale understandable by clinicians<br>• Decision-making rationale understandable by AI experts only<br>• Decision-making rationale not understandable |
| Supporting evidence | • Supported by both observational data and randomized controlled trial (RCT)<br>• Supported by RCT<br>• Supported by observational data |
| Diversity of training and validation data | • 50% of patients are well-represented in data<br>• 75% of patients are well-represented in data<br>• 100% of patients are well-represented in data |

Following the identification of the final set of attributes, pretest interviews were conducted with a separate sample of four PCPs to ensure that the content was presented in an understandable fashion. We then constructed a DCE comprising 15 choice tasks, each presenting two hypothetical BCS alternatives defined by varying levels of the six attributes such that the participant must trade off better performance on some attributes for worse performance on others. The levels in each alternative for each choice task were determined by an experimental design developed in SAS 9.4 (Cary, North Carolina) using D-efficiency criteria that maximize the information gathered by the DCE.[28,29]

Users had the opportunity to indicate after each choice task whether they preferred mammography interpretation by radiologist alone to both AI-based products described in the choice task, thereby allowing us to estimate the desirability of the status quo relative to the

introduction of AI. To control for excluded attributes, participants were asked to imagine that all alternatives cost the same and that all positive screens followed the same pattern of diagnostic work-up. Figure 2.1 contains a sample task from the DCE.

**Which artificial intelligence-based product would you say is better?**

* must provide value

○ Product A
○ Product B

reset

| **Product characteristic** | **Product A** | **Product B** |
|---|---|---|
| *Sensitivity (radiologists miss 15% of cancers)* | Misses 15% of cancers | Misses 6% of cancers |
| *Specificity (11% of women without cancer receive a false-positive with radiologist)* | 6% of women without cancer receive false-positive | 15% of women without cancer receive false-positive |
| *Radiologist involvement* | Radiologist confirms all screens | 30% of mammograms determined by artificial intelligence to be most likely to contain cancer sent for radiologist confirmation |
| *Supporting evidence* | New product with supporting evidence from both observational data analysis and RCT | New product with supporting evidence from RCT |
| *Transparency of decisions by artificial intelligence* | Decision-making by artificial intelligence is transparent and understandable by clinicians | Decision-making by artificial intelligence is understandable only by experts in artificial intelligence |
| *Representativeness of training and validation data* | 50% of your patients are well-represented in the training and validation data | 75% of your patients are well-represented in the training and validation data |

**Which would you recommend to your patient?**

* must provide value

○ Your preferred artificial intelligence product above
○ Screening by radiologist alone

reset

Figure 2.1. A sample task from the discrete choice experiment

### 2.1.2 *Study Population*

Physicians and nurse practitioners working as PCPs were recruited for qualitative and pretest interviews from a convenience sample of the authors' professional networks. We calculated a minimum sample size of 30 for estimating main effects in the DCE.[30] We randomly selected 350 physician and nurse practitioners working as PCPs around the US to receive a mailed invitation to take part in the internet-based DCE. The names and work addresses of potential participants were gathered from the websites of university-affiliated practices, which we identified using a search engine. Snowball recruitment was also allowed.

### 2.1.3 *Statistical Analysis*

The outcome of interest in our statistical analysis was the impact of each attribute level on the probability that PCPs would recommend AI-augmented screening over screening by radiologist alone. We first used a random parameters logit model, which controls for the heterogeneity in respondents' preferences. These preferences, in turn, indicate the amount of value a given individual derives from a change in the level of one or more attributes. We included in the regression an alternative-specific constant for the option to recommend radiologist alone. This was intended to account for unobserved attributes associated with screening by radiologist alone and to test for status quo bias. All attributes of each alternative were modeled as categorical variables using effects coding.

We next used a latent class model to identify latent preference segments in the data. The latent class model allows us to explain the heterogeneity that had been controlled for in the random parameters logit model.[31] We began with a 2-class model and increased the number of classes until the model no longer converged. The model included the same utility specification as the random parameters logit model, plus the respondent-level covariates collected in the demographic questions after the DCE. Both regressions were conducted in Stata 16.1 (College Station, Texas).

We calculated the relative importance of each attribute by subtracting the coefficient of the least-preferred level for that attribute from the coefficient of the most-preferred level of that attribute. Then, each difference was divided by the sum of differences for all six attributes to arrive at a percentage describing relative importance.

We modeled baseline preference for AI in both regressions with the following formula:

$$Pr(AI) = \frac{1}{1 + e^{-ASC}}$$

where ASC is the coefficient of the alternative-specific constant for recommending radiologist alone.[32,33]

The influence of each attribute on the willingness to recommend AI-based screening was tested by entering their respective coefficients, one-by-one, into the following formula:

$$Pr(AI) = \frac{1}{1 + e^{-(\beta - ASC)}}$$

where β is a coefficient from the regression results. We report from the random parameters logit model the difference in willingness to recommend AI-augmented screening versus screening

by radiologist alone based on attribute levels. From the latent class model, we report the relative importance to each class of each attribute.

## 2.2 RESULTS

### 2.2.1 *Participants*

Ninety-one PCPs responded to the survey.[34] Among these, 22 responded to mailed invitations, and 69 took part in response to snowball sampling. An overall response rate cannot be calculated due to the use of snowball sampling, but 6% of recipients of mailed invitations had responded before we closed data collection. Respondents had a mean of 8.8 (standard deviation: 9.8) years of clinical practice. A plurality (44.0%) was from the Western US region, and a majority (62.6%) identified as female. Generally, respondents expressed a neutral attitude toward technology with 60 (65.9%) describing themselves as using technology when others around them do. They reported mostly positive experiences working with radiologist colleagues. Asked about the ease of following up on questions with a radiologist, 73 (80.2%) said that it was somewhat or very easy. Similarly, 82 (90.1%) reported having a somewhat or very high level of trust in their radiologist colleagues. Detailed respondent characteristics are available in **Table 2.2**.

Table 2.2. Respondent characteristics (n = 91)

| **Parameter** | **n (%)** |
|---|---|
| Attitude toward technology | |
| *I love new technology…* | 3 (3.3)% |
| *I like new technology…* | 19 (20.9%) |
| *I use technology when others around me do* | 60 (65.9%) |
| *I am usually one of the last to use a new technology* | 6 (6.6%) |
| *I am skeptical of new technology…* | 3 (3.3%) |
| Ease of contacting radiologist colleagues | |
| *Very easy* | 37 (40.7%) |
| *Somewhat easy* | 36 (39.6%) |
| *Neither easy nor difficult* | 9 (9.9%) |
| *Somewhat difficult* | 9 (9.9%) |
| *Very difficult* | 0 (0%) |
| Trust in radiologist colleagues | |
| *Very high* | 39 (42.9%) |
| *Somewhat high* | 43 (47.3%) |
| *Moderate* | 8 (8.8%) |
| *Somewhat low* | 1 (1.1%) |
| *Very low* | 0 (0%) |
| Region | |
| *Midwest* | 24 (26.4%) |
| *Northeast* | 10 (11.0%) |
| *South* | 17 (18.7%) |
| *West* | 40 (44.0%) |
| Gender | |
| *Female* | 57 (62.6%) |
| *Male* | 33 (36.3%) |
| *Other* | 1 (1.1%) |
| Years of clinical practice | Mean: 8.8 (Range: 0 – 44) |

Two respondents (2.2%) always selected the alternative with higher sensitivity, and one respondent (1.1%) always selected the alternative with higher specificity. Four (4.4%) participants always chose radiologist alone over both AI products.

### 2.2.2 *Effect of AI Attributes on Screening Recommendation*

We found in the random parameters logit model that the attribute most likely to decrease PCPs' willingness to recommend AI was a lack of improvement in sensitivity over radiologist

alone: this reduced the probability of recommending AI by 0.36 (95% confidence interval (CI): 0.31 to 0.38). Improving sensitivity by 9 percentage points over radiologist alone, on the other hand, was the attribute that most increased willingness to recommend AI. This attribute was associated with an increase of 0.36 (95% CI: 0.29 to 0.42) in the probability of recommending AI.

The changes in probability of recommending AI associated with all attributes can be seen in **Figure 2.2**, with full regression coefficients available in **Table 2.3**. Respondents were equally likely to recommend an AI-based product if all its decisions were double-checked by a radiologist and if the AI alone interpreted negative screens without a radiologist ever examining the images. Finally, supporting AI-augmented screening with both observational studies and randomized controlled trials was not preferred over randomized controlled trials alone.

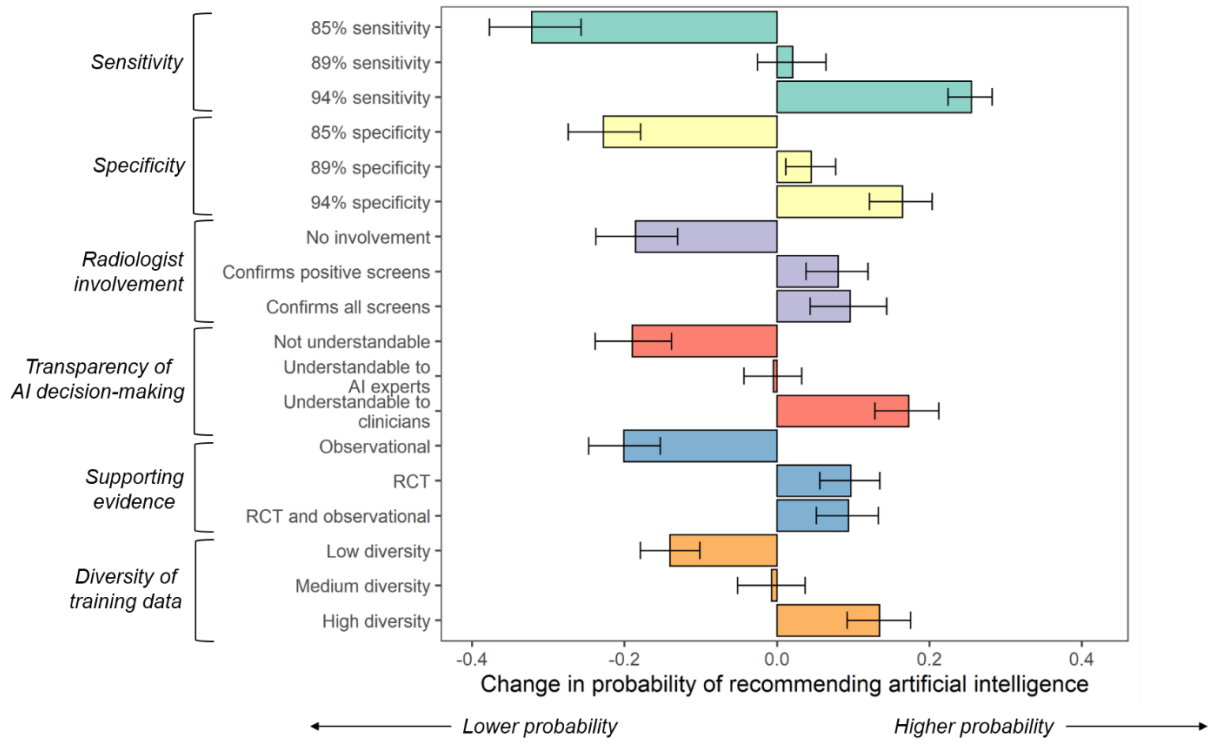


Figure 2.2. Change in probability of recommending artificial intelligence (AI)-augmented breast cancer screening over radiologist alone, given different performance attributes of the AI product.

Table 2.3. Coefficients (with 95% confidence intervals) from the random parameters logit and 3-class latent class models

| Attribute | Random Parameters Logit Model | Latent Class Model | | |
|---|---|---|---|---|
| | | **Class 1: "Sensitivity First"** | **Class 2: "Against AI Autonomy"** | **Class 3: "Uncertain Trade-Offs"** |
| Radiologist alone (alternative-specific constant) | -0.32 (-0.65 to 0.02) | 0.74 (0.45 to 1.04)** | 0.51 (-0.12 to 1.13) | 0.97 (0.70 to 1.24)** |
| 85% Sensitivity | -1.38 (-1.70 to -1.06)** | -1.73 (-2.10 to -1.37)** | -1.21 (-1.83 to -0.60)** | -0.58 (-0.92 to -0.25)** |
| 89% Sensitivity | 0.08 (-0.10 to 0.27) | 0.21 (-0.07 to 0.49) | 0.34 (-0.21 to 0.90) | -0.12 (-0.41 to 0.17) |
| 94% Sensitivity | 1.30 (1.09 to 1.51)** | 1.53 (1.32 to 1.74)** | 0.89 (0.54 to 1.25)** | 0.70 (0.42 to 0.98)** |
| 85% Specificity | -0.93 (-1.14 to -0.72)** | -0.66 (-0.90 to -0.43)** | -1.19 (-1.59 to -0.79)** | -0.59 (-0.90 to -0.29)** |
| 89% Specificity | 0.19 (0.05 to 0.33)* | -0.05 (-0.29 to 0.18) | 0.04 (-0.34 to 0.43) | 0.26 (-0.03 to 0.55) |
| 94% Specificity | 0.75 (0.53 to 0.96)** | 0.71 (0.45 to 0.98)** | 1.15 (0.55 to 1.76)** | 0.33 (0.08 to 0.58)* |
| No radiologist involvement | -0.75 (-0.98 to -0.53)** | -0.30 (-0.55 to -0.04)* | -1.52 (-2.03 to -1.01)** | -0.78 (-1.09 to -0.46)** |
| Radiologist confirms likely positives | 0.34 (0.16 to 0.52)** | 0.29 (0.08 to 0.50)* | 0.23 (-0.10 to 0.56) | 0.30 (0.00 to 0.59) |
| Radiologist confirms all screens | 0.41 (0.18 to 0.64)** | 0.01 (-0.23 to 0.24) | 1.29 (0.71 to 1.87)** | 0.48 (0.21 to 0.75)** |
| No transparency | -0.77 (-0.98 to -0.56)** | -0.60 (-0.87 to -0.33)** | -0.79 (-1.23 to -0.35)** | -0.67 (-1.00 to -0.34)** |
| Understandable to AI experts only | -0.02 (-0.18 to 0.13) | -0.09 (-0.29 to 0.12) | 0.13 (-0.19 to 0.45) | 0.04 (-0.23 to 0.31) |
| Understandable to clinicians | 0.79 (0.56 to 1.01)** | 0.68 (0.41 to 0.95)** | 0.65 (0.18 to 1.12)* | 0.63 (0.35 to 0.92)** |
| Observational data | -0.82 (-1.02 to -0.62)** | -0.77 (-1.04 to -0.51)** | -0.82 (-1.23 to -0.42)** | -0.63 (-0.96 to -0.30)** |
| Randomized controlled trial (RCT) | 0.42 (0.24 to 0.60)** | 0.49 (0.24 to 0.74)** | 0.76 (0.08 to 1.44)* | 0.05 (-0.23 to 0.33) |
| RCT + observational data | 0.40 (0.22 to 0.59)** | 0.28 (0.07 to 0.49)* | 0.07 (-0.30 to 0.44) | 0.58 (0.29 to 0.86)** |
| Low diversity | -0.57 (-0.72 to -0.41)** | -0.37 (-0.58 to -0.15)** | -0.32 (-0.63 to 0.00) | -0.90 (-1.26 to -0.55)** |
| Medium diversity | -0.03 (-0.21 to 0.15) | -0.25 (-0.50 to 0.00) | -0.22 (-0.83 to 0.39) | 0.30 (0.01 to 0.59)* |
| High diversity | 0.59 (0.39 to 0.80)** | 0.62 (0.34 to 0.89)** | 0.54 (-0.09 to 1.17) | 0.60 (0.30 to 0.91)** |
| **Class membership** | | **Class 1: "Sensitivity First"** | **Class 2: "Against AI Autonomy"** | **Class 3: "Uncertain Trade-Offs"** |
| Practice years | | 0.15 (0.03 to 0.27)* | 0.10 (-0.05 to 0.24) | Ref. |
| Negative (versus neutral or positive) attitude toward new technologies | | -28.16 (-74.22 to 17.90) | -15.00 (-23.47 to -6.53)** | Ref. |
| Moderate to high (versus low) difficulty in contacting radiologist colleagues | | -0.68 (-2.48 to 1.12) | 0.44 (-1.81 to 2.68) | Ref. |
| Moderate to high (versus low) level of trust in radiologist colleagues | | 39.41 (incalculable CI) | 7.66 (incalculable CI) | Ref. |
| Urban (versus rural) practice environment | | 0.02 (-1.64 to 1.68) | -0.64 (-2.87 to 1.59) | Ref. |
| Midwest (versus West) region | | -2.63 (-4.41 to -0.84)** | -0.46 (-2.85 to 1.93) | Ref. |
| Northeast (versus West) region | | -2.11 (-4.58 to 0.36) | 0.05 (-2.84 to 2.94) | Ref. |
| South (versus West) region | | 10.32 (0.12 to 20.51)* | 15.50 (7.41 to 23.58)** | Ref. |
| Female | | -2.85 (-4.95 to -0.75)* | -2.35 (-5.16 to 0.45) | Ref. |
| **Goodness of Fit** | | | | |
| Log-likelihood | -1051.4 | -1031.8 | | |
| AIC | 2154.9 | 2185.8 | | |
| BIC | 2391.1 | 2545.9 | | |
| R-squared | 0.109 | 0.156 | | |

### 2.2.3 *Classes of Decision Makers*

We successfully modeled 2- and 3-class versions of the latent class model; the 4-class version did not converge. Both operational models had similar goodness-of-fit, but the 3-class model offered more information. Thus, we present the relative importance of attributes for the 3-class model here (**Figure 2.3**). The regression coefficients for both models and relative importance of attributes for the 2-class model are in **Table 2.4**.

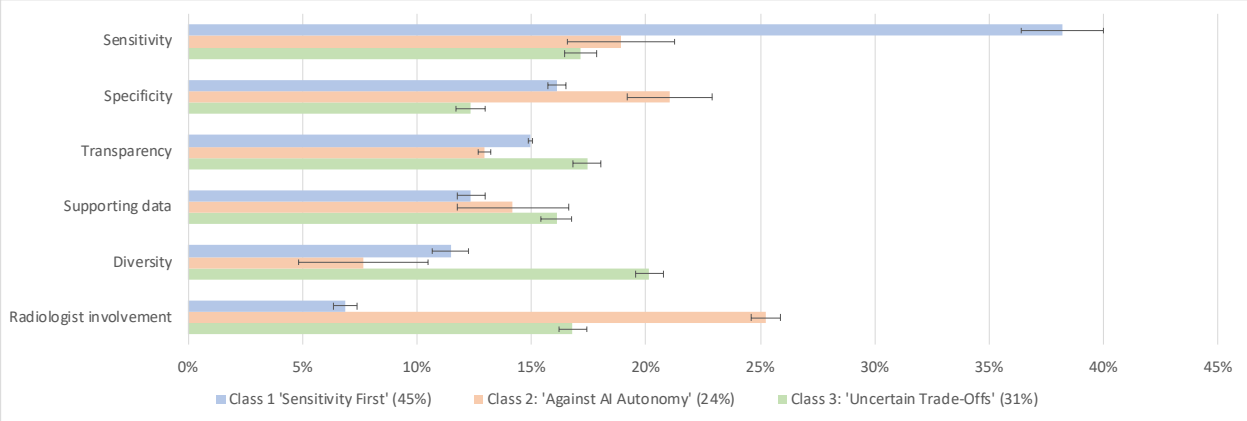


Figure 2.3. Change in probability of recommending artificial intelligence (AI)-augmented breast cancer screening over radiologist alone, given different performance attributes of the AI product.

Table 2.4. Coefficients (with 95% confidence intervals) from the 2-class latent class model

| Attribute | Latent class model | |
|---|---|---|
| | **Class 1: "Sensitivity First"** | **Class 2: "Uncertain Trade-Offs"** |
| Radiologist alone (alternative-specific constant) | 0.55 (0.32 to 0.78)** | 0.93 (0.68 to 1.19)** |
| 85% Sensitivity | -1.46 (-1.75 to -1.17)** | -0.53 (-0.58 to -0.21)** |
| 89% Sensitivity | 0.21 (-0.02 to 0.43) | -0.14 (-0.42 to 0.14) |
| 94% Sensitivity | 1.25 (1.09 to 1.42)** | 0.67 (0.40 to 0.95)** |
| 85% Specificity | -0.77 (-0.94 to -0.61)** | -0.60 (-0.90 to 0.31)** |
| 89% Specificity | -0.03 (-0.21 to 0.16) | 0.28 (-0.01 to 0.56) |
| 94% Specificity | 0.80 (0.60 to 1.00)** | 0.33 (0.08 to 0.58)* |
| No radiologist involvement | -0.63 (-0.83 to -0.42)** | -0.86 (-1.18 to -0.54)** |
| Radiologist confirms likely positives | 0.26 (0.10 to 0.42)** | 0.20 (-0.09 to 0.49) |
| Radiologist confirms all screens | 0.37 (0.18 to 0.56)** | 0.66 (0.39 to 0.93)** |
| No transparency | -0.60 (-0.81 to -0.40)** | -0.63 (-0.94 to -0.31)** |
| Understandable to AI experts only | -0.04 (-0.19 to 0.12) | 0.01 (-0.25 to 0.28) |
| Understandable to clinicians | 0.64 (0.43 to 0.84)** | 0.61 (0.34 to 0.89)** |
| Observational data | -0.72 (-0.91 to -0.52)** | -0.65 (-0.97 to -0.32)** |
| Randomized controlled trial (RCT) | 0.52 (0.31 to 0.74)** | 0.05 (-0.23 to 0.33) |
| RCT + observational data | 0.19 (0.03 to 0.35)* | 0.59 (0.32 to 0.87)** |
| Low diversity | -0.30 (-0.46 to -0.13)** | -0.87 (-1.21 to -0.53)** |
| Medium diversity | -0.20 (-0.41 to 0.02) | 0.31 (0.02 to 0.59)* |
| High diversity | 0.49 (0.27 to 0.72)** | 0.57 (0.27 to 0.86)** |
| **Class membership** | **Class 1: "Sensitivity First"** | **Class 2: "Uncertain Trade-Offs"** |
| Practice years | 0.14 (0.04 to 0.25)* | Ref. |
| Negative (versus neutral or positive) attitude toward new technologies | -2.92 (-5.73 to -0.12)* | Ref. |
| Moderate to high (versus low) difficulty in contacting radiologist colleagues | -0.72 (-2.25 to 0.82) | Ref. |
| Moderate to high (versus low) level of trust in radiologist colleagues | 14.20 (-561.18 to 589.58) | Ref. |
| Urban (versus rural) practice environment | -0.14 (-1.60 to 1.33) | Ref. |
| Midwest (versus West) region | -2.20 (-3.84 to -0.56)* | Ref. |
| Northeast (versus West) region | -1.55 (-4.08 to 0.98) | Ref. |
| South (versus West) region | -0.32 (-2.39 to 1.76) | Ref. |
| Female | -2.84 (-4.41 to -1.26)** | Ref. |
| **Goodness of Fit** | | |
| Log-likelihood | -1062.4 | |
| AIC | 2196.8 | |
| BIC | 2424.3 | |
| R-squared | 0.165 | |

Each class had distinct preferences for the importance of each attribute. Respondents had a 45% probability of assignment to Class 1, which we call "Sensitivity First." Sensitivity was at least twice as important as any other attribute for members of this class. Additionally, members of the "Sensitivity First" class preferred radiologists confirming only likely positives to confirming all screens. Members of Class 2, which we refer to as "Against AI Autonomy," valued radiologist confirmation of all screens as the most important attribute. Sensitivity was not significantly more important than specificity for this class, to which respondents had a 24% probability of being assigned. Finally, we call Class 3 the "Uncertain Trade-Offs" group, because all attributes were of similar importance to these respondents. As such, what could convince them to recommend AI-augmented screening was less clear than for other classes. Respondents had a 31% probability of assignment to this class. Members of the "Sensitivity First" and "Uncertain Trade-Offs" classes did not significantly prefer radiologist confirmation of all images over radiologist confirmation only of likely positives.

Clinicians with more years in practice were significantly more likely to be in the "Sensitivity First" group compared to the "Uncertain Trade-Offs" group (odds ratio: 1.16 per additional year [95% CI: 1.03 to 1.31]). Clinicians who stated that they have a neutral to positive attitude toward technology were more likely to belong to the "Against AI Autonomy" group than to the "Uncertain Trade-Offs" group (odds ratio: $3.05 * 10^{-7}$ [95% CI: $6.88 * 10^{-11}$ to $1.46 * 10^{-3}$]). Most other predictors of class membership were non-significant. **Table 2.3** shows all coefficients from the class membership model.

## 2.3 DISCUSSION

This study demonstrated that there are several potential paths to the development of AI BCS that would be acceptable to PCPs. Improvements in sensitivity have been the focus of many AI researchers and were highly valued by most respondents as well. We found, however, that translating AI BCS into the clinic does not need to wait for these technological improvements. Instead, refining policies around how radiologists work with AI, what data are used in the development of AI, and how studies can support its effectiveness claims all contribute meaningfully to PCPs' decisions around whether to recommend AI to their patients.

Respondents generally agreed on the value of sensitivity as the most important attribute of AI BCS. Developing AI BCS with sensitivity greater than that of radiologists was particularly

valued. This prioritization of sensitivity over specificity by PCPs is in contrast to public health researchers' focus on false positives as a major harm of screening. Another area of broad agreement among respondents was their preference for randomized controlled trials (RCTs) over observational studies. Using both types of studies to support the effectiveness claims of AI BCS was not preferred over RCT alone. This was despite the commonly-held belief among methodologists that clinical trials for diagnostic AI are impractical and among policy-makers that they may be unnecessary.[35]

We also identified meaningful heterogeneity in PCPs' preferences, which we explained by using a latent class model to group respondents into three classes. The area of greatest difference between these groups was in their attitude toward radiologist confirmation of images. Respondents had a 76% probability of belonging to a latent class that supported the use of AI to triage negative screens and to refer likely positives to radiologists. One class, however, defined itself largely by its opposition to any unattended use of AI. The members of this "Against AI Autonomy" class were significantly less likely to identify as having negative attitudes toward technology compared to members of the "Uncertain Trade-Offs" class. We interpret this as reflecting the "Against AI Autonomy" class's open-mindedness toward AI balanced with caution about its potential impacts. This class is the only one, in fact, to not be biased towards the status quo, as shown by the statistically non-significant value of the alternative specific constant of radiologist alone.

The classes also differed in their attitudes toward the potential trade-off between sensitivity and specificity. The "Sensitivity First" class found sensitivity to be at least twice as important as specificity, while the "Against AI Autonomy" class did not see them as significantly different in importance.

Our findings have several implications. First, we found that PCPs are largely supportive of using AI in a triage setting. This novel framework for human-AI collaboration would deploy AI at a low sensitivity / high specificity setting such that likely positives are sent to radiologists for review, while likely negatives receive an immediate determination.[36,37] A recent publication described an AI-based system used in this way that achieved a 99.9% negative predictive value while filtering out approximately 40% of mammograms that did not need radiologist review. Primary care providers' acceptance of using AI in a triage setting means that other workflows with clearer analogies to present-day operations are unnecessary. For example, some studies have used AI as a second reader for BCS images, where AI is treated as if it were a radiologist.[38] Other studies

have used radiologist input as if it were from an AI and integrated it into an ensemble with AI models.[27] The triage workflow instead uses the input of radiologists and AI to optimize for their unique strengths.

Our results also suggest that the PCPs who are most likely to accept the use of a triage system are those who are also most likely to view improved sensitivity as the main appeal of AI BCS. It is therefore necessary to study the overall sensitivity of a collaborative AI-radiologist team to ensure that it improves sensitivity over radiologist alone without negatively impacting specificity. Existing AI algorithms and radiologists detect somewhat different cases, suggesting that combining decisions from both could produce better performance than either one alone.[37] However, putting the triage workflow into practice would meaningfully change the decision task of radiologists by increasing the prevalence of positives as well as by providing a tacit "positive" label from the algorithm. Prior work testing radiologists' accuracy at mammogram interpretation for a non-representative sample of women suggests that training may be necessary to avoid poor performance at interpreting images triaged by AI.[39]

Our work also shows how researchers can create clinically acceptable technologies by refining other attributes if they are only capable of producing minor improvements in AI's sensitivity. In the absence of a technological breakthrough enabling sensitivity over 90%, our study provides evidence that improving other elements of AI BCS may make it clinically appealing for PCPs. Highly diverse training data and explanations of AI decision-making that are understandable by clinicians were both important to most respondents. Both of these attributes contribute to an algorithm's generalizability: diverse data by ensuring that it operates correctly on a wide range of patients, and understandable decision-making by showing when AI is considering inappropriate criteria for a judgment.[40] Improving the quality of clinical trials for AI BCS to include well-designed RCTs would also increase the technology's appeal to PCPs.[20]

Our analysis is novel, as relatively little work has been published on clinicians' preferences for AI. One survey of senior specialist physicians in the United Kingdom found liability, accuracy, understandability, and quality of supporting evidence to be primary concerns.[41] Qualitative work among PCPs in the United Kingdom found that, while many were concerned about AI's potential to disrupt the doctor-patient bond or to miss atypical presentations, participants were hopeful that AI could reduce inefficiencies by triaging uncomplicated patients and providing faster results.[42] Another study among physicians in New Zealand focused on understandability and found that 88%

of respondents were more likely to trust an AI algorithm that produced an understandable explanation of its decisions.[43] With the exception of liability, these are in general agreement with our results.

Our findings regarding the importance of understandability also deserve mention because of the importance that this topic has taken on among clinical commentators. Prior reports have suggested that the understandability of AI decision-making is vital for its application in clinics.[19,44] This has been supported by the survey mentioned above that showed how understandability correlates with acting on AI's recommendations.[43] While our findings bolster current knowledge about clinicians' preferences for explainable AI, they simultaneously undercut the idea of its paramount importance, at least among PCPs. We find, instead, that poor understandability can be offset by other attributes such as high sensitivity, high specificity, or excellent diversity in training data. Interpretable AI supports tests of generalizability that are essential in quality assurance and regulatory processes, however, and therefore should not be ignored.[40,45] Understandability is also connected to providers' liability for using AI.[18]

Our study has several limitations. The attitudes that informed our respondents' choices are likely to change as they encounter more examples of clinical AI. Our study can be used as a baseline to measure these changes, but this means that its results may have an apparent bias toward the negative. Its results also cannot be used to predict uptake, which would depend on the input of other stakeholders such as patients, payers, and radiologists. We were unable to evaluate all potentially relevant attributes, including cost. We attempted to determine the most important attributes, but may have missed attributes with similar importance. Changing the set of included attributes may have shifted the estimated impact of the attributes we included. Finally, the generalizability of our sample is unknown. Despite our efforts at recruiting a representative sample of PCPs, we relied on snowball sampling, which means that respondent selection was not entirely independent. It also means we may have underestimated the size of the confidence intervals.

While our study provides a useful start to quantitative translational research in AI for cancer screening, much work remains to be done. Future studies should examine the preferences of other stakeholders in the decision to implement AI BCS: patients, payers, and radiologists will all play important roles in deciding on the use of AI technologies. Work examining the performance of different configurations for the collaboration between AI and clinicians should be done and, with it, a more detailed examination of clinician's willingness to trade improvements in

AI's accuracy for increased autonomy in decision-making. Finally, we recommend that similar studies be conducted periodically to assess how exposure to clinical AI alters PCPs' willingness to recommend an expanded role for AI, particularly in cancer screening.

## 2.4 CONCLUSION

Much has been written about what AI needs before it can become clinically acceptable. In this study, we have quantified the impact of six highly relevant attributes of AI on PCPs' decision to recommend it to their patients for BCS. We find evidence that technical advances that allow for greater diagnostic accuracy are important, but are not the only way to produce appealing AI products. We also find support for using AI to interpret some mammograms without radiologist confirmation, which opens up the possibility of developing innovative human-machine collaborations that can reduce radiologist burden and improve the efficiency of BCS.

# Chapter 3. HEALTH OUTCOMES MODELING TO ASSESS ARTIFICIAL INTELLIGENCE DIAGNOSTICS: AN APPLICATION TO BREAST CANCER SCREENING

Advances in artificial intelligence (AI) such as convolutional neural networks have enabled its application to medical imaging diagnostics.[46] Assessing the performance of these algorithms is a necessary step as AI research is increasingly translated into commercial products. Discrimination defines an algorithm's ability to identify positive cases and is perhaps the most important performance metric for a diagnostic algorithm that produces a binary output.[35] Tests of discrimination inform comparisons between models as well as the selection of operating points (sometimes called "cut points"), which indicate the predicted probability of disease that optimally separates determinations of positives versus negatives.[47]

Discrimination ability is evaluated using the receiver operating characteristic (ROC) curve, which shows sensitivity as a function of specificity.[48,49] Researchers are able to treat the tasks of model comparison and operating point selection as mathematical properties of the ROC curve itself, thereby simplifying the evaluation of an algorithm's discriminatory ability and producing a metric that allows comparison between algorithms developed for different purposes. This approach has many limitations, however. Comparing models whose ROC curves cross is difficult.[50] It can also overemphasize very low and high specificities that would never be clinically useful.[51] Most relevant to the evaluation of AI, however, is the ROC curve's lack of connection to clinical impact.[52] While clinical trials ideally would provide estimates of AI diagnostics' effects on patient outcomes, this approach is difficult with AI algorithms due to rapid technological advances.[4]

Health outcomes modeling is an approach developed in decision analysis and used extensively in the health economics literature.[53,54] It is especially useful in making indirect comparisons, linking short term surrogate markers to longer term clinical outcomes, including life expectancy and quality of life in the absence of clinical trials, thus making it potentially applicable to AI-based diagnostics.[55] At its core, health outcomes modeling is a way of simulating patients' passages through a sequence of clinical and disease processes such as screening, treatment, and progression. The main process of this model is to track how much time patients spend in discrete

states of health. The summed time in each health state is then multiplied by a utility (quality of life) weight derived from preference studies. The result is expressed in quality-adjusted life years (QALYs) and accounts for changes in both survival and quality of life due to an intervention.

The objective of this study was to demonstrate the use and potential utility of health outcomes modeling by evaluating the discriminatory performance of AI algorithms developed to interpret mammograms for breast cancer screening. These algorithms were submitted to the Digital Mammography DREAM Challenge, a competition sponsored by a consortium of healthcare and technology firms to promote the development of artificial intelligence for breast cancer screening (AI BCS). We conclude with a discussion of the advantages and disadvantages of using health outcomes modeling as applied to AI diagnostic tools.

## 3.1 METHODS

### 3.1.1 *Overall Approach*

We developed a health outcomes model for BCS and estimated the QALYs generated by each AI algorithm across the possible range of operating points. We used the maximum QALYs generated by each algorithm to rank them by performance and used the criterion of QALY-maximization to select an operating point. We then compared our findings to the rankings and operating points derived by a series of ROC-based methods.

### 3.1.2 *Data*

The Digital Mammography DREAM Challenge was a competition that awarded prizes for the development of AI BCS from the public.[27] It was operated by a consortium of healthcare and technology firms between September 2016 and November 2017. Teams and individuals who entered had access to mammograms, cancer diagnoses, and limited patient demographics from Kaiser Permanente Washington and the Karolinska Institute that they could use to develop their algorithms. Competitors from around the world submitted a total of 28 algorithms, which were evaluated by area under the ROC curve (AUC) and specificity when sensitivity is fixed at 86%. The developers of the top eight algorithms subsequently entered a cooperative phase where they created ensembles of their algorithms with and without the inclusion of radiologist readings as part of the ensemble.

We used a dataset with 51,314 mammograms. Each mammogram had a predicted probability of cancer from the algorithms, and an indicator of true cancer status. This indicator was derived from a cross-linkage with Surveillance, Epidemiology, and End Results (SEER) Program data that detected women in the test set who were diagnosed with breast cancer within twelve months of the included mammogram. Linking to SEER Program data allowed for cancer determinations that were closer to the gold standard than radiologist interpretations would have been. We assigned a letter name to each AI algorithm for ease of reference.

### 3.1.3 *Health Outcomes Model*

We began with an open-source health outcomes model of breast cancer screening in the United States and adapted it to generate QALYs. The model simulated the outcomes for women who attend screening in comparison to those who do not.[56,57] Screening benefited women in this model by allowing for the discovery of cancers in an earlier and more treatable stage, relative to when they would be discovered symptomatically, thus improving quality of life by requiring less invasive treatment and improving survival.[58] The probability of undergoing this stage shift was a function of the frequency and sensitivity of screening, and included an adjustment for overdiagnosis.[59] The availability of this open-source model assured that we had a validated source of breast cancer incidence and survival estimates by stage. If we had not obtained this open-source model, we would have needed to create our own and to validate its outputs by comparing them to published figures to ensure proper calibration.[60,61]

The open-source model did not include discrete health states, but instead produced the ages and disease stages at which women with breast cancer would be diagnosed; ages of death; and indicators of whether women's deaths were breast cancer-related, each in the presence and absence of screening. We added a number of health states to the health outcomes model in order to generate QALYs.

See **Figure 3.1** for a schematic of our adapted model. Every two years between the ages of 50 and 74, these women attended screening. Women with negative screens returned to the "well" health state, while women with positive screens underwent further workup, generally consisting of further imaging and possibly biopsy. If the workup resulted in determining that the woman did not have cancer, then she returns to "well"; if there was a positive determination of cancer, then

she proceeded to cancer treatment. Having a positive or negative screen was correlated with true cancer status through the sensitivity and specificity of screening.

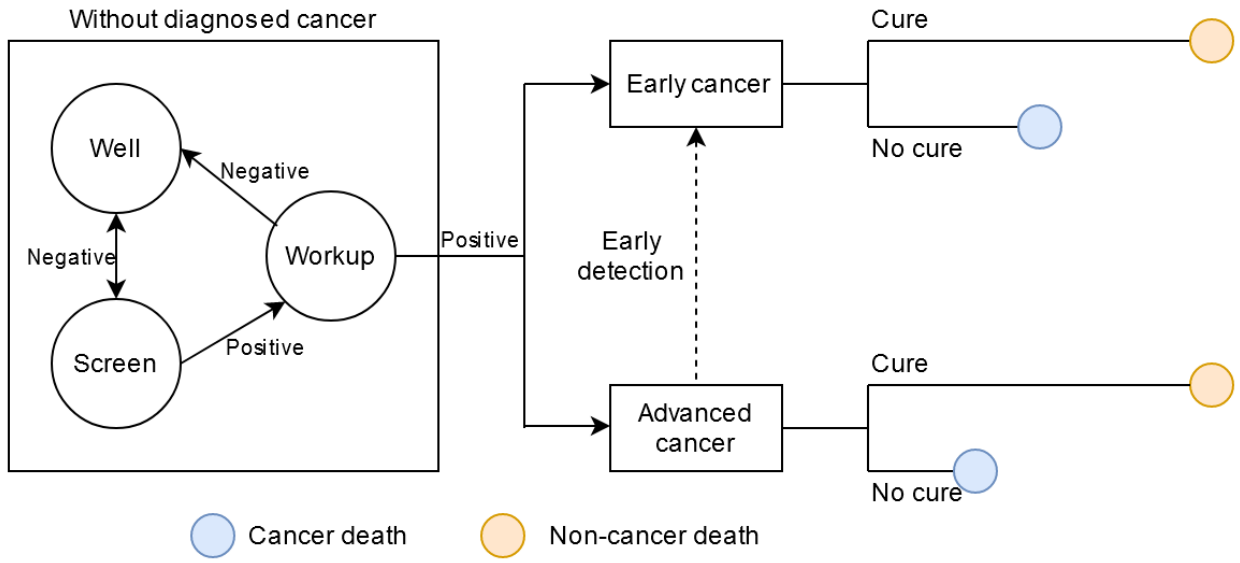


Figure 3.1. Schematic of breast cancer screening and outcomes model.

We next accounted for cancer-related health states, which we simplified to two: early and advanced cancer. Based on the clinical course of treatment, we assumed treatment for early and advanced cancer would both take one year, after which women who survived would go into remission. Women entered a "progressive cancer" state eighteen months prior to breast cancer-related deaths. Women who were cured, on the other hand, lived ten years in the remission state and then transitioned to the cured state.

We assigned utility or quality of life weights (on a scale of 0 to 1, where 0 represents death and 1 represents perfect health) to the health states based on published utility elicitation studies. Starting with the "well" health state, we identified a source that provided age-adjusted utility weights for individuals over 50. We then converted other utility weights into disutilities to be subtracted from the utility weight of that well states – that is, we took the difference between the utility weight of that state and perfect health and subtracted it instead from the age-adjusted "well" state. See **Table 3.1** for a list of health states, their durations, and the utility weight associated with them.

Table 3.1. Utility weights associated with breast cancer screening and breast cancer-related outcomes.

| State | Value | Standard deviation | Duration | Source |
|---|---|---|---|---|
| Well (≥ 50 years) | 0.85 | 0.01 | Until cancer diagnosis or death | Shaw, 2005[56] |
| Screening | -0.07 | 0.02 | Three weeks | Bonomi, 2008[57] |
| Workup of positive screen | -0.14 | 0.02 | Two weeks | Bonomi, 2008[57] |
| Early cancer | -0.14 | 0.04 | One year or until progression | Pataky, 2013[58] |
| Advanced cancer | -0.325 | 0.05 | One year or until progression | Pataky, 2013[58] |
| Cancer remission | -0.035 | 0.05 | Nine years for patients with non-cancer death; until progression for patients with cancer-related death. | Pataky, 2013[58] |
| Progressive disease | -0.62 | 0.05 | Up to 18 months before death. | Schleinitz, 2006[59] |
| Cured | -0.00 | 0 | Until other cause death | Assumption |

We assessed uncertainty in the simulated sample and around the cancer-related parameters, including utility weights.[62,63] To account for the first, we gradually increased the sample size for each model simulation to 30,000, at which point we found that the average time spent in each state stopped changing. We next repeated samples of 30,000 simulated women until variation stabilized at 10,000 runs in order to account for parameter uncertainty.

We therefore repeated 10,000 model runs for each potential combination of sensitivity and specificity, with each consisting of 30,000 women starting when screening begins at age 50 and ending when all women in the sample have died. Each future year was discounted by 3% to account for the lower value of the future relative to the present.[64,65] Repeatedly calculating QALYs for each combination of sensitivity and specificity allowed us to examine how outcomes change with improvements in sensitivity and specificity. All measures of QALY gain were relative to no screening, so as to provide a fixed point of reference. We validated the resulting outcomes against published estimates created by the CISNET consortium, whose members have created six different health outcomes models for BCS.[66]

We also evaluated parameter uncertainty around the sensitivity at each specificity value in the algorithms as well. We used a normally distribution within the 95% confidence interval of each algorithm's ROC curve to produce 1,000 randomly drawn sensitivity values across the range of specificity values from 0 to 100% for each algorithm. This process was facilitated by the "pROC" package for R.[67] We matched each of these specificity-sensitivity pairs representing algorithm performance to a random outcome generated by the health outcomes model for that combination of sensitivity and specificity.

We ranked the 28 algorithms by the QALYs generated at the operating point with the highest mean number of QALYs - we also used this operating point for comparison with ROC-based methods of operating point selection.

We evaluated the health outcomes model's sensitivity to changes in the acceptable cost-benefit ratio by tripling and then sextupling the disutility of workup for positive screens. This allowed us to explore potential explanations for our findings by penalizing false positives more than in the base case. This increases the relative importance of specificity and may therefore change our performance evaluations. Comparing the differences in the rankings between the base case and these secondary analyses will also help us to understand the generalizability of our findings to other disease areas.

### 3.1.4 *ROC-Based Analysis*

We used two ROC-based methods for ranking algorithms by performance and three methods for selecting operating points. We first produced an ROC curve for each algorithm to replicate the standard measure of performance. Each algorithm's ROC curve included the 95% confidence interval to reflect uncertainty in performance based on sample size. We then calculated the AUC, which we used to produce an initial ranking of the algorithms by performance. The algorithms submitted to the Digital Mammography DREAM Challenge were originally judged by specificity when sensitivity was 86% (i.e., equal to the average radiologist). The ranking produced by this method is also included and compared to the health outcomes model.

We used three different methods for selecting an operating point without using a health outcomes model: Youden's index, decision curve analysis, and a fixed operating point of 86% sensitivity. Youden's index uses the point on the ROC curve that is closest to perfect performance (i.e., 100% sensitivity and 100% specificity – the upper left corner of the ROC curve) as the

operating point. This method is based, however, on the assumption that false positives and false negatives are equally undesirable, which is not true for breast cancer screening.[5] We thus used decision curve analysis, which requires establishing an acceptable ratio of false positives to false negatives based on the relative harm caused by each.[6] This ratio can be estimated by literature review, stakeholder surveys, or analysis of current practice. We estimated that a ratio of 100 false positives to 1 false negative would be acceptable based on current breast cancer screening practices. Specifically, given a breast cancer prevalence of 0.6% in women of screening age, six women per thousand have breast cancer.[68] The average radiologist interprets mammograms with 85.9% sensitivity and 90.5% specificity, suggesting that one of those women with breast cancer would receive a false negative screen while 109 without breast cancer would receive a false positive.[27] We rounded this estimate to our final ratio of 100 to 1, which we use in the decision curve analysis.

Finally, we compared operating points selected by maximizing QALYs in the health outcomes modeling to the operating point which corresponds to 86% sensitivity.

## 3.2 RESULTS

### 3.2.1 *Model validation*

The CISNET models estimated between 470 and 1,150 QALYs would be gained per 10,000 women screened biennially by radiologist alone between the ages of 50 and 74.[69] Our model predicted a gain of 960 QALYs in this scenario, placing it slightly above the average value of 860 QALYs found in the six CISNET models.

### 3.2.2 *Health Outcomes Model Results*

Improving sensitivity by one percentage point improved outcomes by between 8.04 and 24.86 QALYs per 10,000 women screened. Improvements in sensitivity had the highest associated QALY gain when baseline sensitivity was low. Improving specificity by one percentage point was associated with a consistent gain of 4.99 QALYs per 10,000 women, regardless of baseline performance. We illustrate the relationship between sensitivity and stage shift, as well as the change in outcomes caused by improvements in specificity and sensitivity in **Figure 3.2**. Absolute outcomes by sensitivity and specificity are available in **Table 3.2**.

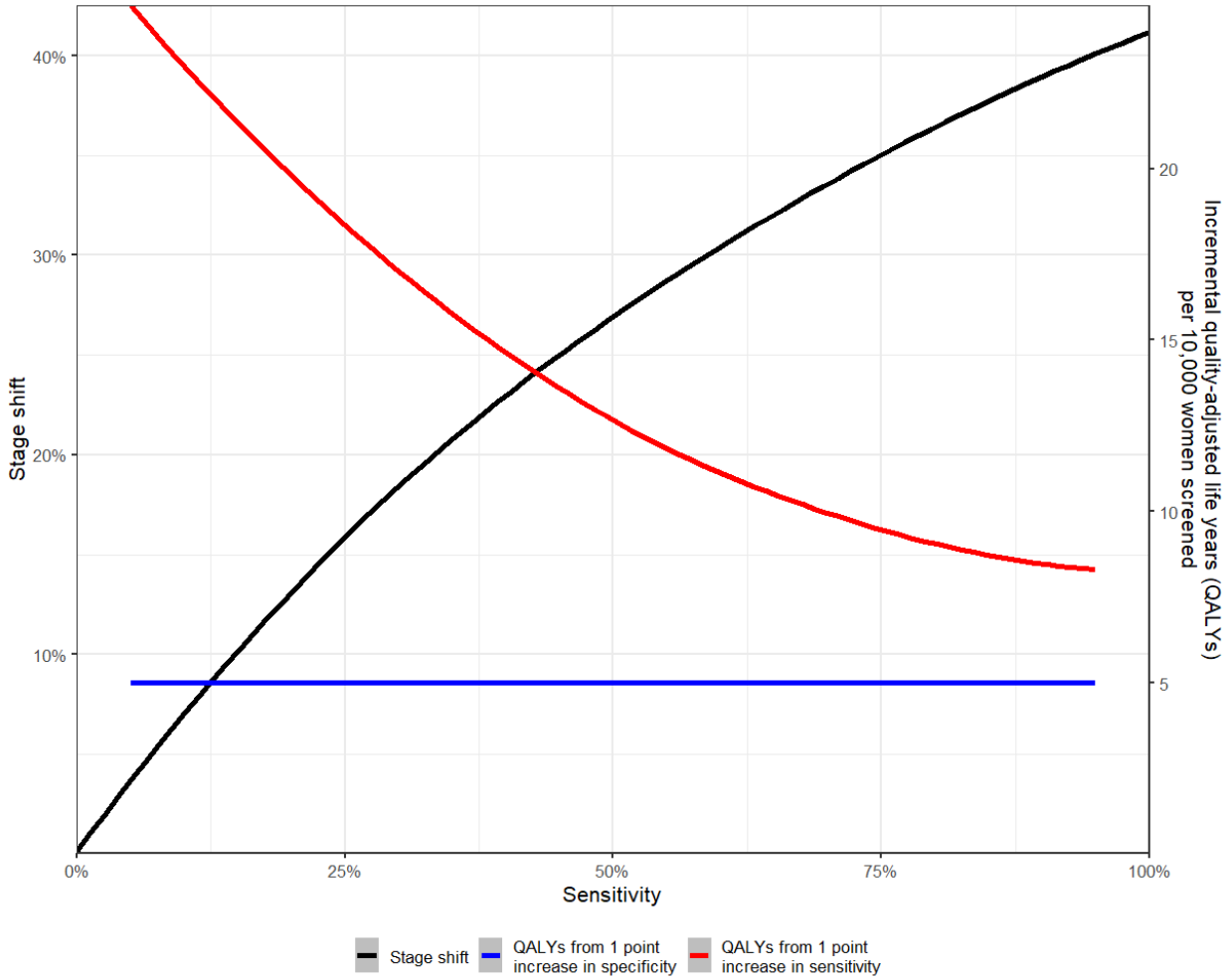


Figure 3.2. Incremental per capita quality-adjusted life years (QALYs) of screening with variable sensitivity / specificity, as compared to no screening.

Table 3.2. Per capita quality-adjusted life years (QALYs) of screening with variable sensitivity and specificity, as compared to no screening.

| Specificity | 0.05 | 0.1 | 0.15 | 0.2 | 0.25 | 0.3 | 0.35 | 0.4 | 0.45 | 0.5 | 0.55 | 0.6 | 0.65 | 0.7 | 0.75 | 0.8 | 0.85 | 0.9 | 0.95 | 1 | Sensitivity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.037 | 0.071 | 0.103 | 0.132 | 0.159 | 0.184 | 0.208 | 0.23 | 0.25 | 0.269 | 0.287 | 0.304 | 0.32 | 0.336 | 0.35 | 0.364 | 0.377 | 0.389 | 0.401 | 0.412 | Stage shift |
| 0 | -0.074 | -0.062 | -0.050 | -0.039 | -0.029 | -0.020 | -0.012 | -0.003 | 0.004 | 0.011 | 0.017 | 0.024 | 0.029 | 0.035 | 0.040 | 0.045 | 0.050 | 0.055 | 0.059 | 0.063 | |
| 0.05 | -0.072 | -0.059 | -0.047 | -0.037 | -0.027 | -0.018 | -0.009 | -0.001 | 0.006 | 0.013 | 0.020 | 0.026 | 0.032 | 0.038 | 0.043 | 0.048 | 0.053 | 0.057 | 0.062 | 0.066 | |
| 0.1 | -0.069 | -0.057 | -0.045 | -0.034 | -0.024 | -0.015 | -0.007 | 0.001 | 0.009 | 0.016 | 0.022 | 0.029 | 0.034 | 0.040 | 0.045 | 0.050 | 0.055 | 0.060 | 0.064 | 0.068 | |
| 0.15 | -0.067 | -0.054 | -0.042 | -0.032 | -0.022 | -0.013 | -0.004 | 0.004 | 0.011 | 0.018 | 0.025 | 0.031 | 0.037 | 0.043 | 0.048 | 0.053 | 0.058 | 0.062 | 0.066 | 0.071 | |
| 0.2 | -0.064 | -0.052 | -0.040 | -0.029 | -0.019 | -0.010 | -0.002 | 0.006 | 0.014 | 0.021 | 0.027 | 0.034 | 0.039 | 0.045 | 0.050 | 0.055 | 0.060 | 0.065 | 0.069 | 0.073 | |
| 0.25 | -0.062 | -0.049 | -0.037 | -0.027 | -0.017 | -0.008 | 0.001 | 0.009 | 0.016 | 0.023 | 0.030 | 0.036 | 0.042 | 0.048 | 0.053 | 0.058 | 0.063 | 0.067 | 0.071 | 0.076 | |
| 0.3 | -0.059 | -0.047 | -0.035 | -0.024 | -0.014 | -0.005 | 0.003 | 0.011 | 0.019 | 0.026 | 0.032 | 0.039 | 0.044 | 0.050 | 0.055 | 0.060 | 0.065 | 0.070 | 0.074 | 0.078 | |
| 0.35 | -0.057 | -0.044 | -0.032 | -0.022 | -0.012 | -0.003 | 0.006 | 0.014 | 0.021 | 0.028 | 0.035 | 0.041 | 0.047 | 0.053 | 0.058 | 0.063 | 0.068 | 0.072 | 0.076 | 0.081 | |
| 0.4 | -0.054 | -0.042 | -0.030 | -0.019 | -0.009 | 0.000 | 0.008 | 0.016 | 0.024 | 0.031 | 0.037 | 0.044 | 0.049 | 0.055 | 0.060 | 0.065 | 0.070 | 0.075 | 0.079 | 0.083 | |
| 0.45 | -0.052 | -0.039 | -0.027 | -0.017 | -0.007 | 0.002 | 0.011 | 0.019 | 0.026 | 0.033 | 0.040 | 0.046 | 0.052 | 0.058 | 0.063 | 0.068 | 0.073 | 0.077 | 0.081 | 0.085 | |
| 0.5 | -0.049 | -0.037 | -0.025 | -0.014 | -0.004 | 0.005 | 0.013 | 0.021 | 0.029 | 0.036 | 0.042 | 0.049 | 0.054 | 0.060 | 0.065 | 0.070 | 0.075 | 0.080 | 0.084 | 0.088 | |
| 0.55 | -0.047 | -0.034 | -0.022 | -0.012 | -0.002 | 0.007 | 0.016 | 0.024 | 0.031 | 0.038 | 0.045 | 0.051 | 0.057 | 0.063 | 0.068 | 0.073 | 0.078 | 0.082 | 0.086 | 0.090 | |
| 0.6 | -0.044 | -0.032 | -0.020 | -0.009 | 0.001 | 0.010 | 0.018 | 0.026 | 0.034 | 0.041 | 0.047 | 0.054 | 0.059 | 0.065 | 0.070 | 0.075 | 0.080 | 0.085 | 0.089 | 0.093 | |
| 0.65 | -0.042 | -0.029 | -0.017 | -0.007 | 0.003 | 0.012 | 0.021 | 0.029 | 0.036 | 0.043 | 0.050 | 0.056 | 0.062 | 0.068 | 0.073 | 0.078 | 0.083 | 0.087 | 0.091 | 0.095 | |
| 0.7 | -0.039 | -0.027 | -0.015 | -0.004 | 0.005 | 0.015 | 0.023 | 0.031 | 0.039 | 0.046 | 0.052 | 0.058 | 0.064 | 0.070 | 0.075 | 0.080 | 0.085 | 0.090 | 0.094 | 0.098 | |
| 0.75 | -0.037 | -0.024 | -0.012 | -0.002 | 0.008 | 0.017 | 0.026 | 0.034 | 0.041 | 0.048 | 0.055 | 0.061 | 0.067 | 0.073 | 0.078 | 0.083 | 0.088 | 0.092 | 0.096 | 0.100 | |
| 0.8 | -0.034 | -0.022 | -0.010 | 0.001 | 0.010 | 0.020 | 0.028 | 0.036 | 0.044 | 0.051 | 0.057 | 0.063 | 0.069 | 0.075 | 0.080 | 0.085 | 0.090 | 0.095 | 0.099 | 0.103 | |
| 0.85 | -0.032 | -0.019 | -0.007 | 0.003 | 0.013 | 0.022 | 0.031 | 0.039 | 0.046 | 0.053 | 0.060 | 0.066 | 0.072 | 0.078 | 0.083 | 0.088 | 0.093 | 0.097 | 0.101 | 0.105 | |
| 0.9 | -0.029 | -0.017 | -0.005 | 0.006 | 0.015 | 0.025 | 0.033 | 0.041 | 0.049 | 0.056 | 0.062 | 0.068 | 0.074 | 0.080 | 0.085 | 0.090 | 0.095 | 0.100 | 0.104 | 0.108 | |
| 0.95 | -0.027 | -0.014 | -0.002 | 0.008 | 0.018 | 0.027 | 0.036 | 0.044 | 0.051 | 0.058 | 0.065 | 0.071 | 0.077 | 0.083 | 0.088 | 0.093 | 0.098 | 0.102 | 0.106 | 0.110 | |
| 1 | -0.024 | -0.012 | 0.000 | 0.011 | 0.020 | 0.030 | 0.038 | 0.046 | 0.054 | 0.061 | 0.067 | 0.073 | 0.079 | 0.085 | 0.090 | 0.095 | 0.100 | 0.105 | 0.109 | 0.113 | |

### 3.2.3 *Comparison of Algorithm Performance*

The algorithms generated a maximum of between 540 and 860 additional QALYs per 10,000 women screened vs. no screening. The AUCs ranged from 0.496 to 0.874, while specificity when sensitivity was fixed at 86% ranged from 9% to 49%. Rankings of the 28 algorithms were effectively identical using health outcomes modeling, AUC, or specificity at 86% sensitivity (**Figure 3.3**).

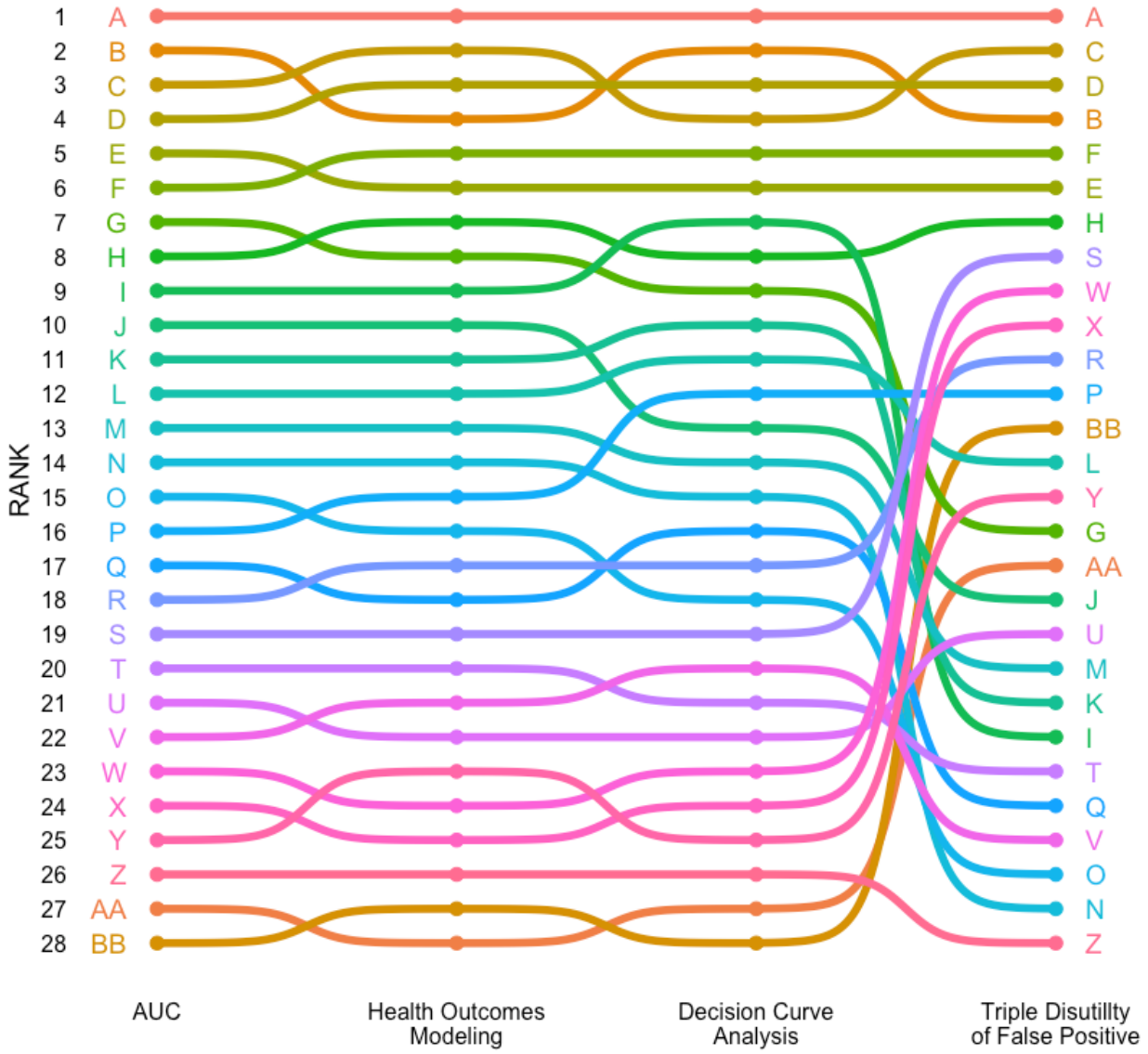


Figure 3.3. Change in model rankings by performance metric.

Our secondary analyses in which we modified the disutility of false positives resulted in a substantially different ranking of the algorithms. The algorithms that the other methods found to be in the top six remained in the top six positions, but lower-ranking algorithms were almost entirely reordered.

### 3.2.4 *Operating Point Selection*

The QALY-maximizing operating point generated by the health outcomes model and the empirically selected operating point of 86% used in judging the Digital Mammography DREAM Challenge produced the most similar operating points (**Table 3.3**). These operating points were, on average, 28 percentage points apart in specificity, but generated a statistically non-significant ($p < 0.05$) difference in the number of QALYs every algorithm produced.

Table 3.3. Mean (range) of differences between operating point selection methods and maximum quality-adjusted life years (QALYs) gained per capita versus no screening.

| | **Youden's index** | **Decision curve analysis** | **Fixed operating point at 86% sensitivity** |
|---|---|---|---|
| Difference in operating point (percentage points on specificity scale) | 26 (2 to 72) | 54 (11 to 92) | 28 (0 to 38) |
| Difference in QALYs produced at operating point | 0.010 (0.000 to 0.096) | 0.053 (0.004 to 0.097) | 0.007 (0.000 to 0.009) |
| Number of models with significant difference in QALYs | 8 | 24 | 0 |

Youden's index performed next best. On average, the operating point was 26 percentage points away from the QALY-maximizing operating point but, for four of the algorithms, it generated significantly fewer QALYs. Both Youden's index and the fixed operating point discussed above tended to select operating points with lower specificity and higher sensitivity than the QALY-maximizing operating point (**Figure 3.4**).

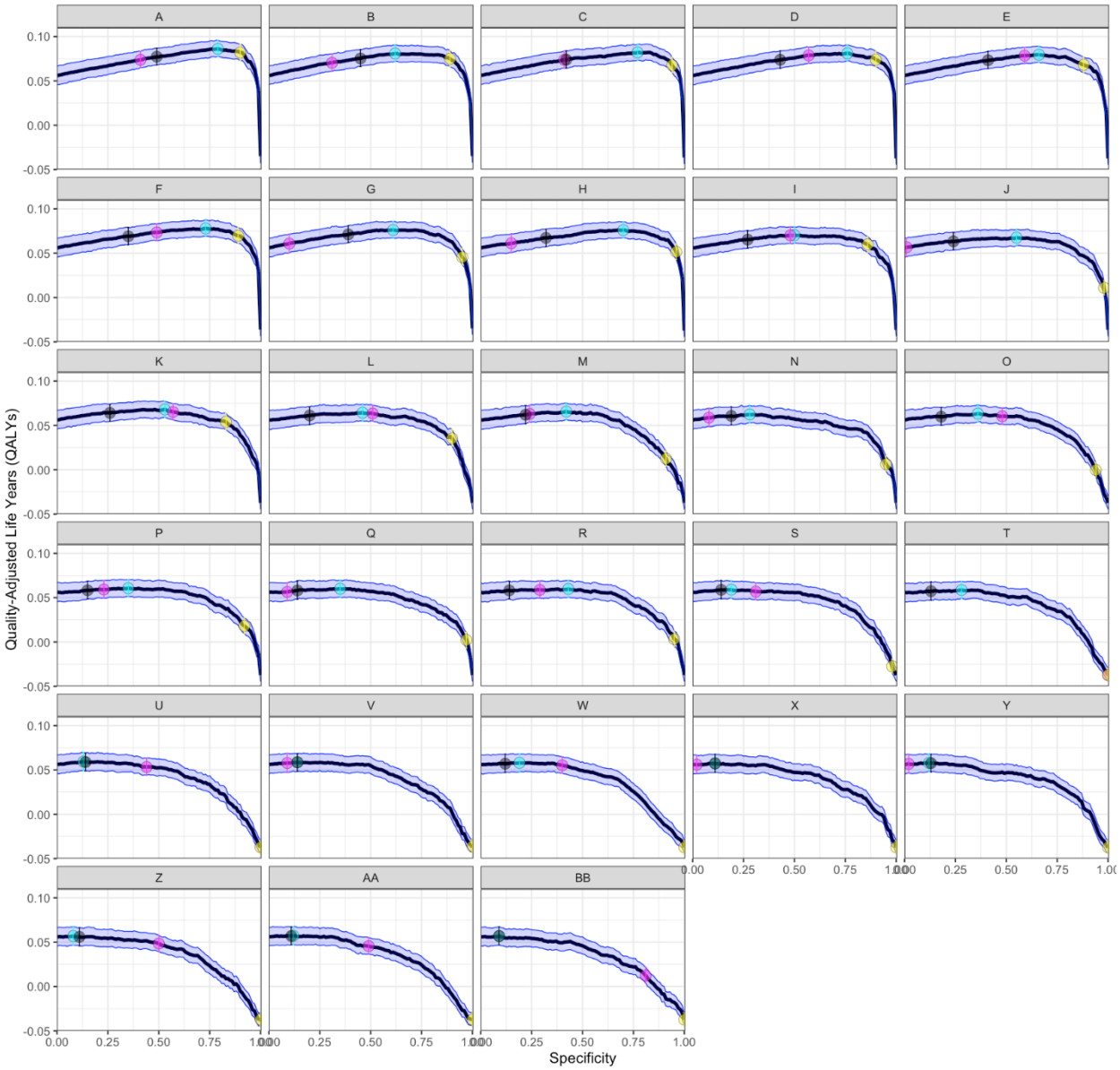


Figure 3.4. Comparison of operating points selected for the 28 models by using Youden's index (pink), decision curve analysis (yellow), fixed sensitivity of 86% (black), and maximization of quality-adjusted life years (blue).

Decision curve analysis was farthest from the QALY-maximizing operating point. On average, these two operating points were 54 percentage points apart in sensitivity, with QALYs produced by the two operating points having a statistically significant difference for 24 out of 28 algorithms. Unlike the ROC-based methods and decision curve analysis, it generally produced

operating points with higher specificity and lower sensitivity than the QALY-maximizing operating point.

## 3.3 DISCUSSION

We developed a health outcomes model based on an open-source microsimulation of breast cancer outcomes. We used the health outcomes model to compare the performance of AI BCS algorithms and to select operating points in relation to the area under the ROC curve (AUC), Youden's index, decision curve analysis, and the operating point corresponding to a fixed sensitivity of 86%.

We found that comparing rankings of AI algorithms by QALY-maximization in the health outcomes model, AUC, specificity at 86% sensitivity, and produced similar rankings of the 28 models. We hypothesize the similarity in rankings is due to the fact that improving sensitivity and specificity led to similar QALY gains. Indeed, the QALYs produced by these improvements became more similar as sensitivity approached 100%.

Our observation of the diminishing returns on improvements to sensitivity likely stems from the non-linear relationship between sensitivity and stage-shift. Improving sensitivity from a low baseline caused a relatively great increase in stage-shift. The benefit of screening, however, was capped by the screening interval: biennial screening with perfect discrimination leads to only 41% of advanced cancers being shifted to an earlier stage. This limitation on the stage-shift of screening is due to the distribution of times between when a cancer becomes detectable by screening and when it would become symptomatic - the lead time.[70] We ran a post-hoc analysis of the health outcomes model to test our proposed explanation, in which the disutility of false positives was inflated to three and then six times its value in the base case. These cases showed little variation among top performing algorithms, which generally were less affected by changes to the utility of false positives due to their high specificity. Lower-performing models were substantially reranked, however. The ranking by performance that we observed suggests the potential for diseases with different risk-benefit ratios to generate more differences in ROC- versus health outcomes-based methods. The ability to test this explanation by altering a single part of the health outcomes model points to one of its potential advantages: namely, that it can be used as a tool to explore what is unique about different diseases and how one could reasonably expect ROC-based tools to perform in response.

### 3.3.1 *Operating Point Selection*

We found substantial differences in the operating points that would be selected using the different evaluation methods. The operating point corresponding to a fixed sensitivity of 86% produced a statistically non-significant difference from the QALY-maximizing operating point for every model. Operating points selected by Youden's index showed statistically significant differences from the QALY-maximizing operating point in fewer cases than decision curve analysis.

These findings suggest that comparing these algorithms on the basis of performance at an equivalent sensitivity to the average radiologist – as was done in the Digital Mammography DREAM Challenge - corresponds well to their likely effect on patient outcomes. The usefulness of an operating point based on clinician performance for comparing models in other diseases is not clear. Rather, this finding makes the case for using a health outcomes model as a confirmatory check for one's decision to use a certain method of operating point selection, given the unique balance of risks and benefits for that disease.

Patient concerns should be central to contemporary research into clinical AI, which has been criticized as being focused only on accuracy metrics without considering measured impact on domains that foreground stakeholder concerns.[71] An example of this focus on accuracy metrics was seen in a recent study on AI BCS that selected its operating point on the basis of AUC. [37] Integrating health outcomes modeling into the process of operating point selection would ensure that patient preferences are represented in how the model is applied to clinical situations.

### 3.3.2 *Challenges of AI for Health Outcomes Modeling*

Artificial intelligence algorithms present unique challenges for health outcomes modeling. The need to explore outcomes generated by a range of the ROC curve is computationally expensive. Evaluating these 28 algorithms required us to run our health outcomes model approximately 420 million times, for example. Another complication is assessing the complementarity of AI and clinician judgment.[72] That assessment would mean not only determining the overlap between the cases detected by AI and by clinicians – an AI that finds the same cases as a clinician would be potentially less valuable – but also estimating the impacts of variable configurations of the AI-clinician collaboration.[73,74] Modeling how AI and clinicians work

together should also consider the behavioral response to AI, which is currently understudied. When used in subgroup analysis, health outcomes modeling also highlights the clinical impact that arises from spectrum bias (i.e., differential performance among subgroups with different risk levels) more directly than ROC-based methods can.[75] Including subgroup analysis, however, would require potentially complex causal inference to estimate counterfactual outcomes associated with early detection or delayed diagnosis.[76,77] Finally, many potential uses of AI do not produce binary diagnoses. These applications, such as image segmentation, pose a challenge to health outcomes modeling, because they have a less direct connection to the potential clinical response and are therefore harder to model.[78,79]

### 3.3.3 *Limitations*

Our work has several limitations. First, we did not consider all methods of assessing performance or selecting the operating point. We did not, for example, include Pauker and Kassirer's treatment threshold approach, nor Pepe's predictiveness curve.[50,80,81] These approaches informed the development of decision curve analysis, but the differences in their approaches may have led to different results. We also excluded several methods that integrate costs, such as Phelps and Mushlin's merging of ROC analysis with medical decision theory or Laking and Lord's non-linear expansion path approach.[82,83] We did not include these approaches because they would have substantially broadened this study's scope.

We also note that the QALY metric is controversial.[84] Perhaps the most common critique is that small benefits accruing to a large number of people are not distinguished from a smaller number receiving larger benefits, which may produce decisions that are unfair to society's more vulnerable members. Approaches to addressing this critique vary, but may include requiring that QALY gains pass some minimal threshold before being counted or modifying how QALYs are summed across groups to account for distributional equity.[85,86] There are also several alternatives to the QALY that address its shortcomings.[87] For example, Basu's "Health Years in Total (HYT)" concept separately aggregates gains to quality of life and life expectancy, meaning that individuals with low baseline quality of life can still benefit from longer life expectancy.[88] Had we attempted to include a modification of or alternative to the QALY, it may have given somewhat different results.

*Future Work*

Substantial work remains to be done on the use of health outcomes modeling for clinical AI. First, health outcomes modeling should be used in algorithms that display spectrum bias, which occurs when algorithm performance is unequal between different subgroups and is not accounted for in the interpretation of the model's output.[89,90] Subgroup analysis is frequently used in health outcomes modeling and could be a tool for assessing the impact of spectrum bias. We hypothesize that an assessment of spectrum bias in AI would likely show the advantages of health outcomes modeling over ROC-based methods in starker contrast than we have been able to. While ROC-based methods can be used to assess differences in algorithm performance among different subgroups, it does not make the clinical implications of these differences apparent. This work, however, is predicated on the availability of detailed patient data, which privacy or intellectual property concerns may make difficult to share.[91,92]

More insight on how to use health outcomes modeling for AI will also be gained by performing comparative studies such as this one in disease areas with different trade-offs versus BCS. Our study found substantial similarity between health outcomes modeling and ROC-based methods that assume that false positives and false negatives are equally undesirable, which matched with our finding that improving sensitivity and specificity produced similar benefits. Other use cases for clinical AI would likely have substantially different ratios of benefits. Algorithms have been developed to detect depression and other mental health problems from a patient's speech and movement patterns.[93] A false positive for this application would mean needlessly contacting the user, potentially via a chatbot; a missed opportunity to diagnose a burgeoning mental health crisis, though, could have much greater costs.

A major disadvantage of health outcomes modeling is that it is more labor-intensive than ROC-based methods. Little work has been done to measure the cost of performing health outcomes modeling relative to simpler methods, however. This information will be vital in order to rationally weigh the trade-offs between the costs and benefits of using health outcomes modeling to acquire greater precision around how to use clinical AI.

## 3.4 Conclusion

A valuable step in translating AI into clinical settings likely will be to evaluate impacts on patient morbidity, mortality, and quality of life, in addition to diagnostic accuracy. Health outcomes modeling is an excellent tool to aid in this transition. In this study, we demonstrated how

health outcomes models can be built and applied to the evaluation of AI algorithms as a supplement to ROC-based assessments. Our findings showed the main advantages of health outcomes modeling over ROC-based methods: it allows users to explore potential explanations for their findings and better integrates patient preferences. At the same time, evaluating AI with health outcomes modeling poses unique challenges. Bridging the disciplinary divides that separate AI researchers and health outcomes modelers can meaningfully advance both fields, while bringing the potential advantages of AI to health systems.

# Chapter 4. CONCLUSION

I aimed to use this dissertation as a way of exploring the role that health economics can play in bringing AI to clinics and to patients. This meant performing one experiment to assess the predictors of primary care provider acceptance and another to evaluate outcomes modeling as a tool to optimize AI for patient benefit.

In highly publicized statements, AI researchers have suggested that it may be relatively easy for AI to replace radiologists.[73] Our findings, however, indicate that they will play an essential role in bringing AI to clinical settings, as most PCPs are reluctant to accept AI without some radiologist confirmation of some images. Our respondents' acceptance of "human in the loop" AI for breast cancer screening opens up new possibilities for exploring novel workflows that incorporate the unique benefits of both human and machine. Much work remains to be done on providers' behavioral responses to the introduction of AI and on assessing the combined performance of human-AI collaborative work.

The DCE we performed establishes a useful approach to soliciting other stakeholders' preferences for AI. We found in this work that PCPs were engaged in the research and came to it with ideas about AI that we were able to integrate. These ideas were later validated as important attributes in our analysis of the responses. Similar future work could be meaningfully done with patients, payers, and radiologists. Other future work could explore how generalizable our findings about breast cancer screening would be to other potential applications of AI.

Our second experiment, likewise, opens the possibility of further research in the application of health economic methods to AI. We applied outcomes modeling to the tasks of model comparison and operating point selection for machine learning models, and found that it performs as well as AUC and decision curve analysis methods. This validates the use of outcomes modeling. Further work should explore cases where outcomes modeling may be superior as in, for example, when disease outcomes are heterogeneous and some AI models may have lower detection rates in high-risk subgroups. The data and analysis costs of such a project would be high. This means that developing methods for measuring the costs of applying outcomes modeling should be developed in tandem, which will allow researchers to decide when heuristic methods represent a better value than the more labor- and data-intensive methods of health outcomes modeling.

It is no longer adequate for AI and healthcare research to proceed along separate research tracks. To tackle translational issues, AI needs researchers with domain knowledge of the healthcare system. Likewise, healthcare research depends on AI to broaden the accessibility of high-quality care and to reduce costs. The applications of health economics tools to AI is an attempt at building a bridge between the two fields, which we hope will be bolstered by more future work.

# BIBLIOGRAPHY


1. Beam AL, Kohane IS. Translating Artificial Intelligence Into Clinical Care. *JAMA*. 2016;316(22):2368. doi:10.1001/jama.2016.17217

2. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? *Am J Med*. 2018;131(2):129-133. doi:10.1016/j.amjmed.2017.10.035

3. Jha S, Cook T. Artificial Intelligence in Radiology—The State of the Future. *Acad Radiol*. 2020;27(1):1-2. doi:10.1016/j.acra.2019.11.003

4. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25(1):30-36. doi:10.1038/s41591-018-0307-0

5. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biom J J Math Methods Biosci*. 2005;47(4):458-472.

6. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making*. 2006;26(6):565-574. doi:10.1177/0272989X06295361

7. Myers ER, Moorman P, Gierisch JM, et al. Benefits and harms of breast cancer screening: a systematic review. *Jama*. 2015;314(15):1615-1634.

8. Hubbard RA, Kerlikowske K, Flowers CI, Yankaskas BC, Zhu W, Miglioretti DL. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study. *Ann Intern Med*. 2011;155(8):481-492.

9. Ong M-S, Mandl KD. National expenditure for false-positive mammograms and breast cancer overdiagnoses estimated at $4 billion a year. *Health Aff (Millwood)*. 2015;34(4):576-583.

10. Nelson HD, Pappas M, Cantor A, Griffin J, Daeges M, Humphrey L. Harms of breast cancer screening: systematic review to update the 2009 US Preventive Services Task Force recommendation. *Ann Intern Med*. 2016;164(4):256-267.

11. Houssami N, Lee CI, Buist DS, Tao D. Artificial intelligence for breast cancer screening: Opportunity or hype? *The Breast*. 2017;36:31-33.

12. Trister AD, Buist DS, Lee CI. Will machine learning tip the balance in breast cancer screening? *JAMA Oncol*. 2017;3(11):1463-1464.

13. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*. 2007;356(14):1399-1409.

14. Dromain C, Boyer B, Ferre R, Canale S, Delaloge S, Balleyguier C. Computed-aided diagnosis (CAD) in the detection of breast cancer. *Eur J Radiol*. 2013;82(3):417-423.

15. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med*. 2015;175(11):1828-1837.

16. Business Wire. FDA-cleared artificial intelligence breast cancer diagnosis system launched by Paragon Biosciences. Published September 12, 2019. Accessed May 25, 2020. https://www.mpo-mag.com/contents/view_breaking-news/2019-09-12/fda-cleared-artificial-intelligence-breast-cancer-diagnosis-system-launched-by-paragon-biosciences/

17. Geras KJ, Mann RM, Moy L. Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives. *Radiology*. 2019;293(2):246-259.

18. Price WN, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. *JAMA*. Published online October 4, 2019. doi:10.1001/jama.2019.15064

19. Reyes M, Meier R, Pereira S, et al. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiol Artif Intell*. 2020;2(3):e190043. doi:10.1148/ryai.2020190043

20. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. Published online March 25, 2020:m689. doi:10.1136/bmj.m689

21. Nsoesie EO. Evaluating Artificial Intelligence Applications in Clinical Settings. *JAMA Netw Open*. 2018;1(5):e182658. doi:10.1001/jamanetworkopen.2018.2658

22. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342

23. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA*. 2018;320(1):2.

24. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. *Int J Hum-Comput Stud*. 2003;58(6):697-718. doi:10.1016/S1071-5819(03)00038-7

25. Clark MD, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete Choice Experiments in Health Economics: A Review of the Literature. *PharmacoEconomics*. 2014;32(9):883-902. doi:10.1007/s40273-014-0170-x

26. Ho MP, Gonzalez JM, Lerner HP, et al. Incorporating patient-preference evidence into regulatory decision making. *Surg Endosc*. 2015;29(10):2984-2993. doi:10.1007/s00464-014-4044-2

27. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw Open*. 2020;3(3):e200265. doi:10.1001/jamanetworkopen.2020.0265

28. Kuhfeld WF. Experimental design, efficiency, coding, and choice designs. *Mark Res Methods Sas Exp Des Choice Conjoint Graph Tech*. Published online 2005:47-97.

29. Johnson F, Lancsar E, Marshall D, et al. Constructing Experimental Designs for Discrete-Choice Experiments: Report of the ISPOR Conjoint Analysis Experimental Design Good Research Practices Task Force. *Value Health*. 2013;16(1):3-13. doi:10.1016/j.jval.2012.08.2223

30. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. *Patient - Patient-Centered Outcomes Res*. 2015;8(5):373-384. doi:10.1007/s40271-015-0118-z

31. Boeri M, Saure D, Schacht A, Riedl E, Hauber B. Modeling Heterogeneity in Patients' Preferences for Psoriasis Treatments in a Multicountry Study: A Comparison Between Random-Parameters Logit and Latent Class Approaches. *PharmacoEconomics*. Published online 2020:1-14.

32. Hall J, Kenny P, King M, Louviere J, Viney R, Yeoh A. Using stated preference discrete choice modelling to evaluate the introduction of varicella vaccination. *Health Econ*. 2002;11(5):457-465. doi:10.1002/hec.694

33. van Dam L, Hol L, de Bekker-Grob EW, et al. What determines individuals' preferences for colorectal cancer screening programmes? A discrete choice experiment. *Eur J Cancer*. 2010;46(1):150-159. doi:10.1016/j.ejca.2009.07.014

34. Hendrix N. Replication Data for: "Artificial intelligence in breast cancer screening: Primary care provider preferences", v1. Harvard Dataverse. Published online July 16, 2020. doi:10.7910/DVN/EX4NG2

35. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology*. 2018;286(3):800-809. doi:10.1148/radiol.2017171920

36. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol*. 2019;29(9):4825-4832. doi:10.1007/s00330-019-06186-9

37. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89-94. doi:10.1038/s41586-019-1799-6

38. Adamson AS, Welch HG. Machine Learning and the Cancer-Diagnosis Problem — No Gold Standard. *N Engl J Med*. 2019;381(24):2285-2287. doi:10.1056/NEJMp1907407

39. Miglioretti DL, Ichikawa L, Smith RA, et al. Correlation Between Screening Mammography Interpretive Performance on a Test Set and Performance in Clinical Practice. *Acad Radiol*. 2017;24(10):1256-1264. doi:10.1016/j.acra.2017.03.016

40. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ArXiv160204938 Cs Stat*. Published online February 16, 2016. Accessed October 17, 2019. http://arxiv.org/abs/1602.04938

41. Petkus H, Hoogewerf J, Wyatt JC. What do senior physicians think about AI and clinical decision support systems: Quantitative and qualitative analysis of data from specialty societies. *Clin Med*. 2020;20(3):324-328. doi:10.7861/clinmed.2019-0317

42. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial Intelligence and the Future of Primary Care: Exploratory Qualitative Study of UK General Practitioners' Views. *J Med Internet Res*. 2019;21(3):e12802. doi:10.2196/12802

43. Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *J Am Med Inform Assoc*. Published online February 27, 2020:ocz229. doi:10.1093/jamia/ocz229

44. Nundy S, Montgomery T, Wachter RM. Promoting Trust Between Patients and Physicians in the Era of Artificial Intelligence. *JAMA*. 2019;322(6):497. doi:10.1001/jama.2018.20563

45. Gastounioti A, Kontos D. Is It Time to Get Rid of Black Boxes and Cultivate Trust in AI? *Radiol Artif Intell*. 2020;2(3):e200088. doi:10.1148/ryai.2020200088

46. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. *Annu Rev Biomed Eng*. 2017;19:221-248.

47. Vickers AJ. Prediction models in cancer care. *CA Cancer J Clin*. 2011;61(5):315-326.

48. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.

49. Park SH, Goo JM, Jo C-H. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J Radiol*. 2004;5(1):11-18.

50. Moons KGM, Stijnen T, Michel BC, et al. Application of Treatment Thresholds to Diagnostic-test Evaluation: An Alternative to the Comparison of Areas under Receiver Operating Characteristic Curves. *Med Decis Making*. 1997;17(4):447-454. doi:10.1177/0272989X9701700410

51. Lobo JM, Jiménez-Valverde A, Real R. AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr*. 2008;17(2):145-151.

52. Kerr KF, Janes H. First things first: risk model performance metrics should reflect the clinical application. *Stat Med*. 2017;36(28):4503-4508.

53. Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M. Conceptualizing a model: a report of the ISPOR-SMDM modeling good research practices task force–2. *Med Decis Making*. 2012;32(5):678-689.

54. Rochau U, Jahn B, Qerimi V, et al. Decision-analytic modeling studies: an overview for clinicians using multiple myeloma as an example. *Crit Rev Oncol Hematol*. 2015;94(2):164-178.

55. Habbema JDF, Wilt TJ, Etzioni R, et al. Models in the Development of Clinical Practice Guidelines. *Ann Intern Med*. 2014;161(11):812. doi:10.7326/M14-0845

56. Birnbaum J, Gadi VK, Markowitz E, Etzioni R. The effect of treatment advances on the mortality results of breast cancer screening trials: a microsimulation model. *Ann Intern Med*. 2016;164(4):236-243.

57. Birnbaum JK, Duggan C, Anderson BO, Etzioni R. Early detection and treatment strategies for breast cancer in low-income and upper middle-income countries: a modelling study. *Lancet Glob Health*. 2018;6(8):e885-e893.

58. Connor RJ, Chu KC, Smart CR. Stage-shift cancer screening model. *J Clin Epidemiol*. 1989;42(11):1083-1095.

59. Pashayan N, Pharoah P, Tabár L, et al. Validation of a modelling approach for estimating the likely effectiveness of cancer screening using cancer data on prevalence screening and incidence. *Cancer Epidemiol*. 2011;35(2):139-144. doi:10.1016/j.canep.2010.07.012

60. Vanni T, Karnon J, Madan J, et al. Calibrating models in economic evaluation. *Pharmacoeconomics*. 2011;29(1):35-49.

61. Menzies NA, Soeteman DI, Pandya A, Kim JJ. Bayesian methods for calibrating health policy models: a tutorial. *Pharmacoeconomics*. 2017;35(6):613-624.

62. Cronin KA, Legler JM, Etzioni RD. Assessing uncertainty in microsimulation modelling with application to cancer screening interventions. *Stat Med*. 1998;17(21):2509-2523.

63. Sharif B, Kopec JA, Wong H, et al. Uncertainty analysis in population-based disease microsimulation models. *Epidemiol Res Int*. 2012;2012.

64. Weitzman ML. Why the Far-Distant Future Should Be Discounted at Its Lowest Possible Rate. *J Environ Econ Manag*. 1998;36(3):201-208. doi:10.1006/jeem.1998.1052

65. Gravelle H, Smith D. Discounting for health effects in cost–benefit and cost-effectiveness analysis. *Health Econ*. 2001;10(7):587-599.

66. Arleo EK, Hendrick RE, Helvie MA, Sickles EA. Comparison of recommendations for screening mammography using CISNET models. *Cancer*. 2017;123(19):3673-3680.

67. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77. doi:10.1186/1471-2105-12-77

68. Tabár L, Fagerberg G, Day N, Duffy S, Kitchin R. Breast cancer treatment and natural history: new insights from results of screening. *Lancet Br Ed*. 1992;339(8790):412-414.

69. Mandelblatt J, Cronin K, de Koning H, Miglioretti D, Schechter C, Stout N. Collaborative modeling of US breast cancer screening strategies. *Rockv MD Agency Healthc Res Qual US Dep Health Hum Serv*. Published online 2015.

70. Knudsen AB, McMahon PM, Gazelle GS. Use of modeling to evaluate the cost-effectiveness of cancer screening programs. *J Clin Oncol*. 2007;25(2):203-208.

71. Wagstaff K. Machine learning that matters. *ArXiv Prepr ArXiv12064656*. Published online 2012.

72. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda*. The University of Chicago Press; 2019.

73. Langlotz CP. Will Artificial Intelligence Replace Radiologists? *Radiol Artif Intell*. 2019;1(3):e190058. doi:10.1148/ryai.2019190058

74. Liew C. The future of radiology augmented with Artificial Intelligence: A strategy for success. *Eur J Radiol*. 2018;102:152-156. doi:10.1016/j.ejrad.2018.03.019

75. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *J Clin Oncol*. 2016;34(21):2534-2540. doi:10.1200/JCO.2015.65.5654

76. Valeri L, Chen JT, Garcia-Albeniz X, Krieger N, VanderWeele TJ, Coull BA. The Role of Stage at Diagnosis in Colorectal Cancer Black-White Survival Disparities: A Counterfactual Causal Inference Approach. *Cancer Epidemiol Biomarkers Prev*. 2016;25(1):83-89. doi:10.1158/1055-9965.EPI-15-0456

77. Swanson SA, Holme Ø, Løberg M, et al. Bounding the per-protocol effect in randomized trials: an application to colorectal cancer screening. *Trials*. 2015;16(1):541. doi:10.1186/s13063-015-1056-8

78. Robertson S, Azizpour H, Smith K, Hartman J. Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. *Transl Res*. 2018;194:19-35.

79. Sadoughi F, Kazemy Z, Hamedan F, Owji L, Rahmanikatigari M, Azadboni TT. Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. *Breast Cancer Targets Ther*. 2018;10:219.

80. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980;302(20):1109-1117.

81. Pepe MS, Feng Z, Huang Y, et al. Integrating the Predictiveness of a Marker with Its Performance as a Classifier. *Am J Epidemiol*. 2007;167(3):362-368. doi:10.1093/aje/kwm305

82. Phelps CE, Mushlin AI. Focusing Technology Assessment Using Medical Decision Theory. *Med Decis Making*. 1988;8(4):279-289. doi:10.1177/0272989X8800800409

83. Laking G, Lord J, Fischer A. The economics of diagnosis. *Health Econ*. 2006;15(10):1109-1120. doi:10.1002/hec.1114

84. Neumann PJ, Greenberg D. Is the United States ready for QALYs? *Health Aff (Millwood)*. 2009;28(5):1366-1371.

85. Savulescu J, Persson I, Wilkinson D. Utilitarianism and the Pandemic. *Bioethics*. Published online May 20, 2020:bioe.12771. doi:10.1111/bioe.12771

86. Asaria M, Griffin S, Cookson R. Distributional Cost-Effectiveness Analysis: A Tutorial. *Med Decis Making*. 2016;36(1):8-19. doi:10.1177/0272989X15583266

87. Brouwer E, Carlson J. PNS43 IDENTIFYING AND ASSESSING THE FEASIBILITY OF PROPOSED ALTERNATIVE APPROACHES TO QALY ESTIMATION WITHIN COST-EFFECTIVENESS MODELS USING A SYSTEMATIC LITERATURE REVIEW: AN UPDATED ANALYSIS. *Value Health*. 2019;22:S293-S294.

88. Basu A, Carlson J, Veenstra D. Health Years in Total: A New Health Objective Function for Cost-Effectiveness Analysis. *Value Health*. 2020;23(1):96-103.

89. Mulherin SA, Miller WC. Spectrum Bias or Spectrum Effect? Subgroup Variation in Diagnostic Test Evaluation. *Ann Intern Med*. 2002;137(7):598. doi:10.7326/0003-4819-137-7-200210010-00011

90. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *ArXiv190912475 Cs Stat*. Published online September 26, 2019. Accessed October 16, 2019. http://arxiv.org/abs/1909.12475

91. Balthazar P, Harri P, Prater A, Safdar NM. Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics. *J Am Coll Radiol*. 2018;15(3):580-586.

92. Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med*. 2019;25(1):37-43.

93. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.

# VITA

Nathaniel Hendrix received a Doctorate in Pharmacy from the University of Washington in 2015, and enrolled in the Doctor of Philosophy program in the Comparative Health Outcomes, Policy & Economics (CHOICE) Institute soon after. His research has focused on health information technologies, pharmacoepidemiology, cost-effectiveness analysis, and vaccines in low- and middle-income countries. He is the co-founder of the CHOICE Institute's student blog, *Incremental Thoughts*, and has developed curricula for teaching cost-effectiveness and claims analysis to his colleagues.