Published in final edited form as:
Stat Methods Med Res. 2010 October ; 19(5): 429–449. doi:10.1177/0962280209359842.

Modeling the cumulative risk of a false-positive screening test

Rebecca A. Hubbard[1], **Diana L. Miglioretti**[1], and **Robert A. Smith**[2]

[1] Group Health Research Institute, Biostatistics Unit and Department of Biostatistics, University of Washington, Seattle, WA

[2] American Cancer Society, Atlanta, GA

Abstract

The goal of a screening test is to reduce morbidity and mortality through the early detection of disease; but the benefits of screening must be weighed against potential harms, such as false-positive (FP) results, which may lead to increased healthcare costs, patient anxiety, and other adverse outcomes associated with diagnostic follow-up procedures. Accurate estimation of the cumulative risk of a FP test after multiple screening rounds is important for program evaluation and goal setting, as well as informing individuals undergoing screening what they should expect from testing over time. Estimation of the cumulative FP risk is complicated by the existence of censoring and possible dependence of the censoring time on the event history. Current statistical methods for estimating the cumulative FP risk from censored data follow two distinct approaches, either conditioning on the number of screening tests observed or marginalizing over this random variable. We review these current methods, identify their limitations and possibly unrealistic assumptions, and propose simple extensions to address some of these limitations. We discuss areas where additional extensions may be useful. We illustrate methods for estimating the cumulative FP recall risk of screening mammography and investigate the appropriateness of modeling assumptions using 13 years of data collected by the Breast Cancer Surveillance Consortium. In the BCSC data we found evidence of violations of modeling assumptions of both classes of statistical methods. The estimated risk of a FP recall after 10 screening mammograms varied between 58% and 77% depending on the approach used, with an estimate of 63% based on what we feel are the most reasonable modeling assumptions.

Keywords

cumulative false-positive rate; dependent censoring; discrete hazard rate; mammography; screening evaluation

1 Introduction

The goal of a screening test is to reduce morbidity and mortality by detecting disease among asymptomatic individuals early, when treatment may be most successful[1]. Some screening tests may also lead to the prevention of disease through the removal of precursor lesions or the treatment of conditions that cause disease. However, apart from this principal goal, screening programs also need to consider the balance of benefits and harms. Harms may result directly from the screening test itself, such as radiation exposure from an imaging test, or may result from the diagnostic workup of a positive screening test. Many screening tests are designed to be more sensitive than specific, with true-positive rates that are a small percentage of the total positive rate. Those who test positive usually will undergo more specific, though possibly more

Correspondence: Rebecca A. Hubbard, PhD, Group Health Research Institute, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101, Phone: (206) 287-2066, Fax: (206) 287-2017, hubbard.r@ghc.org.

expensive and invasive, diagnostic tests. Thus, it is important to understand and quantify the risk of false-positive (FP) results, which are associated with increased healthcare costs, patient anxiety, and other adverse outcomes arising from diagnostic follow-up procedures among patients without disease.

In breast cancer screening, the benefits of mammography are well established and include reducing the risks of being diagnosed with an advanced breast cancer and of dying from breast cancer and increasing the range of treatment options such as breast-conserving therapy[2,3]. Harms associated with screening mammography include recall for either immediate or short-interval follow-up, biopsy of benign lesions, and the anxiety associated with these diagnostic workups[4,5]. FP recalls–i.e., recall for the additional workup of a screening mammogram among women without breast cancer–are the most prevalent harm. FPs are unavoidable if mammography screening programs are to succeed at detecting small breast cancers, but it is important that the rate not be greater than necessary to achieve that purpose. Women should be informed about the inevitability of FPs and their risk of experiencing a FP mammogram if they undergo regular screening. In fact, it is a worthy goal to tailor information about the likelihood of experiencing a FP based on risk factors associated with the probability of a FP result.

Estimating the cumulative risk of a FP result is challenging for several reasons. First, there may be significant differences within and between programs in the FP rate, and these differences may be influenced by the screening population, interpretive skill, and resources. Second, because individuals may not receive all recommended screening rounds within the study period. This may be due to: administrative censoring, if the study ends before all individuals receive all exams; non-adherence with recommended screening intervals; or individuals dropping out of the screening program. Differences in the FP risk between subjects who choose to attend more screening rounds and those who drop out represents a form of dependent censoring. In this case, estimates conditional on the number of screening rounds attended will not reflect the FP rate that would be observed if the entire population adhered to the recommended screening regimen.

Previous research has estimated the cumulative risk of a FP screening mammogram result from right censored observations under the assumption that the probability of participating in subsequent screening rounds is independent of prior exam results[6,7,8]; however, evidence suggests this assumption may not hold for screening mammography[9,10]. Extensions by Xu et al.[10] for estimating the FP risk while allowing for differences among those who attend different numbers of screening rounds rely heavily on parametric assumptions about the distribution of the time to the first FP result for censored individuals that may not be appropriate in the context of screening mammography.

In this paper, we review existing statistical methods used to estimate the cumulative risk of a FP result. We discuss limitations and unrealistic assumptions of current approaches and propose extensions. We compare inference on the cumulative FP recall rate of screening mammography based on these statistical methods and investigate the appropriateness of modeling assumptions using 13 years of data collected by the Breast Cancer Surveillance Consortium (BCSC). We end with a discussion of our findings and suggested areas for future research.

2 Review of existing statistical methods

2.1 Definitions and notation

Let n denote the number of subjects under observation and S_i denote the number of screening tests received by the ith subject. We denote the screening round of the first FP as W_i. Let T_{ij} =

1 be an indicator of receiving a jth screen and $Y_{ij} = 1$ indicate a FP at that screen for the ith subject. Thus, if $S_i = k$ then $T_{ij} = 1$ for j less than or equal to k and $T_{ij} = 0$ for all j greater than k. We additionally denote the complete screening history for the ith subject up to the jth test as $\mathbf{Y}_{ij} = (Y_{i1}, \ldots, Y_{ij})$. For ease of notation we will suppress the subscript i relating to subject throughout.

We are interested in estimating the cumulative probability of an individual receiving a FP after a specified number of screening rounds. This cumulative probability can be expressed in terms of W as $p_j = P(W \leq j)$ or in terms of Y as $p_j = 1 - P(Y_1 = 0, \ldots, Y_j = 0)$; that is, the probability that after undergoing j tests, the subject has received at least one FP test result. In addition to the overall cumulative FP probability, we may also be interested in understanding how individual characteristics influence this probability. In the discussion below, we address estimation of both of these quantities.

If all subjects received the recommended regimen of screening tests, that is, if $T_{ij} = 1$ for all i and j, it would be straightforward to estimate the cumulative FP probability. For instance, one could use the empirical distribution of W as an estimator. However, in practice not all subjects will attend all recommended screening rounds. This results in right censoring of W. That is, the time of the first FP will be unobserved for some subjects. If the censoring mechanism is not independent of W then naive estimates of p_j will be biased relative to the true probability of a FP test result under conditions of regular screening. Below we discuss two distinct approaches to estimating p_j when W is right censored.

2.2 Conditional estimation of the false-positive probability

One approach to estimating p_j under right censoring of the data developed by Gelfand and Wang[7] is to define the cumulative risk of a FP screening test conditional on attending j screening rounds. That is,

$$p_j^* = P(W \leq j | S \geq j) = 1 - P(Y_1 = 0, \ldots, Y_j = 0 | T_1 = 1, \ldots, T_j = 1). \quad (1)$$

They additionally define the probability of receiving a first FP at the jth screening round conditional on participating in at least j screens as

$$q_j = P(Y_j = 1 | Y_1 = 0, \ldots, Y_{j-1} = 0, T_1 = 1, \ldots, T_j = 1). \quad (2)$$

If we assume that $q_j = P(Y_j = 1 | Y_1 = 0, \ldots, Y_{j-1} = 0)$, that is the probability of receiving a first FP is independent of the number of rounds of screening a subject participates in, then we can express the cumulative FP probability in terms of q_j as

$$p_j = 1 - \prod_{k=1}^{j} (1 - q_k). \quad (3)$$

As noted by Xu et al.[10], a necessary and sufficient condition for equation (3) to hold is that T_{j+1} is conditionally independent of $I(\mathbf{Y}_j = \mathbf{0}_j)$ given $S \geq j$. Xu et al.[10] suggest evaluating this assumption using a hypothesis test for independence of $I(\mathbf{Y}_j = \mathbf{0}_j)$ and $I(S = j)$ among those attending at least j rounds of screening. Independence of these two binary random variables can be tested using Pearson's χ^2-test or Fisher's exact test for each value of j.

If the independence assumption holds, it is straightforward to develop the likelihood for q_j. Let s_j denote the number of subjects participating in at least j screening rounds with no FPs in the first j screens and r_j denote the number of subjects participating in at least j screens with no FPs in the first $j-1$ screens. Note this implies that $r_j - s_j$ subjects have a FP at the jth screen. It also implies that the number of subjects at risk for a FP screen decreases by $s_j - r_{j-1}$ between the $(j-1)$th and jth tests. We can now write the likelihood for q_j as

$$L(q_1, \ldots, q_K) = \prod_{j=1}^{K} q_j^{r_j - s_j} (1 - q_j)^{s_j}, \tag{4}$$

where K is the largest number of screening tests observed for any subject.

Estimation of p_j^* has been investigated via both maximum likelihood (ML) and Bayesian approaches. As pointed out by Gelfand and Wang[7], the ML estimate is simply the actuarial estimator with

$$\widehat{q}_j = 1 - \frac{s_j}{r_j} \text{ and } \widehat{p}_j^* = 1 - \prod_{k=1}^{j} (1 - \widehat{q}_k),$$

with variance given by Greenwood's formula,

$$\mathrm{Var}(\widehat{p}_j^*) = (1 - \widehat{p}_j^*)^2 \sum_{k=1}^{j} \frac{r_k - s_k}{r_k s_k}.$$

Bayesian estimation is similarly straightforward for the conditional model. In this case, we can either assign prior distributions to q_j or, if prior information is not available on the screen-specific FP rates, a hierarchical approach in which q_j are assumed to arise from some common distribution with unknown parameters can be adopted. The latter was explored by Gelfand and Wang[7] who suggest using $q_j \sim \mathrm{Beta}(a, b)$, as this is the conjugate density in this case with independent gamma hyperpriors on a and b. Estimation can be carried out by using Markov Chain Monte Carlo (MCMC) via the Gibbs sampler to sample from the posterior distribution [11]. In the context of screening mammography, Elmore et al.[6] used this method to estimate the cumulative FP risk.

2.2.1 Covariate effects in the conditional model—The conditional model can be extended to allow for estimation of possibly time-varying covariate effects on the risk of a FP test result. To do so, we define $q(\boldsymbol{X}_j) = P(Y_j = 1 | Y_1 = 0, \ldots, Y_{j-1} = 0, \boldsymbol{X}_j)$, the probability of a first FP at screen j for a subject with covariate vector $\boldsymbol{X}_j$ and specify a regression equation expressing $q(\boldsymbol{X}_j)$ as a function of covariates and regression parameters.

ML estimation for the conditional model with covariates can be undertaken using standard software for binary regression. For example, Baker et al.[12] propose a logistic regression model for $q(\boldsymbol{X}_j)$. In the frequentist framework, existing research has only addressed fixed effects regression models for $q(\boldsymbol{X}_j)$; it would be straightforward to incorporate random effects via the methods of generalized linear mixed models.

A Bayesian estimation method for $q(\boldsymbol{X}_j)$ was proposed by Gelfand and Wang[7]. They suggest using

$$1 - q(\boldsymbol{X}_j) = (1 - q_j)^{\exp(\boldsymbol{X}_j\beta)}$$

because this form can be thought of as a discrete time analogue to the proportional hazards model with hazard function $h(\boldsymbol{X}_j) = h_j \exp(\boldsymbol{X}_j\beta)$ and $h_j = -\log(1 - q_j)$ interpreted as discrete baseline hazards. We can express the likelihood in terms of $q(\boldsymbol{X}_j)$ as

$$L(q_1, \ldots, q_k) = \prod_{j=1}^{k} (q(\boldsymbol{X}_j))^{y_j} (1 - q(\boldsymbol{X}_j))^{1-y_j}.$$

Gelfand and Wang[7] note that this is a Bernoulli likelihood and hence for $q(\boldsymbol{X}_j)$ close to zero can be reasonably approximated by a Poisson distribution with $P(Y_j = 0) = 1 - q(\boldsymbol{X}_j) = \exp(-\lambda_j)$, where $\lambda_j = h_j \exp(\boldsymbol{X}_j\beta)$. Estimates for $q(\boldsymbol{X}_j)$ can be used to obtain the covariate specific cumulative risk,

$$p_j^*(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j) = 1 - \prod_{k=1}^{j} (1 - q(\boldsymbol{X}_k)).$$

Using either the Bernoulli likelihood or the Poisson approximation to the likelihood, Bayesian estimation can be carried out by assigning prior distributions to h_j and β. Gelfand and Wang [7] propose a hierarchal approach with $h_j \sim \text{Gamma}(a, b)$. In this model we can either assume a and b known based on prior information or assign hyperpriors for a and b. Random effects can easily be incorporated into the Bayesian framework by introducing subject- or cluster-specific covariate effects. In the context of screening mammography, this Bayesian estimation method with random effects used to capture between-radiologist variability in FP rates was implemented by Christiansen et al.[8].

2.2.2 Summary of the conditional model—The conditional model addresses difficulties in estimating the cumulative FP probability under right censoring by assuming independence of the number of screening rounds a subject attends and the probability of a FP test result. A straightforward test exists that allows us to evaluate whether this assumption is appropriate in practice. In Section 4.3 we evaluate the independence assumption for the BCSC population.

A conceptual limitation of the conditional model is that $\widehat{p_j^*}$ is an estimate of the probability of a FP screening test result for subjects who choose to participate in at least j tests during the study period. This may not correspond to the FP risk in subjects who do not comply with screening recommendations or receive fewer than j tests for other reasons. Ideally, we would like to estimate the unconditional FP probability, p_j, that is, the cumulative FP rate for all subjects, regardless of their observed screening behavior. In Section 2.3, we discuss approaches to estimating p_j by marginalizing over the number of screening rounds attended.

2.3 Marginal estimation of the false-positive probability

Conceptually, it is desirable to estimate p_j, the unconditional cumulative probability of a FP. This can be achieved in the presence of censoring by marginalizing over the distribution of number of screening rounds attended, S. In this approach we define

$$p_j = P(W \leq j) = \sum_{k=1}^{j} \xi_k, \tag{5}$$

where $\xi_k = P(W = k)$.

Consider the data available for estimating p_j to be (S, Z, δ) where $Z = \min(W, S)$ and δ is an indicator of $W < S$. These data are similar to the data available in a typical survival context with right censored data. However, in the case of screening tests, in addition to observing Z we also observe S regardless of whether or not the event (a FP) precedes the censoring time. This additional information, which is not available in the typical context of right censored data, allows us to relax the assumption of independence of S and W.

To estimate p_j without assuming independence of S and W we must make some assumption about the relationship between the time of the first FP and the number of screening rounds attended for subjects who have not received a FP prior to censoring. Xu et al.[10] propose the model

$$P(W = j | S = k) = \theta_k (1 - \theta_k)^{j-1}, \tag{6}$$

when $j > k$. That is, following censoring, the time to the first FP screen is geometrically distributed with parameter $\theta_k = P(Y_j = 1 | Y_1 = 0, \ldots, Y_{j-1} = 0, S = k)$ dependent on the number of screening rounds attended. Note that by making this assumption we facilitate estimation of the risk of a first FP test for screens subsequent to the maximum number of screens observed for an individual. However, while we are now allowing for variation in the FP rate for subjects who participate in different total numbers of screens, we have introduced a new assumption. We now assume that, given the total number of screening rounds attended, FP results occur at a constant rate across screening rounds following censoring. At the least, for screening mammography we expect the FP rate to be higher for the first versus subsequent rounds of screening[13]. We discuss an extension of this method making use of an assumption that may be more appropriate in the context of screening mammography in Section 3.2.

Let n_{kj} denote the number of subjects participating in k screening rounds with a first FP at the jth screen and n_{k0} denote the total number of subjects with k screens and no FP. Based on the above assumption of geometrically distributed time to first FP given the number of screens received for subjects censored prior to a FP, we can construct the likelihood arising from the data as

$$\begin{aligned} L &= \prod_{k=1}^{K} \left\{ (P(S = k))^{\sum_{j=0}^{k} n_{kj}} \left(\prod_{j=1}^{k} (P(W = j | S = k))^{n_{kj}} \right) (P(W > k | S = k))^{n_{k0}} \right\} \\ &= \prod_{k=1}^{K} (P(S = k))^{\sum_{j=0}^{k} n_{kj}} \prod_{k=1}^{K} \left\{ \left(\prod_{j=1}^{k} (P(W = j | S = k))^{n_{kj}} \right) (1 - \theta_k)^{k n_{k0}} \right\}. \end{aligned} \tag{7}$$

Note that if parameters of $P(S = k)$ are distinct from the parameters of $P(W = j|S = k)$ and θ_k, we can carry out separate estimation for these two portions of the likelihood.

Xu et al.[10] derived maximum likelihood estimators for this model. Under the maximum likelihood approach they obtained estimates for the length of screening

$$\widehat{P}(S=k)=\frac{\sum_{j=0}^{k} n_{kj}}{\sum_{k=1}^{K}\sum_{j=0}^{k} n_{kj}},$$

and for the probability of a FP conditional on the total length of screening

$$\widehat{P}(W=j|S=k)=\frac{n_{kj}}{\sum_{l=0}^{k} n_{kl}}$$

and

$$\widehat{\theta}_k=1-\left(\frac{n_{k0}}{\sum_{l=0}^{k} n_{kl}}\right)^{1/k}.$$

We can then go about estimating p_j via $\widehat{p}_j=\sum_{k=1}^{j}\widehat{\xi}_k$, where

$$\widehat{\xi}_k=\sum_{l=1}^{K}\widehat{P}(S=l)\widehat{P}^{*}(W=k|S=l),$$

and $\hat{P}^*(W = k|S = l) = \hat{P}(W = k|S = l)$ when $k \leq l$ and $\hat{P}^*(W = k|S = l) = \hat{\theta}_l(1 - \hat{\theta}_l)^{k-1}$ when $k > l$. An important advantage of this approach is that it allows for estimation of $\hat{p}_k$ even when k exceeds the maximum number of screens observed in the data.

Xu et al.[10] describe a frequentist estimation approach; however, Bayesian methods can also be used to obtain estimates of p_j using the likelihood in equation (7) by assigning prior distributions to $P(S = k)$, $P(W = j|S = k)$, and θ_k. We can then obtain the posterior distribution for these quantities. If MCMC sampling were used it would be straightforward to obtain a sample from the posterior for p_j using the relationships presented above.

2.3.1 Covariate effects in the marginal model—Xu et al.[10] proposed an approach to incorporating covariates into estimates of the marginal probability of a FP screening test by defining the outcome vector $\{I(W = 1), I(W = 2), \ldots, I(W = k), I(W > k)\}$. Conditional on k this will be multinomially distributed. We can then specify a regression equation relating covariates $\boldsymbol{X}_j$ to $P(W = j|S = k, \boldsymbol{X}_j)$, for instance via a logistic formulation. A separate regression equation is estimated for each value of S. We can similarly obtain estimates of $P(S = k|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$ by using a multinomial likelihood for $\{I(S = 1), \ldots, I(S = K)\}$. Standard multinomial logistic regression software can be used to estimate the regression parameters in this likelihood and estimates of $P(W = j|S = k, \boldsymbol{X}_j)$ and $P(S = k|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$ can be used to estimate $\xi(\boldsymbol{X})$. Bayesian

estimation can also be used to obtain risk factor parameter estimates making use of the multinomial likelihood conditional on number of screens received.

2.3.2 Summary of the marginal model—The marginal model provides a framework for estimating the cumulative probability of a FP screening test result without requiring that we condition on the number of screening rounds a subject attended. This is desirable in the context of evaluating the performance of a recommended screening program because it provides an indication of the FP risks introduced by the program if all subjects complied with recommendations. We contrast this to the conditional model which provides an estimate of the FP risk only for the sub-group who choose to comply with screening recommendations. Because this sub-group may not be representative of the population at large, the marginal model may be preferred in evaluating screening recommendations.

Existing methods for marginal modeling of the cumulative FP probability have several notable limitations. Specifically, these methods have assumed a constant FP rate across screening rounds for censored subjects. This is particularly problematic for subjects participating in only a single screening round because information from this single test will be used to project the FP risk for these subjects in later rounds of screening. This will provide a poor estimate if the FP risk at the first test differs from that expected at later screening rounds. We demonstrate the extent of this problem in the context of screening mammography in Section 4.3 using data from the BCSC. An extension of this method to accommodate screening round dependent variation of the FP rate is needed.

3 Extensions

3.1 Covariate adjusted tests of the independence assumption

As discussed above, the primary limitation of the conditional method is its reliance on the assumption of independence of the number of screening rounds attended and the history of FP results. The test proposed by Xu et al.[10] allows us to formally evaluate this assumption. However, this test does not account for violations of the assumption that may be mediated by conditioning on covariates.

Violations of the independence assumption could arise for several reasons. Differences in the number of screening rounds attended could be caused by prior screening results. For instance, subjects receiving a FP may be less likely to return for additional screening. This would be a direct violation of the independence assumption. Alternatively, differences may be attributable to confounders. For instance, in the context of screening mammography, women at higher risk might be more likely to return for additional screening and might have a different probability of a FP result than lower risk women. Another possibility is that differences in the cumulative FP risk could be due to different screening intervals. Individuals who are screened more frequently will have more observed screening exams during the study period by definition, and previous research for screening mammography has found that the FP probability decreases as the screening interval decreases[13]. Under the latter two scenarios, one way to mediate violations of the independence assumption would be to condition the FP probability estimates on covariates. If the history of FP test results among women who continue to attend screening versus those with fewer observed screening tests is the same after conditioning on covariates, the method of estimation for the conditional cumulative probability estimate described above would remain valid by basing estimates on covariate adjusted $\hat{q}(\boldsymbol{X})$.

Tests of the independence assumption can be extended to adjust for possible confounding by covariates. We propose a test of the independence assumption after adjusting for covariates by fitting a logistic regression model to data for all subjects attending at least k screening rounds for each value of k, the number of screens attended, with outcome $I(W > k)$ and predictors I

($S > k$) and possible confounders. If the regression parameter associated with $I(S > k)$ is significantly different from zero, this would indicate a violation of the independence assumption after accounting for confounding variables. However, if the independence assumption is satisfied, then estimates of $p_j^*(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_j)$ will provide valid inference about the cumulative FP rate.

In addition to carrying out hypothesis testing of the independence assumption, we should also evaluate the clinical importance of variations in the FP probability as a function of number of screens attended after adjusting for confounding by covariates. Covariate adjusted estimates of the probability of a FP conditional on the number of screens obtained and covariates are available via logistic regression models using marginal standardization, also known as predictive margins[14,15]. This calculation entails first estimating the probability of a FP for each combination of covariates based on the fitted logistic regression model. We then combine these estimates weighted by the overall proportion of subjects falling in each stratum. Standard errors for these probability estimates are available via the delta method.

3.2 More flexible marginal models

The marginal model is appealing because it allows us to estimate the FP probability associated with a recommended screening program. However, the model proposed by Xu et al.[10] makes strong assumptions about the probability of a future FP screen for subjects with no FP results prior to the end of screening. This assumption could be relaxed by allowing FP probabilities to vary as a function of screening round as well as total number of screening rounds attended. That is, rather than assume θ_k constant across screening rounds, we specify a screening round-dependent function for $P(Y_j = 1|Y_1 = 0, \ldots, Y_{j-1} = 0, S = k)$, $f_j(\alpha_k)$. This implies that the probability of censoring prior to receiving a FP is

$$P(W>k|S=k)=\prod_{j=1}^{k}(1-f_j(\alpha_k)).$$

The marginal likelihood given by equation (7) holds for this model. Depending on the functional form of $f_j(\alpha_k)$, explicit estimates for $\hat{\alpha_k}$ may be available or estimates can be obtained using numerical maximization of the likelihood.

In the context of screening mammography, we expect the FP probability at the first screening test to be highest followed by a decrease in the FP probability at each subsequent screen. We propose to model this using a linear function on the logistic scale in which FP probability at the first screen is estimated separately dependent on the total number of screening rounds attended followed by a change in the FP probability for subsequent screens and a linear trend thereafter associated with repeated screening. Specifically, we define

$$f_j(\alpha_k)=\frac{\exp(\alpha_{k0}+\beta_0 I(j>1)+\beta_1(j-1))}{1+\exp(\alpha_{k0}+\beta_0 I(j>1)+\beta_1(j-1))}, \quad (8)$$

where $\exp(\alpha_{k0}/(1 + \exp(\alpha_{k0}))$ represents the FP probability at the first screening test for subjects participating in a total of k screening rounds and β_0 and β_1 represent variations in this probability associated with repeated screening which are assumed constant across k. This model could be further relaxed by allowing β_0 and β_1 to depend on k as well. However, in this case additional assumptions would be required for identifiability

3.3 Expanded methods for time-varying covariates

Methods for incorporating time-varying covariates in the marginal model could also be extended. Although time-varying covariates can be incorporated into the model of Xu et al. [10], this method is somewhat problematic. In their model, $P(W = j|S = k, \boldsymbol{X}_j)$ depends only on current covariate values. However, receiving a first FP result at the jth screen is dependent upon the results of the previous $j - 1$ screens. Because $P(W = j|S = k) = P(\boldsymbol{Y}_{j-1} = \boldsymbol{0}, Y_j = 1|S = k)$, it is clear that this joint probability may depend on a complex function of past and current covariate values. Alternatively, one could construct a model for $P(W = j|S = k)$ via the decomposition

$$P(W=j|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_j,S=k)=P(Y_1=0|\boldsymbol{X}_1,S=k)\left(\prod_{l=2}^{j-1}P(Y_l=0|\boldsymbol{Y}_{l-1}=\boldsymbol{0},\boldsymbol{X}_l,S=k)\right)P(Y_j=1|\boldsymbol{Y}_{j-1}=\boldsymbol{0},\boldsymbol{X}_j,S=k).$$

We can then construct a joint model for the binary event $P(Y_j = 1|\boldsymbol{Y}_{j-1} = \boldsymbol{0}, \boldsymbol{X}_j, S = k)$ for all j. Covariate adjustment of $P(Y_j = 1|\boldsymbol{Y}_{j-1} = \boldsymbol{0}, S = k)$ could be incorporated in $f_j(\alpha_k)$ to yield covariate adjusted FP risk estimates that also account for trends in the FP rate associated with variations in the FP probability across screening rounds.

4 Application to the BCSC

4.1 Description of the BCSC population

We illustrate the existing methodology as well as our proposed extensions using data collected by seven mammography registries in the National Cancer Institute-funded Breast Cancer Surveillance Consortium (BCSC) [16] (http://breastscreening.cancer.gov): (1) the Carolina Mammography Registry, (2) the Colorado Mammography Project, (3) Group Health Cooperative in Washington; (4) the New Hampshire Mammography Network, (5) the New Mexico Mammography Project, (6) the San Francisco Mammography Registry, and (7) the Vermont Breast Cancer Surveillance System. These registries link information on women who receive a mammogram at a participating facility to regional cancer registries and pathology databases to determine cancer outcomes. Information on patient and mammogram characteristics collected at the time of the mammogram included patient age, clinical history, breast cancer risk factors, and time since last mammogram.

We included women who had their first screening mammogram between the ages of 40 and 59 at a participating BCSC facility. We included this first screening mammogram along with subsequent screening mammograms meeting inclusion criteria performed from 1994 to the most recent year with complete breast cancer capture, which varied from 2004-2006 across the seven mammography registries. A screening mammogram was defined as a bilateral mammogram that the interpreting radiologist indicated was for routine screening. To avoid misclassifying diagnostic exams as screening exams, we excluded mammograms performed within 9 months of a prior breast imaging exam. We censored women if and when they were diagnosed with breast cancer; received breast augmentation or reconstruction; or self-reported a time since their last mammogram that differed from that in the database by more than six months, because women could go to a non-BCSC facility to receive a mammogram.

Mammograms were classified as positive or negative using standard BCSC definitions (see BCSC Glossary of Terms accessed at http://breastscreening.cancer.gov/data/bcsc_data_definitions.pdf) based on the initial Breast Imaging Reporting and Data Systems (BI-RADS) assessment[17] and recommendations assigned by the radiologist. As defined by the American College of Radiology BI-RADS Breast

Imaging Atlas[17], a positive mammogram was defined as a mammogram with initial BI-RADS assessment of 0, 4, or 5. To correct for inconsistencies in the use of the BI-RADS lexicon early in the study period, BI-RADS 3 assessments with a recommendation for immediate follow-up were reclassified as BI-RADS 0 and considered positive[18]. A positive mammogram was considered to be a FP if the woman was not diagnosed with invasive carcinoma or ductal carcinoma *in situ* within 1 year of the mammogram and prior to the next screening mammogram.

4.2 Summary of observed FP rates

We included 159,574 women who each received 1 to 13 screening mammograms over the study period for a total of 346,082 total mammograms and 303,761 mammograms that did not follow a prior FP result. Characteristics of this cohort are presented in Table 1. Half of the women received only a single mammogram while 5.4% received in excess of 5 mammograms. The majority of women attended screening with an average interval between screens of either 9 to 18 months (39.2%) or 19 - 30 months (37.7%). The interquartile range for the average interval between screens was 15 to 29 months.

Table 2 shows the probability of a first FP result for women with no prior FPs across screening rounds, $\hat{q_j} = \hat{P}(Y_j = 1 | \mathbf{Y}_{j-1} = \mathbf{0}, S \geq j)$, and the empirical cumulative probability of a FP result, $\hat{P}(W \leq j | S \geq j)$. Overall, 12.7% of the 346,082 screening mammograms were recalled for additional workup among women without breast cancer. The probability of a first FP was highest (16.2%) during the first screening round and decreased to between 5 and 10% at subsequent screens. The observed cumulative probability of at least one FP mammogram for women who had at least 10 screening mammograms was 44.5%. Interestingly, the observed cumulative probability of a FP is not monotonically increasing, suggesting the probability of returning for subsequent screening exams may depend on the prior test result.

Two modeling assumptions are central to the methods presented in Section 2: (1) under the conditional approach, the probability of a first FP at the *j*th screen is assumed independent of the total number of screening rounds a woman participates in; and (2) under the marginal approach we assume that women censored prior to a FP result would have experienced a constant FP risk at subsequent screening rounds. Assumption (1) may be violated if characteristics associated with participating in more frequent screening or screening over a longer time span are also associated with the FP probability. In screening mammography, assumption (2) is likely violated because prior research has shown that the FP rate at the first screening mammogram is higher than at subsequent mammograms[13]. This is especially problematic for estimation of the cumulative probability of a FP for women who attended only one round of screening because the FP probability at future screening rounds for these women will be estimated solely using the FP probability at the first round.

We investigate variation in the probability of receiving a FP across screening rounds and as a function of the total number of screening rounds attended for women in the BCSC (Figure 1). Empirical probabilities of a first FP (Figure 1, left) indicate strong time trends in FP risk associated with screening round. The probability of a FP is highest at the first screen, with highest rates among women who attended only one screen. The cumulative FP probability also varies as a function of total number of screening rounds a woman attended. Women attending more screening rounds had lower cumulative probabilities of a FP result then women attending fewer screening rounds (Figure 1, right). These results suggest that assumptions of both the marginal and conditional models are violated in the BCSC population.

4.3 Modeling the cumulative FP rate

Before undertaking model-based estimation of the cumulative FP probability, we formally evaluate the independence assumption for the conditional model. Figure 1 suggests that this assumption may not be appropriate for the BCSC data. We formally test the assumption of independence of the number of screening rounds attended and the history of screening exam results required by the conditional model using the test proposed by Xu et al.[10]. This test is equivalent to asking, among women who participated in at least k screening rounds ($S \geq k$), is a woman who has not yet had a FP result ($W > k$) equally as likely to return for a $(k + 1)$st exam ($S > k$) as a woman who has previously had a FP ($W \leq k$)? We carry out this test using a χ^2 test with one degree of freedom for each screening round (Table 3). In the BCSC data, this assumption is clearly violated for all screening rounds except for the eighth screen.

In Section 3.1, we proposed a covariate-adjusted test of the independence assumption. Covariate adjusted estimates of the FP probability for women who do versus those who do not attend additional screening are presented in Table 3 along with p-values corresponding to a hypothesis test for independence adjusted for baseline age less than 50, calendar year of first screen, average screening interval, and mammography registry. We modeled year of first screen as a categorical variable with separate categories for each year and categorized average interval between screens as shown in Table 1. After adjusting for these covariates, the independence assumption is still violated for all screening rounds except for the first.

In Figure 2 and Table 4 we compare marginal and conditional estimates of the cumulative probability of a FP screening exam in the BCSC population. The marginal curve is an estimate of the probability of receiving a first FP at or before the kth screening round while the conditional curve is an estimate of the probability of receiving a first FP at or before the kth screen, given that a woman chose to participate in at least k rounds of screening. The estimate of the cumulative probability of a FP at the tenth screening round based on the conditional model is 58.2%, while the estimate based on the marginal model is 77.0%. The substantial separation between estimates from the two models may be due to violations of model assumptions. As discussed above, the independence assumption of the conditional model does not hold for the BCSC population. We believe that a more appropriate estimate for these data would allow for variations in the FP probability associated with total number of screening rounds a woman attends. Moreover, the marginal method is also likely to be inappropriate for the BCSC population because it assumes a constant FP probability across screening rounds for women censored prior to their first FP, following censoring. This is an untestable assumption. However, empirical estimates of the probability of a first FP across screening rounds suggests this is likely to be false for these data (Figure 1). An estimate of the cumulative FP probability for screening mammography should allow for trends in FP probability across screening rounds.

To address these concerns, we implemented an adjusted marginal estimator using the proposed model for risk of a first FP given by equation (8). Point estimates were obtained via numerical maximization of the likelihood using a quasi-Newton method, and variances were computed using the delta method. The cumulative FP probability estimated for the BCSC lies in between the marginal and conditional estimates (Figure 2). The adjusted model estimates a cumulative FP probability at the tenth screening round of 63.3%. This estimate is intermediate between the conditional and marginal estimates because the adjusted estimate allows for variation in the FP probability associated with screening round and total number of screening rounds attended. It accommodates higher FP rates among women participating in only one screen, while also allowing for decreases in the FP probability across screening rounds. Both the marginal model proposed by Xu et al.[10] and our adjusted marginal model rely on untestable assumptions concerning the FP probability for women who are censored prior to their first FP. However, observed trends in FP rates (Figure 1) suggest that the adjusted marginal estimate is more appropriate for the BCSC population. A comparison of estimated cumulative FP

probabilities at the first, fifth, and tenth screening round based on the conditional, marginal, and adjusted marginal models is presented in Table 4.

5 Discussion

Estimating the cumulative risk of at least one FP screening test after repeated rounds of screening is important for understanding the potential harms associated with a screening program. However, estimating this risk is challenging, because typically not all individuals will receive all recommended screening rounds within the study period; and some may drop out of the screening program altogether. An additional complication arises if the FP risk differs depending on the number of screening rounds attended. We reviewed existing statistical methods that fall under two general frameworks: conditional approaches, which estimate risk for the subgroup of subjects who choose to attend a specified number of screening tests; and marginal approaches, which marginalize over the number of screening tests subjects chose to attend. The conditional approaches rely on the assumption that the probability of a FP result at each round of screening is independent of the total number of screening rounds attended. By contrast, the marginal approach allows for variation in the FP risk as a function of number of screening rounds attended but relies on the assumption that the risk of a first FP result at each screening round after censoring is constant.

We used 13 years of data on screening mammograms that the BCSC collected on over 150,000 women to illustrate available statistical approaches for estimating the cumulative FP risk and to evaluate the appropriateness of modeling assumptions. We found evidence that assumptions of both approaches do not hold for the BCSC population. Specifically, at almost every screening round, women who returned for subsequent screening mammograms had lower FP rates than did those who did not return. These differences were not mitigated by adjusting for baseline age, average interval between screening exams, year of first exam, and registry site. Assumptions of the marginal model are untestable because they refer to the FP risk among women who are censored, which is by definition unobserved. However, strong trends in the probability of a first FP across screening rounds among women who were not censored suggest that this is an unrealistic assumption in this context. Thus, the estimated cumulative probability of a FP result after 10 screening rounds of 58.2% based on the conditional model likely underestimates the true risk. By comparison, the estimate of 77.0% based on the marginal approach likely overestimates the risk. We proposed an extension to Xu and colleagues' approach that allowed for variations in the FP probability associated with total number of screening rounds attended and variation in the FP probability across screening rounds. Based on our extended model, we estimated the cumulative risk of a FP test after 10 screening exams to be 63.3%.

Our estimates of cumulative risk are higher than those reported in previous studies. For instance, using data from Harvard Pilgrim Health Care in Boston, MA and a conditional cumulative risk model, Elmore et al.[6] estimated that 49.1% of women would experience at least one FP by their tenth screening mammogram. However, this estimate does not account for such possible confounders as the sample's higher-than-normal rate of family history, irregular screening intervals, and presence or absence of prior comparison films; nor does it address variability across different risk groups or among radiologists. Also, in this study population, the observed FP rate at a single exam across all rounds of screening was only 6.3%, notably lower than that found by the BCSC and other U.S. studies and populations (see e.g. Rosenberg et al.[19]).

Our estimates are also notably higher than those for European screening programs. The FP risk in the triennial NHS Breast Screening Programme conducted in the U.K. is 7.8% at the first screening mammogram and 2.8% at subsequent screening mammograms[23]. A woman

attending all screening rounds in this program would participate in 7 rounds of screening over 20 years and would experience a cumulative FP risk of 22.2%, assuming independence of the FP risk and duration of screening. In a study of the Norwegian screening program[24], the FP rate after 10 biennial screening mammograms over 20 years was estimated to be 20.8%, projected from 3 screening rounds. The FP rate after 10 biennial screening mammograms over 20 years in a Spanish screening program[25] was estimated to be 32.4%, projected from 4 screening rounds. Increased risk was found to be associated with previous benign breast disease, perimenopausal status, high body mass index, and younger age. In a study of the Danish screening program[26], the FP rate after 10 biennial screening mammograms over 20 years was estimated to be 15.8-21.5% for Copenhagen and 8.1-9.6% for Fyn, projected from 3-5 screening rounds and assuming independence between exams.

The European studies are not directly comparable to performance in the context of American clinical practice because screening practice differs markedly between Europe and the United States[27,28,29]. Specifically, European screening programs typically have biennial screening with a much greater volume of screening mammograms interpreted per radiologist and typically screening mammograms are double-read, resulting in markedly lower callback rates than in U.S. practices. The lower callback rate in these studies results in lower cumulative risk of a FP result. The results from Europe are also based on a relatively small number of rounds of screening (3-5) observed per woman. Estimates of the FP rate over 10 rounds of screening are extrapolated from this course of observation.

In addition to model-based estimates of the cumulative FP risk, empirical estimates in the BCSC cohort are also higher than those previously reported. In a study of women undergoing screening mammography at Massachusetts General Hospital Avon Comprehensive Breast Center, the empirical cumulative FP risk among women receiving 10 screening mammograms within a 10 year period was 29.2%[20]. We contrast this to our empirical cumulative FP risk of 44.5% among women receiving 10 or more screening exams. There are several reasons why we might expect the estimate based on the BCSC data to be higher than that in previous studies. First, our sample excluded women who reported previous screening mammograms prior to the first exam captured by the BCSC. Including women who had undergone previous screening would tend to underestimate the FP risk because the risk is highest at the first mammogram. Additionally, our follow-up period spanned more than 10 years allowing for longer intervals between screening exams, which is associated with an increased FP risk[21,22,13]. Other differences may exist between our study population and that used in previous studies. We believe that the FP risk among BCSC women, a nationally inclusive cross-section of women participating in screening mammography in a community setting, is likely to most closely reflect the FP experience of women in the United States.

To more fully understand the risk of a FP after multiple rounds of screening, additional extensions of existing models are needed. First, it is important to consider how the cumulative risk of a FP depends on baseline and time-varying covariates – and to account for the wide variability that has been observed in radiologist interpretive performance[30,31,32,33,19]. Our analysis of the BCSC data has not accounted for these sources of variability. In an extension of the work of Elmore et al.[6], Christiansen et al.[8] addressed the role of possible confounders and between-radiologist variability in performance using the Harvard Pilgrim population. The predicted risk of a FP after 9 mammograms varied across radiologists and as a function of woman-level risk factors from 5% to 100%. Between-radiologist variability in performance was found to be very large, with radiologist effects swamping the impact of all other covariates included in the model. Analogous extensions are needed for the marginal model. To more fully understand the FP risk in the BCSC population we will undertake analyses incorporating woman-level risk factors and between-radiologist variability in future studies. Second, marginal methods for estimating the cumulative probability of a FP result that allow for greater

flexibility in patterns of FP probabilities across screening rounds are needed. We have proposed a simple extension of the marginal method that makes assumptions about FP rates among censored women that are likely to be more appropriate in the context of screening mammography. However, more general extensions that would be appropriate to other screening tests are needed.

In this analysis of the BCSC data, we have focused on the FP recall rate for mammography. FP recalls represent the most prevalent harm of mammography. However, the actual impact of a FP recall is much smaller than the impact of other types of FP events such as biopsies. Appropriate evaluation of a screening program should take into account both the probability of a given harm and its cost. The FP recall risk discussed in this paper represents a common though non-invasive cost of mammography. Previous research on the impact of FP mammograms suggest that women receiving a FP recall experience elevated anxiety and distress[5]. While the evidence indicates that FP screening results are stressful, for most women the adverse effects are transitory. Moreover, a survey by Schwartz et al.[34] revealed that women were highly aware of FP results and highly accepting of FPs as a necessary cost of breast cancer screening, although the women surveyed significantly underestimated the likelihood of experiencing a FP finding over a 10-year period. In addition to the FP recall risk, statistical methods discussed in this paper can be used to estimate the risk of other FP events associated with screening tests. In future research we plan to apply statistical methods discussed here to estimation of other potential harms of mammography such as the FP biopsy risk.

To date, international studies have shown highly variable risk of a FP result for women receiving routine mammography in regular screening programs. The cumulative risk observed in this analysis of women in the BCSC is substantial, and considerably higher than previously projected rates for U.S. women[6]. Despite the high FP risk estimated in the BCSC population, this number should not be used in isolation to question the balance of benefits and harms of mammography screening programs. The estimates of the cumulative risk of a FP estimated for the BCSC women are average estimates for the BCSC population that do not account for starting ages, screening intervals, differential risk, and other factors that may influence the FP rate. While we believe that women should be informed that their risk of one or more FP mammograms is relatively likely over a decade or more of regular screening, the estimated rate in our analysis is an overall rate that does not account for individual risk or other influencing factors, and thus is not easily tailored to an individual woman.

Insofar as screening exams are not diagnostic exams, a certain rate of FPs must be anticipated and accepted given the limitations of the current technology and the goal of detecting small breast cancers. Thus, the relative harm of an FP recall must be weighed against both the frequency and benefit of early cancer detection. The possibility should not be overlooked that women may experience greater anticipatory concerns about FP results if experts overly emphasize the harms, or present pros and cons as if they were of equal importance. Moreover, the inconvenience and anxiety associated with a FP mammogram is likely to be highly variable.

Accurate estimation of the FP risk under various common conditions is an important part of program evaluation, goal setting, and identification of strategies that might be used to reduce the FP rate without compromising test sensitivity. FP risk estimates also allow us to best inform women undergoing screening what they should expect during their participation in an early detection program. Finally, as with most performance indicators in screening, an observed or estimated rate is hardly immutable. With targeted interventions the FP rate could be reduced without also reducing sensitivity

Acknowledgments

This work was supported by the National Cancer Institute's Breast Cancer Surveillance Consortium (BCSC) (U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, U01CA70040). The collection of cancer incidence data used in this study was supported in part by several state public health departments and cancer registries throughout the United States. For a full description of these sources, please see: http://breastscreening.cancer.gov/work/acknowledgement.html. We thank the BCSC investigators and the participating women, mammography facilities, and radiologists for the data they have provided 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/.

References

1. Morrison, A. Screening in Chronic Disease. Oxford University Press; 1992.
2. Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. Annals of Internal Medicine 2002;137:347–360. [PubMed: 12204020]
3. Smith RA, Duffy SW, Gabe R, Tabar L, Yen AM, Chen TH. The randomized trials of breast cancer screening: what have we learned? Radiologic Clinics of North America 2004;42:793–806. [PubMed: 15337416]
4. Brett J, Bankhead C, Henderson B, Watson E, Austoker J. The psychological impact of mammographic screening: A systematic review. Psycho-oncology 2005;14(11):917–938. [PubMed: 15786514]
5. Brewer NT, Salz T, Lillie SE. Systematic review: the long-term effects of false-positive mammograms. Annals of Internal Medicine 2007;146(7):502–510. [PubMed: 17404352]
6. Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, Fletcher SW. Ten-year risk of false positive screening mammograms and clinical breast examinations. N Engl J Med 1998;338(16):1089–96. [PubMed: 9545356]
7. Gelfand AE, Wang F. Modelling the cumulative risk for a false-positive under repeated screening events. Stat Med 2000;19(14):1865–79. [PubMed: 10867676]
8. Christiansen CL, Wang F, Barton MB, Kreuter W, Elmore JG, Gelfand AE, et al. Predicting the cumulative risk of false-positive mammograms. J Natl Cancer Inst 2000;92(20):1657–66. [PubMed: 11036111]
9. Burman ML, Taplin SH, Herta DF, Elmore JG. Effect of false-positive mammograms on interval breast cancer screening in a health maintenance organization. Annals of Internal Medicine 1999;131(1):1–6. [PubMed: 10391809]
10. Xu JL, Fagerstrom RM, Prorok PC, Kramer BS. Estimating the cumulative risk of a false-positive test in a repeated screening program. Biometrics 2004;60(3):651–60. [PubMed: 15339287]
11. Gelfand AE, Smith AF. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 1990;85:398–409.
12. Baker SG, Erwin D, Kramer BS. Estimating the cumulative risk of false positive cancer screenings. BMC Med Res Methodol 2003;3:11. [PubMed: 12841854]
13. Yankaskas BC, Taplin SH, Ichikawa L, Geller BM, Rosenberg RD, Carney PA, et al. Association betwen mammography timing and measures of screening performance in the United States. Radiology 2005;234(2):363–373. [PubMed: 15670994]
14. Lane P, Nelder J. Analysis of covariance and standardization as instances of prediction. Biometrics 1982;38(3):613–621. [PubMed: 7171691]
15. Graubard B, Korn E. Predictive margines with survey data. Biometrics 1999;55(2):652–659. [PubMed: 11318229]
16. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. American Journal of Roentgenology 1997;169:1001–1008. [PubMed: 9308451]
17. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS) Breast Imaging Atlas. 4th. Reston, VA: American College of Radiology; 2003.
18. Taplin SH, Ichikawa LE, Kerlikowske K, Ernster VL, Rosenberg RD, Yankaskas BC, et al. Concordance of Breast Imaging Reporting and Data System Assessments and Management

Recommendations in Screening Mammography. Radiology 2002;222(2):529–535. [PubMed: 11818624]

19. Rosenberg RD, Yankaskas BC, Abraham LA, Sickles EA, Lehman CD, Geller BM, et al. Performance Benchmarks for Screening Mammography. Radiology 2006;241(1):55–66. [PubMed: 16990671]
20. Blanchard K, Colbert JA, Kopans DB, Moore R, Halpern EF, Hughes KS, et al. Long-term risk of false-positive screening results and subsequent biopsy as a function of mammography use. Radiology 2006;240(2):335–342. [PubMed: 16864665]
21. Hunt KA, Rosen EL, Sickles EA. Outcome analysis for women undergoing annual versus biennial screening mammography: a review of 24,211 examinations. American Journal of Roentgenology 1999;173(2):285–289. [PubMed: 10430120]
22. Michaelson JS, Kopans DB, Cady B. The breast carcinoma screening interval is important. Cancer 2000;88:1282–1284. [PubMed: 10717607]
23. NHS Breast Screening Programme. Annual Review 2008. NHS Cancer Screening Programmes; 2009.
24. Hofvind S, Thoresen S, Tretli S. The cumulative risk of a false-positive recall in the Norwegian Breast Cancer Screening Program. Cancer 2004;101(7):1501–7. [PubMed: 15378474]
25. Castells X, Molins E, Macia F. Cumulative false positive recall rate and association with participant related factors in a population based breast cancer screening programme. J Epidemiol Community Health 2006;60(4):316–21. [PubMed: 16537348]
26. Njor SH, Olsen AH, Schwartz W, Vejborg I, Lynge E. Predicting the risk of a false-positive test for women following a mammography screening programme. J Med Screen 2007;14(2):94–7. [PubMed: 17626709]
27. Smith-Bindman R, Chu PW, Miglioretti DL, Sickles EA, Blanks R, Ballard-Barbash R, et al. Comparison of screening mammography in the United States and United Kingdom. Journal of the American Medical Association 2003;290(16):2129–2137. [PubMed: 14570948]
28. Yankaskas BC, Klabunde CN, Ancelle-Park R, Rennert G, Wang H, Fracheboud J, et al. International comparison of performance measures for screening mammography: can it be done? Journal of Medical Screening 2004;11:187–193. [PubMed: 15624239]
29. Hofvind S, Vacek PM, Skelly J, Weaver DL, Geller BM. Comparing Screening Mammography for Early Breast Cancer Detection in Vermont and Norway. Journal of the National Cancer Institute 2008;100(15):1082–1091. [PubMed: 18664650]
30. Beam CA, Lavde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. Archives of Internal Medicine 1996;156:209–213. [PubMed: 8546556]
31. Wagner RF, Beam CA, Beiden SV. Reader variability in mammography and its implications for expected utility over the population of readers and cases. Medical Decision Making 2004;24:561–572. [PubMed: 15534338]
32. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. Journal of the National Cancer Institute 2004;96(24):1840–1850. [PubMed: 15601640]
33. Smith-Bindman R, Chu P, Miglioretti DL, Quale C, Rosenberg RD, Cutter G, et al. Physician predictors of mammographic accuracy. Journal of the National Cancer Institute 2005;97(5):358–367. [PubMed: 15741572]
34. Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch HG. US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. British Medical Journal 2000;320:1635–1640. [PubMed: 10856064]

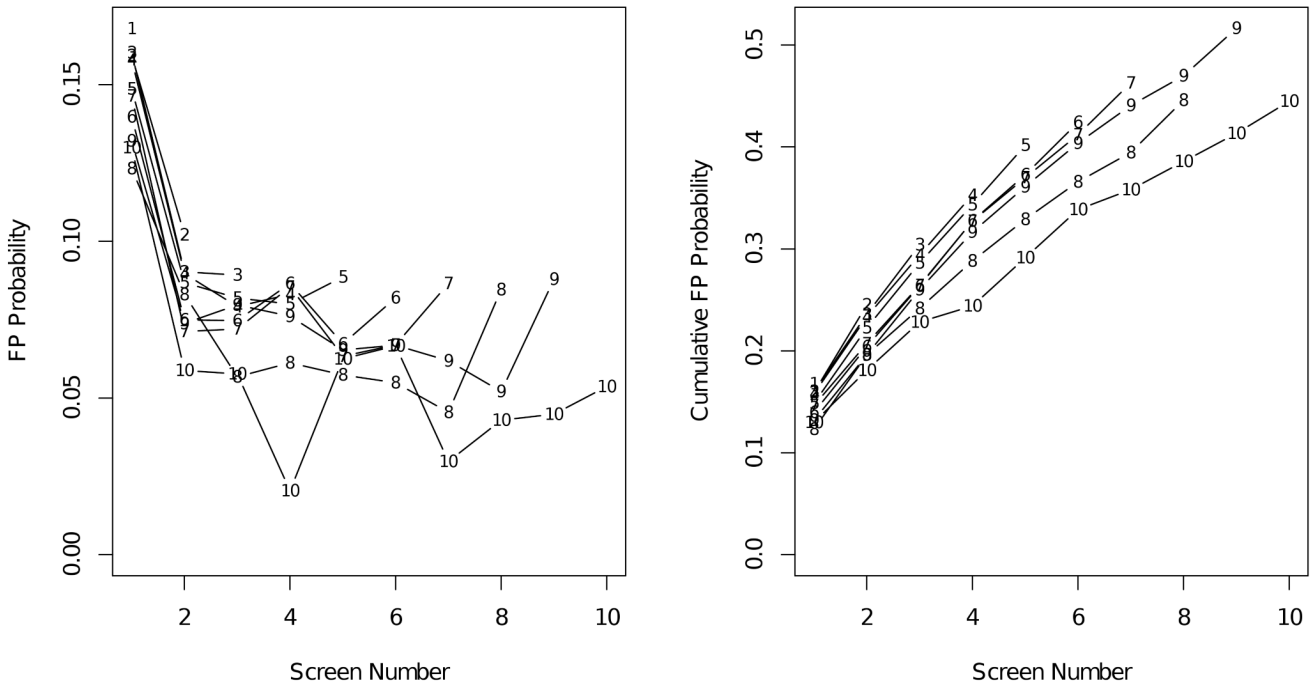

Figure 1.
Empirical estimates of probability of a first FP screening exam (left) and cumulative probability of a first FP screening exam (right) estimated separately by total number of screening rounds attended. Estimates for "10" screens include women receiving 10 or more screening exams.

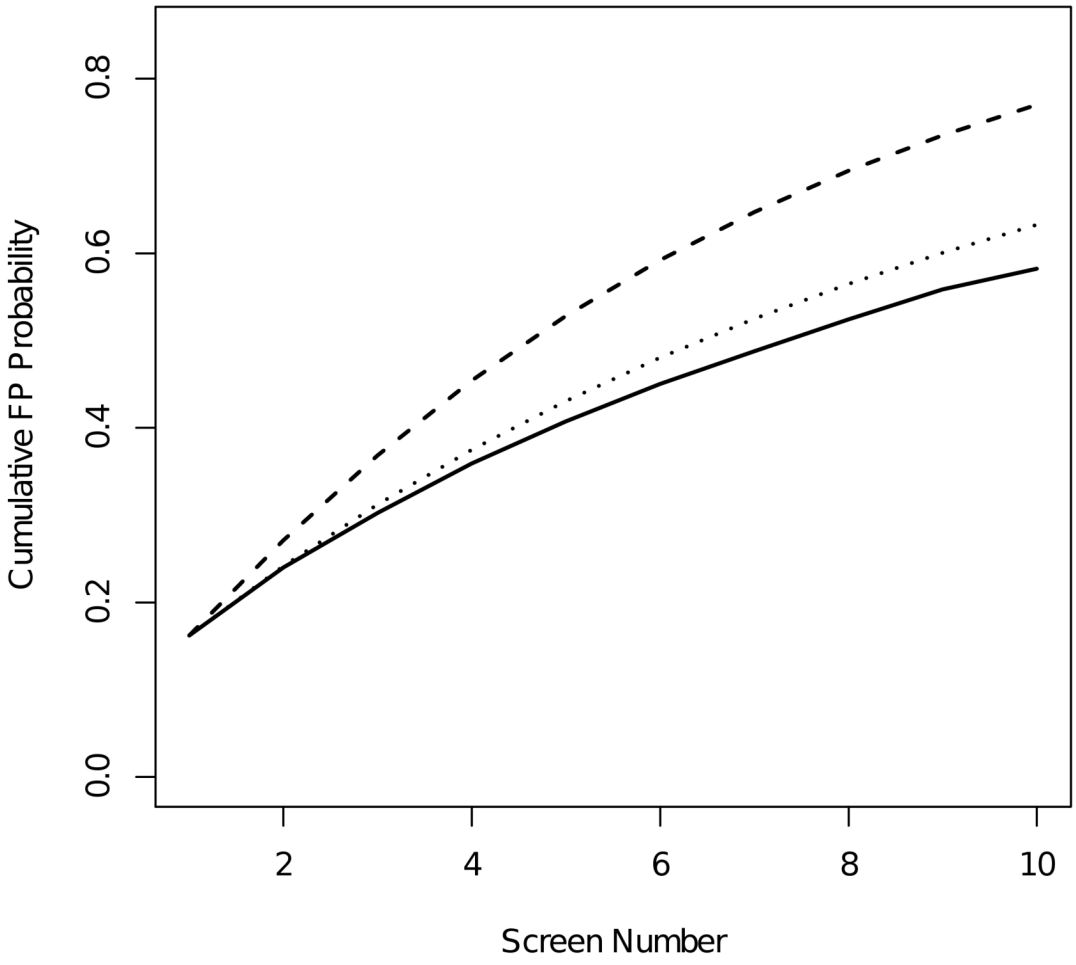

Figure 2.
Conditional (solid), marginal (dashed), and adjusted marginal (dotted) estimates of the cumulative probability of a FP screening exam.

Table 1

Summary of characteristics of women in the BCSC population.

Number of women		%
Age at first mammogram		
40 - 44	95,768	60.0
45 - 49	31,619	19.8
50 - 54	20,689	13.0
55 - 59	11,498	7.2
Average time between screens		
9 - 18 mos	31,298	39.2
19 - 30 mos	30,085	37.7
30 - 42 mos	10,252	12.8
>42 mos	8,254	10.3
Number of screening exams		
1	79,684	49.9
2	33,101	20.7
3	19,139	12.0
4	11,677	7.3
5	7,318	4.6
6	4,220	2.6
7	2,417	1.5
8	1,249	0.8
9	515	0.3
10+	254	0.2

Table 2

Empirical probability of a first FP and empirical cumulative probability of a first FP by screening round.

Screen Number	Number Exams	First FP (%)	Cumulative FP (%)
1	159,574	16.2	16.2
2	67,419	9.3	23.5
3	36,161	8.2	29.1
4	19,873	8.1	34.0
5	10,702	7.5	38.0
6	5,517	7.2	40.9
7	2,688	6.8	43.5
8	1,207	7.1	44.4
9	429	7.2	48.2
10	149	5.4	44.5

Table 3

Cumulative FP probabilities by screening round stratified by participation in additional screening rounds and results of a hypothesis test for association between number of screens and history of screening exam results. Unadjusted probabilities (Unadj) and hypothesis test results are presented in addition to results adjusted for baseline age, average interval between screening exams, year of first exam, and registry (Adj). N = number of women.

Screening round	Additional screens			Last screen			Unadj	Adj
	N	Unadj %	Adj %	*N*	Unadj %	Adj %	*p*	*p*
1	12,471	15.6	16.1	13,387	16.8	16.3	<0.01	0.51
2	10,628	22.7	22.5	8,115	24.5	24.9	<0.01	<0.01
3	7,777	28.1	27.2	5,822	30.4	31.9	<0.01	<0.01
4	5,271	33.0	31.8	4,121	35.3	37.1	<0.01	<0.01
5	3,138	36.3	34.3	2,939	40.2	42.6	<0.01	<0.01
6	1,747	39.4	36.9	1,790	42.4	45.2	0.005	<0.01
7	811	40.2	37.8	1,119	46.3	48.4	<0.01	<0.01
8	340	44.2	40.4	557	44.6	47.0	0.90	0.01
9	105	41.3	38.9	266	51.7	52.9	0.01	<0.01

Table 4

Cumulative probability of a FP screening exam result (95% confidence interval (CI)) after the first, fifth, and tenth screening exams.

	First Round	Fifth Round	Tenth Round
Conditional	16.2 (16.0, 16.4)	40.7 (40.3, 41.2)	58.2 (56.1, 60.4)
Marginal	16.2 (16.0, 16.4)	52.8 (52.5, 53.2)	77.0 (76.7, 77.3)
Adjusted marginal	16.2 (16.1, 16.3)	43.1 (43.0, 43.1)	63.3 (63.2, 63.3)