Cancer Causes & Control (2019) 30:1145–1155
https://doi.org/10.1007/s10552-019-01208-9

ORIGINAL PAPER

# Improving the diagnostic accuracy of a stratified screening strategy by identifying the optimal risk cutoff

John T. Brinton[1,5] · R. Edward Hendrick[2] · Brandy M. Ringham[3] · Mieke Kriege[4] · Deborah H. Glueck[5]

Received: 10 August 2018 / Accepted: 29 June 2019 / Published online: 3 August 2019


## Abstract

**Background** The American Cancer Society (ACS) suggests using a stratified strategy for breast cancer screening. The strategy includes assessing risk of breast cancer, screening women at high risk with both MRI and mammography, and screening women at low risk with mammography alone. The ACS chose their cutoff for high risk using expert consensus.

**Methods** We propose instead an analytic approach that maximizes the diagnostic accuracy (AUC/ROC) of a risk-based stratified screening strategy in a population. The inputs are the joint distribution of screening test scores, and the odds of disease, for the given risk score. Using the approach for breast cancer screening, we estimated the optimal risk cutoff for two different risk models: the Breast Cancer Screening Consortium (BCSC) model and a hypothetical model with much better discriminatory accuracy. Data on mammography and MRI test score distributions were drawn from the Magnetic Resonance Imaging Screening Study Group.

**Results** A risk model with an excellent discriminatory accuracy (c-statistic = 0.947) yielded a reasonable cutoff where only about 20% of women had dual screening. However, the BCSC risk model (c-statistic = 0.631) lacked the discriminatory accuracy to differentiate between women who needed dual screening, and women who needed only mammography.

**Conclusion** Our research provides a general approach to optimize the diagnostic accuracy of a stratified screening strategy in a population, and to assess whether risk models are sufficiently accurate to guide stratified screening. For breast cancer, most risk models lack enough discriminatory accuracy to make stratified screening a reasonable recommendation.

**Keywords** Cancer screening · Stratified screening · Risk assessment · ROC analysis

## Introduction

In cancer screening, more intensive screening strategies are often recommended for people whose risk of cancer exceeds a cutoff value [1, 4, 25, 33]. People at lower risk are given less intense screening or no screening at all. The approach can be termed a *stratified screening strategy*. In the United States, stratified screening strategies are recommended for breast cancer, endometrial cancer, and colorectal cancer [35, 36].

Stratified screening matches screening intensity to risk. The underlying assumption of stratified screening is that for people at high risk, the benefits of detection outweigh the possible harms of additional screening. The early detection of cancers can reduce mortality, and improve the quality and length of life [16, 30]. Yet cancer screening can also lead to false-positive exams, undesirable events for anyone seeking cancer screening. The problem is compounded with regular screening, since even a low false-positive rate can

Data collection and sharing for the BCSC was supported by the National Cancer Institute (U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, U01CA70040, HHSN261201100031C).

✉ John T. Brinton
john.brinton@ucdenver.edu

[1] Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO, USA

[2] Department of Radiology, School of Medicine, University of Colorado Denver, Aurora, CO, USA

[3] Lifecourse Epidemiology of Adiposity and Diabetes (LEAD) Center, University of Colorado Denver, Aurora, CO, USA

[4] Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands

[5] Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO, USA

lead to a high cumulative false-positive rate. For example, regular mammography screening will result in at least one false-positive exam for 50% of women who undergo yearly mammography over 10 years [12].

While stratified screening strategies have some appeal, the question is how to decide who should get intensive screening. In many cases, expert panels choose the risk level at which more intensive screening is recommended. Although expert panels usually select the risk cutoff after a review of evidence, development of guidelines may still be subjective. To illustrate the issue, consider the American Cancer Society (ACS) guidelines for adjunct breast MRI screening [33]. The ACS recommends that women over 30 years of age who have between 20 and 25% lifetime risk of breast cancer be screened with contrast-enhanced breast MRI in addition to mammography [33]. No information is given in the document as to how the cutoff for additional screening was chosen. In addition, a risk cutoff between 20 and 25% may leave women and their personal physicians without clear guidance for how to proceed.

Perhaps in light of this lack of clear guidance, Dr. Otis Brawley, the Chief medical and scientific officer of the American Cancer Society, argues that processes for developing risk cutoffs for cancer screening should be transparent [9]. One approach to make the process more transparent is to use an analytic approach to determine an optimal screening strategy, a tack taken by a variety of authors. Pepe et al. [26] assessed the classification power of biomarkers. Baker et al. [6] suggested using a utility function to assess whether combinations of biomarkers would result in a useful screening regimen. Gail and Pfeiffer [13] used a loss-function approach to model a risk-tool-based decision for further screening. In a loss-function approach, one assigns a penalty to wrong decisions. Wrong decisions include falsely diagnosing disease in someone who is disease-free, and incorrectly declaring no disease in someone who has cancer.

In response, we propose instead a metric designed to measure the diagnostic accuracy of a stratified screening strategy, without consideration of cost or loss. The approach mirrors that used in the medical literature, where criteria based on accuracy often appear before any consideration of cost. For example, it took almost 3 years after Pisano et al. [27] compared the diagnostic accuracies of digital and film mammography for Tosteson et al. [38] to compare the cost-effectiveness of the two modalities. The rapid adoption of digital mammography, despite the lack of documentation of cost-effectiveness, suggests that accuracy, not cost, is the true metric that drives adoption of screening techniques.

In order to provide a method to quantify the population-wide accuracy of a screening strategy, we define a population-based metric of diagnostic accuracy, based on a population-wide receiver-operating characteristic (ROC) curve. A population-wide ROC curve for a strategy is a measure of the sensitivity and specificity of the strategy across all risk spectra of the population. For the population-wide curve, appropriate weighting is used to account for the differential risk of disease within each stratum. In turn, the area under the population-wide ROC curve quantifies the diagnostic accuracy of a screening strategy.

Using area under the ROC curve (AUC) as a metric, we propose a method to optimize the risk cutoff that defines the more intensely screened group. To illustrate the utility of the method, we find an optimal risk cutoff for the Breast Cancer Surveillance Consortium (BCSC) risk model [7]. For comparison, we apply the method to a hypothetical example where the risk-assessment model has much greater discriminatory accuracy than the BCSC risk model. Results of this work are intended to help investigators choose evidence-based risk cutoffs that optimize the diagnostic accuracy of a stratified screening strategy.

## Methods

### Analytic approach

We describe a method to obtain a risk cutoff that maximizes the diagnostic accuracy of a stratified screening strategy in a population. The approach requires two inputs: (1) the joint distribution of screening test scores, and (2) the odds of disease given the risk score.

#### Assumptions and notation

Figure 1 shows a stratified screening strategy similar to the one suggested for breast cancer by Saslow et al. [33]. Every person undergoes risk assessment. People at high risk are given two screening tests. People at low risk are given one screening test. A high score on either or both screening tests indicates that the strategy is positive for disease. Typically, if the strategy is positive, a study participant would be referred for further work-up which may lead to a confirmatory test, such as biopsy. For the purpose of this research, we assume that when a person has two screening tests with different scores, the maximum score of the two tests is used. This "worst-scenario" approach corresponds to that described in Sardanelli et al. [32, p. 98]

We assume that the two screening tests used in a stratified screening strategy have the same scale, and thus, the same threshold for suspicion of disease on either test. The requirement that the two screening tests are scored on the same scale is often made so that clinicians do not have to make assumptions about how one scale compares to another. It is a common practice in breast cancer screening, in which film-screen mammography, digital mammography, and

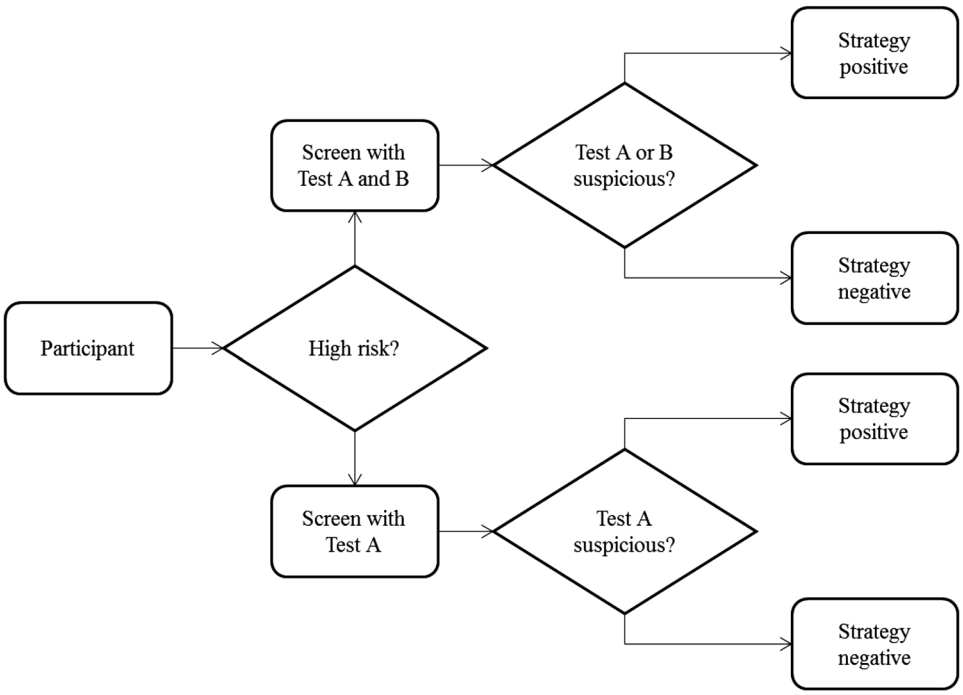



**Fig. 1** Stratified screening strategy with two screening tests

contrast-enhanced breast MRI are usually rated using the same scale [14, 15, 18–22, 32, 43, 44], known as BI-RADS [34].

Finally, we assume that both the receiver-operating characteristic (ROC) curves and the diagnostic accuracies (AUCs) of the screening tests are independent of the risk score. This assumption is reasonable when the physical process by which the screening test operates does not depend on the same process that modifies risk. For example, the sensitivity and specificity of screening breast MRI appear to be independent of family history, a surrogate measure of risk [18–20, 22].

Individuals at high risk are given two ordinal categorical screening tests: screening test $A$ and screening test $B$. Individuals at low risk are given only one screening test, test $A$. For each screening test, assume that the higher the screening test score, the higher the likelihood of disease. We assume that the two screening tests are scored independently.

Let $y_{idj}$ be the screening test score for individual $i \in \{1, 2, \ldots, N\}$, with disease status $d \in \{c, n\}$ on the screening test $j \in \{A, B\}$. Here $c$ indicates the presence of disease and $n$ indicates the absence of disease. Let $T$ be the number of possible outcomes for the two screening tests, so that each test can take on one value from the possible set of outcomes $t \in \{1, 2, \ldots, T\}$. Test scores for individuals with disease and without disease each have a potentially bivariate discrete probability mass function. For a given disease status and for $t_A$ and $t_B \in \{1, 2, \ldots, T\}$, assume that the two tests have a bivariate discrete probability mass function $f_{A,B|d}\left(y_{idA} = t_A, y_{idB} = t_B | d\right) = w_{t_A t_B d}$, where $w_{t_A t_B d}$ is the probability that screening test $A$ and screening test $B$ take on values $t_A$ and $t_B$, respectively. Under these assumptions, we have that

$$\sum_{t_A=1}^{T} \sum_{t_B=1}^{T} w_{t_A t_B n} = 1, \tag{1}$$

and

$$\sum_{t_A=1}^{T} \sum_{t_B=1}^{T} w_{t_A t_B c} = 1. \tag{2}$$

Recall our assumption of a common cutoff score of $\theta$ for both screening tests. For a high-risk participant, the stratified screening strategy is positive if $\left(y_{idA} \geq \theta\right)$, $\left(y_{idB} \geq \theta\right)$, or both. For a low-risk participant, the stratified screening strategy is positive only if $\left(y_{idA} \geq \theta\right)$.

Let $f_Z(z)$ and $F_Z(z)$ be the probability density and cumulative distribution functions for the risk scores, respectively, where $Z \in \left[z_{\min}, z_{\max}\right]$. Here, the risk score is assumed to have a continuous distribution. Let $\lambda \in \left[z_{\min}, z_{\max}\right]$ be the risk cutoff for categorizing an individual as low or high risk for cancer: individuals with risk scores less than $\lambda$ are deemed low risk, while those with risk scores greater than or equal to $\lambda$ are high risk. Let the index $k \in \{l, h\}$ denote risk strata. The probability that a person will be classified as low risk is $r_l(\lambda) = Pr[Z < \lambda]$. The probability of being high risk is $r_h(\lambda) = 1 - r_l(\lambda)$.

Let $\pi$, $\pi_h(\lambda)$ and $\pi_l(\lambda)$ be the prevalences of disease in the general screening population, the high-risk stratum, and the low-risk stratum, respectively. Let $D^+$ be the event an individual has disease. Let $\Pr\{D^+|Z=z\}$ be the conditional probability an individual has disease, given a risk score. From Eqs 6.6, p. 278, [29], the prevalence of disease in each stratum is given by

$$\pi_l(\lambda) = \int_{z_{\min}}^{\lambda} \Pr\{D^+|Z=z\} f_Z(z)dz \tag{3}$$

and

$$\pi_h(\lambda) = \pi - \pi_l(\lambda) = \int_{\lambda}^{z_{\max}} \Pr\{D^+|Z=z\} f_Z(z)dz. \tag{4}$$

Note that with $\lambda = z_{\max}$, $\pi_l(\lambda) = \pi_l(z_{\max}) = \pi$. Similarly, with $\lambda = z_{\min}$, $\pi_h(\lambda) = \pi_l(z_{\min}) = \pi$.

For conciseness, we write $\pi_l(\lambda)$ and $\pi_h(\lambda)$ as $\pi_l$ and $\pi_h$, respectively. Similarly, $r_l$ and $r_h$ denote the proportion of the entire population classified as low risk and high risk, respectively.

### Sensitivity and specificity of the stratified screening strategy

The sensitivity and specificity of the stratified screening strategy depend on the sensitivity and specificity values for the two component screening tests. The sensitivity for screening test $A$ is given by

$$s_A = \sum_{t_A=\theta}^{T} \sum_{t_B=1}^{T} w_{t_A t_B c}. \tag{5}$$

The sensitivity for screening test $B$ is given by

$$s_B = \sum_{t_A=1}^{T} \sum_{t_B=\theta}^{T} w_{t_A t_B c}. \tag{6}$$

The specificity of screening test $A$ is given by

$$p_A = \sum_{t_A=1}^{\theta-1} \sum_{t_B=1}^{T} w_{t_A t_B c} \tag{7}$$

and the specificity for screening test $B$ is given by

$$p_B = \sum_{t_A=1}^{T} \sum_{t_B=1}^{\theta-1} w_{t_A t_B c}.. \tag{8}$$

To derive formulae for the sensitivity and specificity of the stratified screening strategy, we give the probability of every possible outcome for the stratified screening strategy (Table 1).

The number of true positives (TP) for the strategy is the number of people called positive by the strategy who do, in fact, have cancer. False positives (FP), true negatives (TN), and false negatives (FN) are defined similarly. Then the sensitivity of the strategy is given by

$$\text{Sens}(\theta, \lambda) = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{9}$$

and the specificity by

$$\text{Spec}(\theta, \lambda) = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{10}$$

where

$$j_h = r_h \times \pi_h \times \left\{ \sum_{t_A=\theta}^{T} \sum_{t_B=\theta}^{T} w_{t_A t_B c} \right\} \tag{11}$$

**Table 1** Stratified screening strategy outcomes and associated probabilities

| Risk Level | Disease | Test A | Test B | Strategy | Probability |
|---|---|---|---|---|---|
| High | + | + | + | + | $j_h$ |
| High | + | − | + | + | $s_B \cdot \pi_h \cdot r_h - j_h$ |
| High | + | + | − | + | $s_A \cdot \pi_h \cdot r_h - j_h$ |
| High | + | − | − | − | $(\pi_h \cdot r_h - s_B \cdot \pi_h \cdot r_h) - (s_A \cdot \pi_h \cdot r_h - j_h)$ |
| High | − | + | + | + | $(r_h - \pi_h \cdot r_h) - p_A \cdot (r_h - \pi_h \cdot r_h) - (p_B \cdot (r_h - \pi_h \cdot r_h) - g_h)$ |
| High | − | − | + | + | $p_A \cdot (r_h - \pi_h \cdot r_h) - g_h$ |
| High | − | + | − | + | $p_B \cdot (r_h - \pi_h \cdot r_h) - g_h$ |
| High | − | − | − | − | $g_h$ |
| Low | + | + | No value | + | $s_A \cdot \pi_l \cdot r_l$ |
| Low | + | − | No value | − | $\pi_l - s_A \cdot \pi_l \cdot r_l$ |
| Low | − | + | No value | + | $(r_l - \pi_l \cdot r_l) - p_A \cdot (r_l - \pi_l \cdot r_l)$ |
| Low | − | − | No value | − | $p_A \cdot (r_l - \pi_l \cdot r_l)$ |

$$g_h = r_h \times (1 - \pi_h) \times \left\{ \sum_{t_A=1}^{\theta-1} \sum_{t_B=1}^{\theta-1} w_{t_A t_B n} \right\}, \tag{12}$$

$$\begin{aligned} TP = & (j_h) + (s_B \cdot \pi_h \cdot r_h - j_h) \\ & + (s_A \cdot \pi_h \cdot r_h - j_h) + (s_A \cdot \pi_l \cdot r_l), \end{aligned} \tag{13}$$

$$\begin{aligned} FP = & \left[ (r_h - \pi_h \cdot r_h) \right. \\ & - p_A \cdot (r_h - \pi_h \cdot r_h) - \left. (p_B \cdot (r_h - \pi_h \cdot r_h) - g_h) \right] \\ & + \left[ p_A \cdot (r_h - \pi_h \cdot r_h) - g_h \right] \\ & + \left[ p_B \cdot (r_h - \pi_h \cdot r_h) - g_h \right] \\ & + \left[ (r_l - \pi_l \cdot r_l) - p_A \cdot (r_l - \pi_l \cdot r_l) \right], \end{aligned} \tag{14}$$

$$\begin{aligned} FN = & \left[ (\pi_h \cdot r_h - s_B \cdot \pi_h \cdot r_h) - (s_A \cdot \pi_h \cdot r_h - j_h) \right] \\ & + (\pi_l - s_A \cdot \pi_l \cdot r_l), \end{aligned} \tag{15}$$

and

$$\mathrm{TN} = g_h + \left[ p_A \cdot (r_l - \pi_l \cdot r_l) \right]. \tag{16}$$

### Calculating the diagnostic accuracy of the stratified screening strategy

Equations (9) and (10) define a receiver-operating characteristic curve for the stratified screening strategy as a function of the risk cutoff, lambda. We can assess the diagnostic accuracy of the stratified screening strategy for each risk cutoff using the area under the receiver-operating characteristic curve (AUC). We calculate the AUC of the stratified screening strategy for each risk cutoff using a trapezoidal rule approximation [37].

Recall that in a stratified screening strategy, all people are assessed for risk of disease using a risk model, and those people with risk scores above a risk cutoff are given more intensive screening. Both the ROC curve and the AUC for the stratified screening strategy depend on the choice of risk cutoff for the strategy. Because the ROC and AUC depend on the risk cutoff, the diagnostic accuracy of the stratified screening strategy is a function of the risk cutoff.

### Finding the risk cutoff that maximizes the diagnostic accuracy of the stratified screening strategy

We used a grid search approach [37] to identify an optimal risk cutoff for the strategy, $\lambda^*$, which maximizes the AUC of the strategy. When no maximum exists, one approach is to select $\lambda^* = \lambda_{\max}$. This choice corresponds to a screening strategy where all participants are given only Test $A$. Alternatively, one could choose $\lambda^* = \lambda_{\min}$ . This cutoff choice corresponds to adopting a screening strategy where all participants are given both screening tests. Giving everyone in the population both screening tests is a useful strategy when the two screening tests produce better diagnostic accuracy than either component screening test alone.

## Methods for real and hypothetical examples

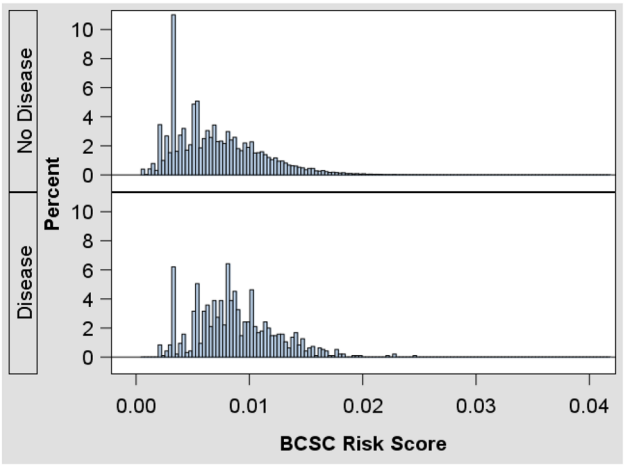
To evaluate the performance of the proposed stratified screening strategy, we require the following information: (1) the joint distribution of screening test scores and (2) the distribution of the risk scores in the general population. From this information, we are able to calculate the AUC of the strategy for a given risk score cutoff.

We provide two examples. In the first example, we use a risk model with moderate discriminatory accuracy developed with data from the Breast Cancer Screening Consortium (BCSC) [7]. In the second example, we use a hypothetical risk model with high discriminatory accuracy. The hypothetical risk model has higher discriminatory accuracy than the BCSC model (c-statistics of 0.947 versus 0.631).

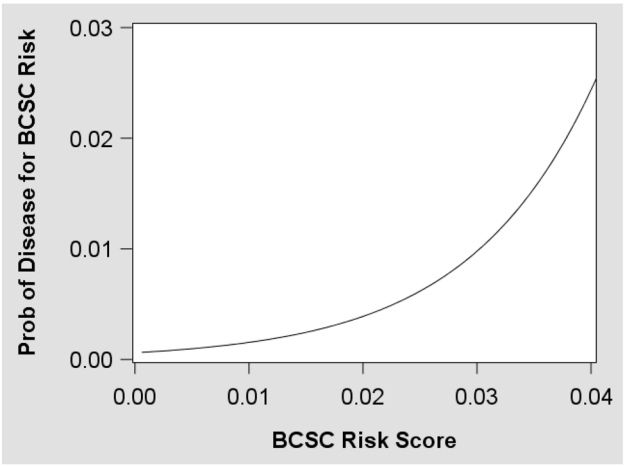
### Breast Cancer Screening Consortium example

We used publicly available data from the Breast Cancer Screening Consortium (BCSC). The data included the presence or the absence of cancer, and a risk score calculated using the BCSC model [7]. The BCSC model is a risk-assessment tool based on data from over 2 million women, aged 35 years or older, who had no previous breast cancer and did not have breast augmentation. The data are from community-based registries and cover a broad geographic swath of the United States. The BCSC model provides an estimate of the 5-year risk of breast cancer. Scores on the BCSC risk model range between 0.0 and 0.04. A score of 0.04 reflects a 5-year probability of cancer of approximately 0.025. The risk-assessment tool had a $c$ statistic of 0.631 (95% confidence interval [CI] = 0.618–0.644) for pre-menopausal women and 0.624 (95% CI = 0.619–0.630) for post-menopausal women [7]. The $c$ statistic is a measure of discriminatory accuracy of a risk prediction model, and ranges between 0.5 for a risk-assessment tool no better than chance and 1.0 for a perfect risk-assessment tool [17].

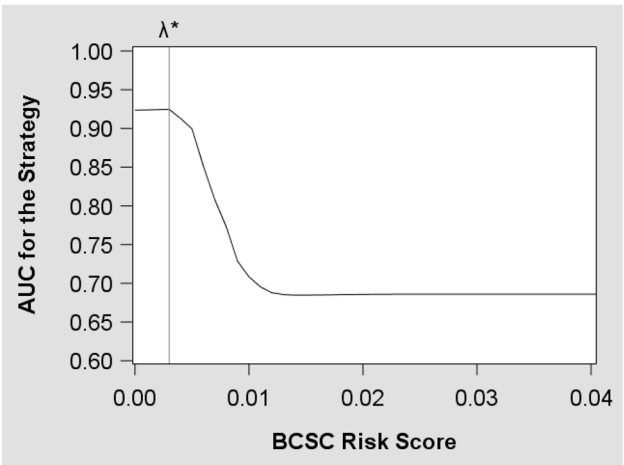
The empirical distribution of risk scores in the screening population, $f_z(z)$, was estimated from the BCSC data [7]. We show the empirical distributions of the risk scores conditional on disease status in Fig. 2a. Note that the calculation of the probability of breast cancer given the risk score uses the unconditional empirical distribution of the risk scores (i.e., the risk scores for both the cancers and the noncancers




(a) Distribution of BCSC Risk Scores




(b) Probability of Disease Given the BCSC Risk Score




(c) AUC of Strategy by Risk Cutoff

**Fig. 2** Optimal risk cutoff for adjunct screening using the BCSC risk model

combined). We display the conditional empirical distributions to illustrate the large amount of overlap between the two distributions. The large overlap is the underlying reason for the BCSC model's modest discriminatory ability.

Most models predict incident risk; however, for the application of our method, we utilize the BCSC risk model data to estimate the prevalence of cancer in a screening population. The probability of breast cancer given the risk score, $\Pr\{D^+|Z=z\}$, was estimated from the BCSC assigned five-year risk scores by regressing the risk score on disease status indicator. The probability of disease given the risk score was estimated with the inverse logit as shown in Eq. (17), with $\beta_0 = -7.4$ and $\beta_1 = 92.7$,

$$\Pr\{D^+|Z=z\} = \frac{\exp(\beta_0 + z\beta_1)}{\exp(\beta_0 + z\beta_1) + 1}. \tag{17}$$

The probability of disease given the BCSC risk score is shown in Fig. 2b.

The joint distribution of MRI and mammography test scores was obtained from unpublished data from the study of Kriege et al. [18]. Briefly, 1909 women with a cumulative lifetime risk of 15% or more were screened every year by mammography and MRI. Each modality was scored independently using a standardized Breast Imaging Reporting and Data System (BI-RADS) scale. Women were followed for the development of breast cancer for a median of 2.9 years. Of the 1909 women, 1795 had data on both mammography and MRI for at least one visit. Most women had multiple rounds of screening and BI-RADS scores for both mammography and MRI. To ensure that each woman appeared in the dataset only once, we used the following approach. For women who eventually developed breast cancer ($N = 45$), the last screening scores before diagnosis were used. For women with no evidence of breast cancer, ($N = 1750$), one pair of BI-RADS scores, including one for mammography and one for MRI, were chosen at random from all of the woman's screening examinations. The AUC for mammography alone was 0.686, while the AUC for screening breast MRI was 0.827 [18].

### Hypothetical Risk Model Example

We were curious to evaluate whether a risk prediction tool with better discriminatory accuracy yielded a clearer choice for a risk model cutoff, and thus we simulated a stratified screening strategy with a better diagnostic accuracy.

Risk models have low discriminatory ability if the distribution of risk scores for women with disease and women without disease substantially overlap, the case is shown for the BCSC risk tool results in Fig. 2a. Instead, we made distributional assumptions for the risk scores so that there was

a strong separation between scores for women with disease and without disease. We assumed that risk scores for women without breast cancer had a beta distribution with parameters $\alpha = 3$ and $\beta = 21$. We assumed that the risk scores for women with breast cancer were beta-distributed with parameters $\alpha = 9$ and $\beta = 21$. We fixed the prevalence of disease at 0.006, the prevalence of breast cancer observed in Pisano et al., [27]. The resulting distribution of risk scores for the entire population, $f_z(z)$, is a mixture of the two beta distributions. The simulated risk score distributions for 100,000 women without disease and 600 women with disease are shown in Fig. 3a.

Using logistic regression on the simulated data, we fit the probability of disease given the risk score $\Pr\{D^+|Z = z\}$. The model fit yielded $\beta_0 = -7.73$, and $\beta_1 = 29.64$, using Eq. (17). The model had good predictive accuracy with a $c$ statistic of 0.947, much better than the $c$ statistic for the BCSC risk-assessment tool [7]. The probability of disease given the risk score is plotted in Fig. 3b.

The joint distribution of test scores for mammography and MRI was obtained as described above.

## Results

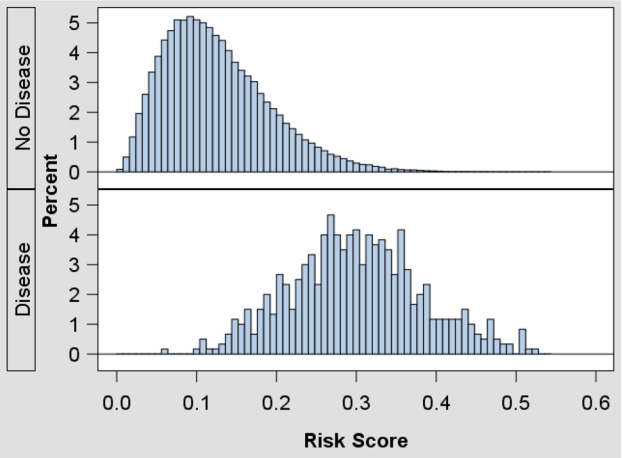
### Optimal risk score cutoff based on the Breast Cancer Screening Consortium (BCSC) Model

As shown in Fig. 2c, the optimal risk score cutoff for maximizing the AUC of a stratified screening strategy that uses the BCSC risk-assessment model is $\lambda^* = 0.0031$, which corresponds to a five-year breast cancer probability of 0.0008. The optimal risk cutoff appears in Fig. 2c as a vertical line occurring at the risk score where the AUC is maximized. Using this risk cutoff value yields an AUC for the strategy of 0.885.

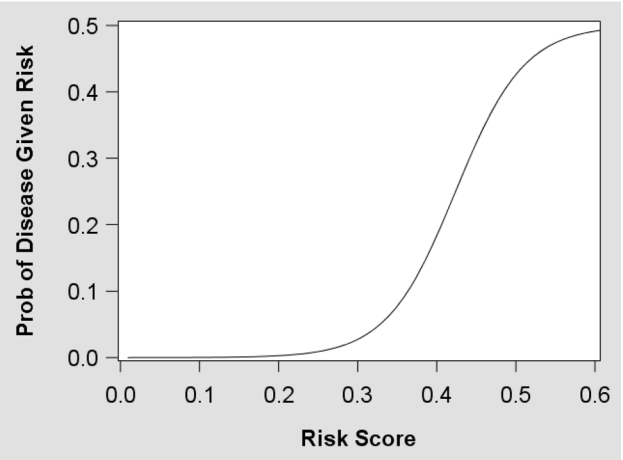
A stratified screening strategy with a risk cutoff of $\lambda^* = 0.0031$ would mean that every woman in the screening population with a BCSC 5-year risk score greater than 0.0031 would be screened with both digital mammography and contrast-enhanced screening breast MRI. Based on the BCSC data this would result in more than 99% of the general screening population receiving both tests, an absurd result. The BCSC model lacks the discriminatory accuracy to differentiate between women at high risk, who require screening with both mammography and MRI, and women at low risk, who require mammography alone.

### Optimal risk cutoff based on a hypothetical risk model with high discriminatory accuracy

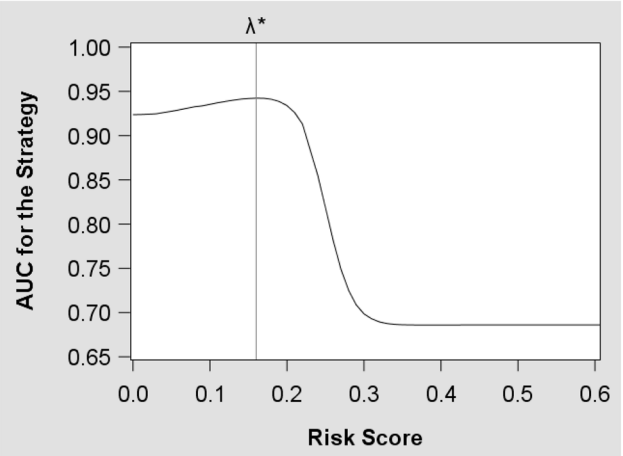
As shown in Fig. 3c, the optimal risk score cutoff for maximizing the AUC of a stratified screening strategy that uses




(a) Distribution of Risk Scores




(b) Probability of Disease Given the Risk Score




(c) AUC of the Strategy by Risk Cutoff

**Fig. 3** Optimal risk cutoff for adjunct screening using a hypothetical model with high discriminatory accuracy

the hypothetical risk-assessment model is $\lambda^* = 0.16$. Again, the optimal risk cutoff appears in Figure 3c as a vertical line and occurs at the point where the AUC is maximized. Using this risk cutoff value yields an AUC for the strategy of 0.90.

The results indicate that in order to optimize the AUC of a stratified screening strategy based on the hypothetical risk-assessment tool considered in this section, every woman with a risk score greater than 0.16 on the hypothetical risk model should be screened with both digital mammography and contrast-enhanced screening breast MRI. This would result in 20% of the population receiving both tests, with the rest screened only with digital mammography. When a risk model has excellent discriminatory accuracy, the choice of a cutoff is clear. Using this cutoff maximizes the accuracy of the stratified screening strategy for the entire screening population.

## Discussion

We demonstrate a single analytic approach for identifying an optimal cutoff for a risk-based stratified screening strategy. The approach maximizes the AUC of the stratified screening approach in a screening population. The approach uses mathematical criteria and empirical data, rather than expert opinion to identify a risk threshold for adjunct screening with a secondary test. The work presented in this manuscript has the potential to inform cancer screening recommendations for a variety of disease sites, in addition to the breast cancer case considered in the examples.

If the only risk models available have poor discriminatory accuracy, using a risk model to determine the intensity of screening may not be a good approach, a finding that mirrors that of Wald et al. [40]. The discriminatory accuracy of a risk model measures the probability that a risk model will correctly differentiate between those who will develop disease and those who will not. Using a risk-assessment model with low discriminatory accuracy means that the risk model often provides incorrect classification. The risk model cannot discriminate between those who will and those who will not develop disease. Because the risk model is often wrong, there is essentially no improvement between a strategy where all women are screened with both tests (shown on the far left of the horizontal axis in Figure 2c), and a stratified screening strategy where women are first risk assessed and then receive screening based on their personal risk.

Using a risk model with poor discriminatory accuracy, any algorithm seeking to choose an optimal risk cutoff for a stratified screening strategy will produce unacceptable results. Using the BCSC risk model, we obtained a risk cutoff that suggested 99% of women presenting for breast cancer screening should be screened with both contrast-enhanced breast MRI and mammography. Yet such a strategy would never be accepted by women and their physicians, or third-party payers. It could be clinically implemented with fast breast MRI used everywhere, but costs would be enormous.

Our results indicate that matching screening intensity to risk is a good strategy only when the risk-assessment tool has good discriminatory accuracy. This result agrees with those of other authors, who used alternative rationales to achieve the same conclusion [6, 13]. While no current risk model for breast cancer has a c-statistic as good as the model we posit in Section 3.2 [2], it is the hope that future models that incorporate genetic and epigenetic information may perform better. With better risk models, stratified screening strategies using our optimization strategy would be practical and would improve cancer screening.

One limitation of our work is that the real data example we chose uses the BCSC risk model. It is important to note that the ACS [33] recommended evaluating lifetime risk of breast cancer using models that are largely dependent on detailed family history, such as the BRCAPRO, Claus, or Tyrer-Cuzick models [8, 10, 39], not the BCSC model. However, evaluation of the performance of any one of these models is not currently possible. Our method requires, as an input, the probability of disease (prevalent screen detectable disease), given the risk score. Yet this distribution for the BRCAPRO, BOADICEA, Claus, or Tyrer-Cuzick models [3, 8, 10, 39] is not readily available. Although Amir et al. [2] evaluated the probability of breast cancer incidence given risk assessment with either the Tyrer-Cuzick or Claus models [2, Figure 1, p 812], their data are not publicly accessible.

Another possible limitation is that we assumed that the risk score was independent of the performance of the two screening tests. Yet the BCSC model includes breast density as an input [7]. In addition, breast density is associated with the sensitivity and specificity of mammography [27]. This contravenes our assumption of the independence of the risk-assessment model and the performance of the screening tests. We chose to keep the BCSC example, since we could find no other published, freely available dataset containing the population distribution of risk scores.

One other potential limitation is our use of the Kriege et al.'s [18] data to estimate the joint distribution of the mammography and MRI scores. The Kriege et al.'s [18] study used mostly screen-film mammography. Since their study, there has been widespread adoption of digital mammography. In addition, since the Kriege et al.'s [18] study occurred, radiologists' experience and their skills with breast MRI have increased. Even with these potential limitations, it is unlikely that the results of our analysis would change much with updated data. In addition, the goal of this manuscript is to demonstrate the applicability of our method. In the future, our results could be updated using

new information once updated data on the diagnostic accuracy of MR and mammography become available.

A potential bias occurs in our results because Kriege et al. [18] only enrolled women at high lifetime risk. We used these data under our assumption that the screening test score is independent of the risk, and hence, any estimate, even an estimate from a high-risk population, would be valid. This in fact may not be true. If we could obtain data from a general, low-risk population on the joint distribution of mammography and MR scores, we could evaluate the validity of the assumption. Such data are difficult to obtain.

Some authors have suggested using partial AUC or the Youden's index instead of the full area under the receiver-operating characteristic curve [23, 24], the metric used in this manuscript. Our rationale for using full area under the curve follows. In most breast cancer papers [18–20, 22, 27], the full area under the curve is used as the metric, due to the nature of the detection task. For a continuous biomarker, typically follow-up testing is only done for extreme values of the biomarker, i.e., for parts of the curve where sensitivity is high and specificity is low. When radiologists review mammography or breast MRI images, sometimes a radiologist will see a salient detail that nintey-nine out of one hundred other readers would miss. Thus, detection of cancer may occur even in cases where the sensitivity is low, and the specificity is high. Thus, consideration of the full curve reflects the true clinical picture. In many cases other than breast cancer screening [41], a partial area under the curve, both for the test and for the population, may have merit. An extension of our method could be achieved by changing the numerical algorithm to use partial AUC, or the Youden's index as the metric.

Extensions to situations with multiple screening tests could certainly be considered in the framework we have set up in this paper. Another extension would be to derive similar probabilistic approaches for more complex stratified screening strategies, such as those with more than one risk-assessment tool or more than two possible screening approaches. Radiologists conducting breast cancer screening might be interested in guidelines for adding either whole-breast ultrasound or breast MRI to mammography. Finally, the method could easily apply to models which predict risk for short fixed time horizons ($\leq$ 10 years), rather than for the remaining lifetime. The advantage of using short time horizon risk models is that most risk models are only validated for short horizons, and are therefore more accurate in the short term [28].

In this manuscript, we provide an approach for optimizing the diagnostic accuracy of a stratified screening strategy by choosing an appropriate risk cutoff. Yet in the end, diagnostic accuracy should not be the only factor used to determine how one should screen for cancer. The most important factor used to determine whether and what kind of screening should be implemented is whether the screening program leads to a mortality reduction and how large the mortality reduction is. Many countries have suggested adding MRI to mammography screening programs for high-risk groups [31, 33]. The argument for adding MRI to the screening regimen has mostly been driven by expert opinion, in turn motivated by data that suggest that screening with MRI may reduce cancer stage[18, 42].

While our approach only considers diagnostic accuracy, and not mortality reduction, our manuscript provides a first step. The next step is to use our approach to choose a risk cutoff, and then use simulation-based models that predict mortality reduction and the risk of false-positive screens to evaluate the effects on mortality, morbidity, and cumulative false-positive rate.

Although this paper applied our method to breast cancer screening with mammography or MRI, the methods could also be applied to evaluate the utility of other modalities in breast cancer screening. In addition, randomized controlled clinical trials have demonstrated that screening yields mortality reduction in colon [5], lung [1] and oral [30] cancer. The methods of this manuscript could be used to find appropriate risk cutoffs to optimize stratified screening strategies for these other disease sites. In fact, the approach has the potential to perform better at other sites since other cancers may have single, and very strong risk factors, such as the odds ratio of 33.6 for oropharyngeal cancer among those who are HPV-16 L1 seropositive, nonsmokers, and non-drinkers [11].

The methods presented in this paper fulfill Dr. Brawley's call for transparent processes for developing risk cutoffs for cancer screening [9]. Instead of using expert opinion to choose a cutoff, standards-setting bodies like the American Cancer Society could use this approach to optimize the diagnostic accuracy of a screening strategy.

**Acknowledgments** This manuscript was submitted to the Department of Biostatistics and Informatics in the Colorado School of Public Health, University of Colorado Denver, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biostatistics for JTB. Partial funding for DHG was provided by a generous grant from the Lundbeck Foundation, who provided a visiting professorship to the University of Copenhagen. The authors thank the BCSC investigators, participating mammography facilities, and radiologists who provided the relevant data for this study. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: http://breastscreening.cancer.gov/.

**Funding** Funding was provided by Lundbeckfonden, National Cancer Institute (Grant Nos. 5K07CA088811, 1R03CA136048-01A1), and National Institute of Dental and Craniofacial Research (Grant No. RC2DE020779).

## References


1. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM, Sicks JD (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365(5):395–409
2. Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, Wilson M, Howell A (2003) Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. J Med Genet 40(11):807–814
3. Antoniou AC, Pharoah PPD, Smith P, Easton DF (2004) The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br J Cancer 91(8):1580–1590. https://doi.org/10.1038/sj.bjc.6602175
4. Armstrong AC, Evans GD (2014) Management of women at high risk of breast cancer. BMJ 348(apr28 26):g2756–g2756. https://doi.org/10.1136/bmj.g2756
5. Atkin WS, Edwards R, Kralj-Hans I, Wooldrage K, Hart AR, Northover JMA, Parkin DM, Wardle J, Duffy SW, Cuzick J, UK Flexible Sigmoidoscopy Trial Investigators (2010) Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. Lancet (London, England) 375(9726):1624–1633. https://doi.org/10.1016/S0140-6736(10)60551-X
6. Baker SG (2000) Identifying combinations of cancer markers for further study as triggers of early intervention. Biometrics 56(4):1082–1087
7. Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, Tice JA, Buist DSM, Geller BM, Rosenberg R, Yankaskas BC, Kerlikowske K (2006) Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst 98(17):1204–1214. https://doi.org/10.1093/jnci/djj331
8. Berry DA, Iversen ES Jr, Gudbjartsson DF, Hiller EH, Garber JE, Peshkin BN, Lerman C, Watson P, Lynch HT, Hilsenbeck SG, Rubinstein WS, Hughes KS, Parmigiani G (2002) BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. J Clin Oncol 20(11):2701–2712
9. Brawley O, Byers T, Chen A, Pignone M, Ransohoff D, Schenk M, Smith R, Sox H, Thorson AG, Wender R (2011) New American Cancer Society process for creating trustworthy cancer screening guidelines. J Am Med Assoc 306(22):2495–2499
10. Claus E (2000) Risk models in genetic epidemiology. Stat Methods Med Res 9(6):589–601
11. D'Souza G, Pawlita M, Westra WH (2007) Case–control study of human papillomavirus and oropharyngeal cancer. N Engl J Med 356(19):1944–1956
12. Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW (1998) Ten-year risk of false positive screening mammograms and clinical breast examinations. N Engl J Med 338(16):1089–1096
13. Gail MH, Pfeiffer RM (2005) On criteria for evaluating models of absolute risk. Biostatistics 6(2):227–239. https://doi.org/10.1093/biostatistics/kxi005
14. Hagen AI, Kvistad KA, Maehle L, Holmen MM, Aase H, Styr B, Vabø A, Apold J, Skaane P, Møller P (2007) Sensitivity of MRI versus conventional screening in the diagnosis of BRCA-associated breast cancer in a national prospective series. Breast (Edinburgh, Scotland) 16(4):367–374. https://doi.org/10.1016/j.breast.2007.01.006
15. Hartman AR, Daniel BL, Kurian AW, Mills MA, Nowels KW, Dirbas FM, Kingham KE, Chun NM, Herfkens RJ, Ford JM, Plevritis SK (2004) Breast magnetic resonance image screening and ductal lavage in women at high genetic risk for breast carcinoma. Cancer 100(3):479–489. https://doi.org/10.1002/cncr.11926
16. Hendrick RE, Smith RA, Rutledge JH, Smart CR (1997) Benefit of screening mammography in women aged 40–49: a new meta-analysis of randomized controlled trials. J Natl Cancer Inst 22:87–92
17. Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley series in probability and statistics. Wiley, New York
18. Kriege M, Brekelmans CTM, Boetes C, Besnard PE, Zonderland HM, Obdeijn IM, Manoliu RA, Kok T, Peterse H, Tilanus-Linthorst MMA, Muller SH, Meijer S, Oosterwijk JC, Beex LVAM, Tollenaar RAEM, de Koning HJ, Rutgers EJT, Klijn JGM (2004) Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. N Engl J Med 351(5):427–437
19. Kuhl CK, Schrading S, Leutner CC, Morakkabati-Spitz N, Wardelmann E, Fimmers R, Kuhn W, Schild HH (2005) Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer. J Clin Oncol 23(33):8469–8476
20. Leach MO, Boggis CRM, Dixon AK, Easton DF, Eeles RA, Evans DGR, Gilbert FJ, Griebsch I, Hoff RJC, Kessar P, Lakhani SR, Moss SM, Nerurkar A, Padhani AR, Pointon LJ, Thompson D, Warren RML (2005) Screening with magnetic resonance imaging and mammography of a UK population at high familial risk of breast cancer. Lancet 365(9473):1769–1778
21. Lehman CD (2012) Diffusion weighted imaging (DWI) of the breast: ready for clinical practice. Eur J Radiol 81(Suppl 1):S80–81. https://doi.org/10.1016/S0720-048X(12)70032-3
22. Lehman CD, Blume JD, Weatherall P, Thickman D, Hylton N, Warner E, Pisano E, Schnitt SJ, Gatsonis C, Schnall M, DeAngelis GA, Stomper P, Rosen EL, O'Loughlin M, Harms S, Bluemke DA (2005) Screening women at high risk for breast cancer with mammography and magnetic resonance imaging. Cancer 103(9):1898–1905
23. Ma H, Bandos AI, Gur D (2015) On the use of partial area under the ROC curve for comparison of two diagnostic tests. Biom J 57(2):304–320. https://doi.org/10.1002/bimj.201400023
24. Ma H, Bandos AI, Rockette HE, Gur D (2013) On use of partial area under the ROC curve for evaluation of diagnostic performance. Stat Med 32(20):3449–3458. https://doi.org/10.1002/sim.5777
25. McFarland EG, Levin B, Lieberman DA, Pickhardt PJ, Johnson CD, Glick SN, Brooks D, Smith RA (2008) Revised colorectal screening guidelines: joint effort of the American Cancer Society, U.S. Multisociety Task Force on Colorectal Cancer, and American College of Radiology. Radiology 248(3):717–720
26. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P (2004) Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 159(9):882–890. https://doi.org/10.1093/aje/kwh101
27. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LL, Bassett L, D'Orsi C, Jong R, Rebner M (2005) Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med 353(17):1773–1783
28. Quante AS, Whittemore AS, Shriver T, Hopper JL, Strauch K, Terry MB (2015) Practical problems with clinical guidelines for breast cancer prevention based on remaining lifetime risk. J Natl Cancer Inst 107(7):djv124. https://doi.org/10.1093/jnci/djv124
29. Ross S (1984) First course in probability, 2nd edn. Macmillan Publishing Company, New York
30. Sankaranarayanan R, Ramadas K, Thomas G, Muwonge R, Thara S, Mathew B, Rajan B (2005) Effect of screening on oral cancer mortality in Kerala, India: a cluster-randomised controlled trial. Lancet 365(9475):1927–1933

31. Sardanelli F, Aase HS, Álvarez M, Azavedo E, Baarslag HJ, Balleyguier C, Baltzer PA, Beslagic V, Bick U, Bogdanovic-Stojanovic D, Briediene R, Brkljacic B, Camps Herrero J, Colin C, Cornford E, Danes J, de Geer G, Esen G, Evans A, Fuchsjaeger MH, Gilbert FJ, Graf O, Hargaden G, Helbich TH, Heywang-Köbrunner SH, Ivanov V, Jónsson Á, Kuhl CK, Lisencu EC, Luczynska E, Mann RM, Marques JC, Martincich L, Mortier M, Müller-Schimpfle M, Ormandi K, Panizza P, Pediconi F, Pijnappel RM, Pinker K, Rissanen T, Rotaru N, Saguatti G, Sella T, Slobodníková J, Talk M, Taourel P, Trimboli RM, Vejborg I, Vourtsis A, Forrai G (2017) Position paper on screening for breast cancer by the European Society of Breast Imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey. Eur Radiol 27(7):2737–2743. https://doi.org/10.1007/s00330-016-4612-z
32. Sardanelli F, Podo F, Santoro F, Manoukian S, Bergonzi S, Trecate G, Vergnaghi D, Federico M, Cortesi L, Corcione S, Morassut S, Di Maggio C, Cilotti A, Martincich L, Calabrese M, Zuiani C, Preda L, Bonanni B, Carbonaro LA, Contegiacomo A, Panizza P, Di Cesare E, Savarese A, Crecco M, Turchetti D, Tonutti M, Belli P, Maschio AD (2011) High breast cancer risk Italian 1 (HIBCRIT-1) Study: multicenter surveillance of women at high genetic breast cancer risk using mammography, ultrasonography, and contrast-enhanced magnetic resonance imaging (the high breast cancer risk italian 1 study): final results. Investig Radiol 46(2):94–105. https://doi.org/10.1097/RLI.0b013e3181f3fcdf
33. Saslow D, Boetes C, Burke W, Harms S, Leach MO, Lehman CD, Morris E, Pisano E, Schnall M, Sener S, Smith RA, Warner E, Yaffe M, Andrews KS, Russell CA (2007) American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography. CA 57(2):75–89
34. Sickles EA, D'Orsi CJ, Bassett LW (2013) ACR BI-RADS—mammography
35. Smith RA, Andrews K, Brooks D, DeSantis CE, Fedewa SA, Lortet-Tieulent J, Manassaram-Baptiste D, Brawley OW, Wender RC (2016) Cancer screening in the United States, 2016: a review of current American Cancer Society guidelines and current issues in cancer screening. CA 66(2):96–114. https://doi.org/10.3322/caac.21336
36. Smith RA, Cokkinides V, Brawley OW (2012) Cancer screening in the United States, 2012: a review of current American Cancer Society guidelines and current issues in cancer screening. CA 62(2):129–142
37. Thisted RA (1988) Elements of statistical computing: numerical computation, 1st edn. Chapman and Hall/CRC, Boca Raton
38. Tosteson ANA, Stout NK, Fryback DG, Acharyya S, Herman BA, Hannah LG, Pisano ED, DMIST Investigators (2008) Cost-effectiveness of digital mammography breast cancer screening. Ann Intern Med 148(1):1–10
39. Tyrer J, Duffy SW, Cuzick J (2004) A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 23(7):1111–1130
40. Wald NJ, Hackshaw AK, Frost CD (1999) When can a risk factor be used as a worthwhile screening test? BMJ 319(7224):1562–1565. https://doi.org/10.1136/bmj.319.7224.1562
41. Wang Z, Luo X, Chang YcI (2015) Assessing the predictive power of newly added biomarkers. Biom J 57(5):797–807. https://doi.org/10.1002/bimj.201400210
42. Warner E (2008) The role of magnetic resonance imaging in screening women at high risk of breast cancer. Top Magn Reson Imaging 19(3):163–169. https://doi.org/10.1097/RMR.0b013e31818bc994
43. Warner E, Plewes DB, Hill KA, Causer PA, Zubovits JT, Jong RA, Cutrara MR, DeBoer G, Yaffe MJ, Messner SJ, Meschino WS, Piron CA, Narod SA (2004) Surveillance of BRCA1 and BRCA2 mutation carriers with magnetic resonance imaging, ultrasound, mammography, and clinical breast examination. JAMA 292(11):1317–1325. https://doi.org/10.1001/jama.292.11.1317
44. Yabuuchi H, Matsuo Y, Sunami S, Kamitani T, Kawanami S, Setoguchi T, Sakai S, Hatakenaka M, Kubo M, Tokunaga E, Yamamoto H, Honda H (2011) Detection of non-palpable breast cancer in asymptomatic women by using unenhanced diffusion-weighted and T2-weighted MR imaging: comparison with mammography and dynamic contrast-enhanced MR imaging. Eur Radiol 21(1):11–17. https://doi.org/10.1007/s00330-010-1890-8