Performance assessment for radiologists interpreting screening mammography

D. B. Woodard [1,*] A. E. Gelfand [1], W. E. Barlow [2] and J. G. Elmore [3]

[1] *Institute of Statistics and Decision Sciences, Duke University, 27708-0251*

[2] *Cancer Research and Biostatistics, Seattle, Washington*

[3] *University of Washington School of Medicine*

*Correspondence to: Institute of Statistics and Decision Sciences, Duke University, 27708-0251, dawn@stat.duke.edu, (919)593-4531, or fax (919) 684-8594

Contract/grant sponsor: AHRQ/NCI; contract/grant number: HS10591

SUMMARY

When interpreting screening mammograms radiologists decide whether suspicious abnormalities exist which warrant the recall of the patient for further testing. Previous work has found significant differences in interpretation among radiologists; their false-positive and false-negative rates have been shown to vary widely. Performance assessments of individual radiologists have been mandated by the U.S. government, but concern exists about the adequacy of current assessment techniques.

We use hierarchical modeling techniques to infer about interpretive performance of individual radiologists in screening mammography. While doing this we account for differences due to patient mix and radiologist attributes (for instance, years of experience or interpretive volume). We model at the mammogram level, and then use these models to assess radiologist performance. Our approach is demonstrated with data from mammography registries and radiologist surveys. For each mammogram, the registries record whether or not the woman was found to have breast cancer within one year of the mammogram; this criterion is used to determine whether the recall decision was correct.

We model the false-positive rate and the false-negative rate separately using logistic regression on patient risk factors and radiologist random effects. The radiologist random effects are, in turn, regressed on radiologist attributes such as the number of years in practice.

Using these Bayesian hierarchical models we examine several radiologist performance metrics. The first is the difference between the false-positive or false-negative rate of a particular radiologist and that of a hypothetical "standard" radiologist with the same attributes and the same patient mix. A second metric predicts the performance of each radiologist on hypothetical mammography exams with particular combinations of patient risk factors (which we characterize as "typical", "high-risk", or "low-risk"). The second metric can be used to compare one radiologist to another, while the first

metric addresses how the radiologist is performing compared to an appropriate standard. Interval estimates are given for the metrics, thereby addressing uncertainty.

The particular novelty in our contribution is to estimate multiple performance rates (sensitivity and specificity). One can even estimate a continuum of performance rates such as a performance curve or ROC curve using our models and we describe how this may be done. In addition to assessing radiologists in the original data set, we also show how to infer about the performance of a new radiologist with new case mix, new outcome data, and new attributes without having to refit the model.

1. INTRODUCTION

When interpreting screening mammograms radiologists decide whether or not a patient should be recalled for further testing. Such additional testing typically consists of a diagnostic mammogram, ultrasound, or biopsy. Since the recall decision necessarily involves some subjectivity, efforts have been made to standardize mammographic interpretation among radiologists. Standards for educational programs have been implemented, and licensing and board certification of radiologists is designed to create consistency among radiologists [1]. However, there are still significant differences in interpretation among radiologists; their recall rates and false-positive and false-negative rates have been shown to vary widely ([2],[3]). Concerns about such inconsistency in the quality of mammography led to the 1992 Mammographic Quality Standards Act [4].

As part of this act, radiology facilities must keep standardized records of patient outcomes

in screening mammography. The records are reviewed annually by the Food and Drug Administration as part of physician accreditation renewal. If a physician has a particularly high rate of missed cancers, or a particularly high false-positive rate, his/her accreditation may be revoked.

Such monitoring of clinician performance has become a widely used technique in health care policy. Unlike licensing exams, it directly measures results rather than knowledge, and can be done on an ongoing basis. Monitoring can provide feedback to clinicians and inform professional educational efforts. However, there are concerns about the fairness of punishing physicians based on empirical performance rates [5]. A physician could have low reported rates of desirable patient outcomes simply due to seeing many "high-risk" patients. In addition, if the disease being treated is rare, individual physicians may have seen very few cases, leading to large random variability in the reported rate at which patients have a desirable outcome.

In order to compare physician performance rates fairly, there have been a number of attempts to adjust for differences in patient risk factors ([6], [7], [8]). Such adjustment for "case mix" can reduce the estimated variability among physicians in performance. In order to assess a particular physician, a statistical test can be done to see whether the physician's performance is significantly above- or below-average. This statistical test is "frequentist," meaning that it is only valid if the physician has seen a "large" number of cases. It is not clear how many cases are necessary, but a cutoff is chosen, and physicians with fewer than that number of cases are not assessed at all. This is a potential problem when assessing the false-negative rate of radiologists reading screening mammograms, since most radiologists encounter few cancer

cases. Another issue with case-mix adjustment is that it is not clear how to directly compare one physician to another.

A method which allows such a comparison between any two health providers and which is valid regardless of the number of cases was introduced by Normand, Glickman, and Gatsonis [9]. They use a hierarchical patient-level model for heart attack patient survival rates at different hospitals, and then use this model to assess hospital performance. The uncertainty associated with hospital performance emerges directly under their fully Bayesian model. Since no frequentist tests are needed, their analysis does not require a minimum number of heart attack patient outcomes recorded at each hospital. In their model, information from all the hospitals helps to infer about those with few cases.

Normand et al. propose two performance metrics for hospitals based on their model. The first is a hospital's predicted survival rate for hypothetical patients with a particular combination of risk factors. If a hospital has a high probability of having survival rate on these patients below 2/3 of the average hospital survival rate, the hospital is considered to be underperforming. The second metric compares the performance of the hospital to that of a hypothetical "standard" hospital, which has the same attributes and patient mix but is of average quality. If the hospital has a high probability of having survival rate below 2/3 that of the "standard" hospital, the hospital is considered to be underperforming.

We use similar models and performance metrics in a new clinical context, namely the performance of radiologists in mammography. Like Normand et al. we use Bayesian hierarchical

patient-level models for the rates at which health providers (in our case radiologists) have desirable patient outcomes. These models include both patient risk factors and provider attributes. Also similarly to Normand et al., we assess each provider by predicting their performance rates on patients with a particular combination of risk factors. Additionally, we follow Normand et al. in comparing the performance of each provider to that of a hypothetical "standard" provider with the same attributes and patient mix.

We elaborate the approach of Normand et al. in several different ways. First, we model multiple performance rates, namely the percentage of cancers which are found (sensitivity or one hundred percent minus the false-negative rate) and the percentage of non-cancer patients who get a correct negative screening result (specificity or one hundred percent minus the false-positive rate). In fact, our method allows extension to a continuum of such performance rates, such as *performance curves* or ROC-curve based measures [10]. Of course, if comparison is to be made by reduction of a performance curve or an ROC curve to a single number, then we merely have a generalized version of the Normand et al. type of assessment.

In terms of clinical interpretation, we clarify the distinction between comparison across providers (in our case radiologists) and comparison of an individual provider to a suitable standard. In traditional performance assessment, case mix adjustment is done, and the resulting adjusted performance rates are compared to each other. By contrast, we compare radiologists using their predicted performance rates on a particular type of patient. In addition, we are able to assess the radiologists individually by comparing them to a standard which is appropriate for them. Of course, that standard is based on the hierarchical model, which is

fitted to the entire collection of radiologists, and in that sense each individual radiologist is being assessed in the context of all the radiologists.

The Bayesian approach provides uncertainty for all of the assessments that are done. There is no minimum number of patients required in order to assess the performance of a radiologist; if the radiologist has seen very few cases, then his/her performance assessment will include the uncertainty which may result from having little evidence. For some radiologists this leads to wide intervals for the performance (see section 5.2). However, there are some radiologists who have seen few cancer cases but who failed to catch any of the cancers. As a result they may be assessed as being below-average, since these few cancer cases actually can give a fair amount of evidence that the radiologist is underperforming. Using traditional assessment techniques that require a minimum number of cases these underperforming radiologists would not be discovered.

We also consider the problem of how to predict the performance of a radiologist who is not included in the original data analysis. That is, if after fitting the model we are provided with a new radiologist's attributes, her/his case mix, and the resultant outcomes with regard to recall and cancer incidence, can we still do a performance assessment? We show how to do this and offer a simple post-model fitting approach to implement it.

We investigate performance for radiologists working at three different sites: western Washington, Colorado, and New Hampshire. These sites all participate in the Breast Cancer Surveillance Consortium [11] and all record screening mammography information in the same

fashion. During the period January 1, 1996 to December 31, 2001, we have more than half a million mammograms across these sites. Again, we confine our performance assessment to screening mammograms. Diagnostic mammograms arise under different clinical circumstances, with different definitions of sensitivity and specificity.

Lastly, we note that all of the methodology we present here is applicable to general screening procedure contexts. Such settings typically reduce the outcomes to a two by two table, screening test result by disease outcome. More generally, we can also consider classification settings where more than two categorical outcomes are possible for both the screening test and the patient health status. Other screening settings that could be investigated using our methodology include cardiac (ECHO) examinations, clinical breast examinations, magnetic resonance imaging (MRI), and ultrasound.

The format of the paper is as follows. In section 2 we describe the motivating mammography data. Section 3 presents the Bayesian hierarchical models for sensitivity and specificity upon which we base the performance assessments. The proposed performance assessments are formalized in section 4. Section 5 summarizes the results of the modeling and the performance metric estimation. Section 6 turns to prediction problems associated with performance assessment. Section 7 notes extension to more complicated multiple performance rates while section 8 concludes with a summary and some future directions.

2. SCREENING MAMMOGRAPHY

Mammogram data was taken from three registries that are members of the Breast Cancer Surveillance Consortium (BCSC) [11]. The mammograms date from January 1996 through December 2001 inclusive. The registries, namely the New Hampshire Mammography Network, the Colorado Mammography Program, and the Breast Cancer Surveillance Project at the Group Health Cooperative (Washington state), cover a diverse geographic area. The registries record for each patient a variety of risk factors. In addition they note the identity of the radiologist and the outcome of the mammogram (whether or not the patient was recalled for further testing). In order to assess the accuracy of these results, the registry data was aligned with breast cancer registry data. Following the BCSC, we define breast cancer as either invasive carcinoma or ductal carcinoma *in situ*. We also use the definition of false positive and false negative cases that has been adopted by the BCSC. According to this definition, if a patient is found to have breast cancer within one year of a mammogram the cancer is presumed to have been present at the time of the mammogram.

The mammographic data was de-identified to protect patient privacy. The study activities were approved by institutional review boards at the University of Washington School of Medicine, the Cooper Institute (Colorado), Dartmouth College (New Hampshire), and the Group Heath Cooperative (Washington).

Only screening mammograms were included in the analysis. A screening mammogram is defined to be one designated as such by the radiologist. Unilateral mammograms, as well as those occurring less than nine months after a previous mammogram, were excluded because

they were presumed to be for diagnosis rather than screening. Women younger than 40, with a previous history of breast cancer, or reporting breast implants at the time of the exam were also excluded.

The mammogram outcomes are coded using the BI-RADS^{TM} assessment system. The system has categories 0 through 5. For the purposes of this paper an outcome was considered positive if it was categorized as 0 (needs additional imaging evaluation), 4 (suspicious abnormality), or 5 (highly suggestive of malignancy). In addition, an assessment of 3 (probably benign) is considered to be a positive result if there was a recommendation for immediate work-up.

There is considerable variation in sensitivity and specificity for the radiologists who were represented in the mammography databases. In order to explain some of this variation, we gathered survey data on radiologist clinical practice, demographic attributes, and level of concern about malpractice claims. Surveys were sent to 181 radiologists who were represented in the registries. The radiologists were told that their responses would be linked to observed performance, but that they would not be identified. 139 radiologists responded to the survey, yielding a 77% response rate, high for physician surveys. Furthermore, no significant differences were found between responders and non-responders with respect to recall rates, sensitivity, or specificity [2]. Approval for the radiologist surveys was obtained from the previously mentioned institutional review boards.

Of the radiologist attributes from the survey, only the clinical practice attributes have been

found to have a strong link to radiologist sensitivity and specificity [2]. Three of these attributes were found to be relevant either in [2] or in our preliminary analysis; a summary of these is shown in Table I. The table shows the number of radiologists in each category, along with the empirical sensitivity and specificity of the radiologists in that category. In calculating the performance rates each mammogram is given equal weight, so radiologists who have seen more mammograms have more influence on these measures.

The first radiologist attribute shown is the number of years that the radiologist had been interpreting mammograms since completing training. The second is the annual mammographic volume of the radiologist, measured by the total number of mammograms that the radiologist estimated having interpreted in 2001. The last is whether the radiologist is affiliated with an academic medical center.

Out of the 139 responding radiologists, we considered only those who had read at least 480 of the mammograms in the registries between 1996 and 2001. This level is used as a criterion for licensure to read mammographic film in the U.S. [4], and is indicative of a radiologist who has a substantial commitment to such activity. There were 124 such radiologists. Of these, 120 responded to all three of the questions in Table I; only their mammograms have been included in the table and in the subsequent analysis. In total, there were 550,648 mammograms in the registries which were interpreted by these radiologists.

Patient risk factors were recorded by the registries. Those that are thought to be important determinants of sensitivity or specificity are listed in Table II. Sensitivity is thought to increase

with patient age [12]. We categorized age into four intervals as shown in the table. Dense or heterogeneous breast tissue has been shown to decrease sensitivity and specificity [2]. Breast tissue density/heterogeneity was recorded by the radiologist and coded according to the four levels in the BI-RADS^{TM} system, as listed in the table. First-time mammograms have a lower specificity [2]. Whether or not the patient had a mammogram previously was based on self-reporting and/or on registry records.

Additional factors that have been found to affect recall rates include menopausal status, the use of hormone replacement therapy, and the presence of breast cancer symptoms or a patient history of breast biopsy or surgery [13]. Menopausal status was by self-report unless the woman was aged 55 or older, in which case she was presumed to be post-menopausal. The use of hormone replacement therapy was by self-report. Mammograms with missing patient information were excluded leaving 399,014 available for this analysis. Most of this missingness was because breast density was frequently not recorded. Due to this removal of the missing data, some radiologists had fewer than 480 mammograms in the analysis.

We fit the sensitivity model for all of the remaining 2,043 cancer cases. However, to reduce computational demand, we did not fit the specificity model for all 396,971 non-cancer cases. Instead, we randomly subsampled 250 of the non-cancer mammograms for each radiologist, or all of their non-cancer mammograms if they had fewer than 250. This led to a total of 29,864 mammograms for fitting the specificity model. This subsampling does not bias the coefficient estimates, although it could decrease the power of the specificity analysis by leading to wider interval estimates for the model parameters.

3. HIERARCHICAL MODELS FOR SENSITIVITY AND SPECIFICITY

We characterize mammographic accuracy by fitting two models, one for the probability of a patient not being recalled given that she did not have cancer (specificity) and one for the probability of her being recalled given that she had cancer (sensitivity).

For each model we use a logistic regression on the patient risk factors and radiologist random effects as follows.

$$\text{Specificity} : logit(Pr(R_{ij} = 0|C_{ij} = 0)) = \mathbf{X}'_{ij}\beta_i^{(0)} + \tau_i^{(0)} \quad (1)$$

$$\text{Sensitivity} : logit(Pr(R_{ij} = 1|C_{ij} = 1)) = \mathbf{X}'_{ij}\beta_i^{(1)} + \tau_i^{(1)} \quad (2)$$

Here the radiologists are indexed by $i = 1, ..., I$ while the mammograms associated with radiologist i without and with cancer, respectively, are indexed by $j = 1, ..., n_i^{(0)}$ and $j = 1, ..., n_i^{(1)}$. Additionally, $R_{ij} = 1$ denotes recall for mammogram (i, j), $C_{ij} = 1$ denotes *cancer present* with respect to mammogram (i, j), and $\mathbf{X}_{ij}$ denotes the risk factors of the patient associated with mammogram (i, j). The patient risk factors are described in section 2. They include an indicator of the registry in which the mammogram is recorded. This is because, despite the standards for patient care established by the Mammographic Quality Standards Act [4], there could potentially still be geographic differences in patient outcomes.

The radiologist random effects are then explained using a set of radiologist attributes $\mathbf{W}_i$ (again, listed in section 2), leading to the second stage regression models which are as follows.

$$\tau_i^{(0)} \overset{iid}{\sim} \mathcal{N}(\mathbf{W}_i'\gamma^{(0)}, \sigma^{2(0)}) \tag{3}$$

$$\tau_i^{(1)} \overset{iid}{\sim} \mathcal{N}(\mathbf{W}_i'\gamma^{(1)}, \sigma^{2(1)}) \tag{4}$$

Paliwal et al. [13] find no evidence for the use of radiologist-specific patient risk factor coefficients, so we set $\beta_i^{(0)} = \beta^{(0)}$ and $\beta_i^{(1)} = \beta^{(1)}$.

The likelihood associated with the specificity model, i.e., with (1) and (3) is as follows. A similar expression for the sensitivity arises using (2) and (4).

$$\begin{aligned} & L(\beta^{(0)}, \{\tau_i^{(0)}\}, \gamma^{(0)}, \sigma^{2(0)}; \{R_{ij}\}) \\ = & \prod_{i=1}^{I} \left[\prod_{j=1}^{n_i^{(0)}} Pr(R_{ij}|C_{ij} = 0, \beta^{(0)}, \tau_i^{(0)}) \right] \times \mathcal{N}(\tau_i^{(0)}; \mathbf{W}_i'\gamma^{(0)}, \sigma^{2(0)}) \end{aligned} \tag{5}$$

Here $Pr(R_{ij}|C_{ij} = 0, \beta^{(0)}, \tau_i^{(0)})$ is the probability of the recall status for mammogram (i, j) given no cancer. Using (1) this is:

$$Pr(R_{ij}|C_{ij} = 0, \beta^{(0)}, \tau_i^{(0)}) = \left[\frac{exp(\mathbf{X}_{ij}^T\beta^{(0)} + \tau_i^{(0)})}{1 + exp(\mathbf{X}_{ij}^T\beta^{(0)} + \tau_i^{(0)})} \right]^{1-R_{ij}} \left[\frac{1}{1 + exp(\mathbf{X}_{ij}^T\beta^{(0)} + \tau_i^{(0)})} \right]^{R_{ij}} \tag{6}$$

The Bayesian framework is attractive in analyzing such multilevel models since it avoids asymptotic assumptions in assessing variability. In this framework, model specification is completed when the prior distributions of the parameters are chosen. We used vague independent normal priors for the regression coefficients $\gamma^{(0)}$, $\beta^{(0)}$, $\gamma^{(1)}$, and $\beta^{(1)}$, with mean 0 and variance 10. A variance of 10 was thought to be large enough because in previous studies involving regression of radiologist performance rates on radiologist and patient attributes all

of the regression coefficient intervals were contained in the range (-3,3) (see [2], [13]). We also found that the results of the analysis were not strongly affected by the choice of a larger prior variance for the regression coefficients.

For the prior of the residual variance of the radiologist effects (σ^2) we chose an inverse gamma distribution with a data-based mean but infinite variance. The mean was chosen by taking the empirical sensitivities or specificities of the radiologists, and transforming them to the log-odds scale. On this scale they were regressed linearly on the radiologist attributes. The mean squared error of this regression was taken as the prior mean for the random effect variance. Inference of the performance metrics was not strongly affected by this choice.

These prior and likelihood specifications determine the posterior distributions of the parameters. To implement inference in the Bayesian framework, samples from these posterior distributions are obtained using Gibbs sampling. See the appendix for details on the required full-conditional distributions. Lastly, note that the two models (sensitivity and specificity) can be fitted separately since there are no common mammograms and no common model parameters.

4. PERFORMANCE METRICS

Within the model specification and notation of section 3, we assess performance in the spirit of Normand, Glickman, and Gatsonis [9]. Specificity and sensitivity will be denoted respectively

as follows.

$$sp(\mathbf{X}, \beta^{(0)}, \tau^{(0)}) = Pr(R = 0|C = 0, \mathbf{X}, \beta^{(0)}, \tau^{(0)}) \tag{7}$$

$$se(\mathbf{X}, \beta^{(1)}, \tau^{(1)}) = Pr(R = 1|C = 1, \mathbf{X}, \beta^{(1)}, \tau^{(1)}) \tag{8}$$

The right hand sides of these expressions are calculated using (1) and (2) respectively. We define the case-mix averaged specificity and sensitivity, respectively, for radiologist i as follows.

$$\overline{sp}_i(\beta^{(0)}, \tau_i^{(0)}) = \frac{\sum_{j=1}^{n_i^{(0)}} sp(\mathbf{X}_{ij}, \beta^{(0)}, \tau_i^{(0)})}{n_i^{(0)}} \tag{9}$$

$$\overline{se}_i(\beta^{(1)}, \tau_i^{(1)}) = \frac{\sum_{j=1}^{n_i^{(1)}} se(\mathbf{X}_{ij}, \beta^{(1)}, \tau_i^{(1)})}{n_i^{(1)}} \tag{10}$$

Here $n_i^{(0)}$ and $n_i^{(1)}$ are the number of non-cancer case and cancer case mammograms respectively, interpreted by radiologist i.

Employing the bracket notation of Gelfand and Smith [14] to represent densities, we obtain samples from $[\overline{sp}_i(\beta^{(0)}, \tau_i^{(0)})|\mathcal{D}^{(0)}]$ and $[\overline{se}_i(\beta^{(1)}, \tau_i^{(1)})|\mathcal{D}^{(1)}]$, where $\mathcal{D}^{(0)}$ and $\mathcal{D}^{(1)}$ denote, respectively, all of the data used to fit these models. We do this through the posterior samples of $(\beta^{(0)}, \tau_i^{(0)})$ and $(\beta^{(1)}, \tau_i^{(1)})$ obtained from fitting the models in (1)-(4).

Note that the averages (9) and (10) do not "adjust" radiologist specificity and sensitivity for case mix across radiologists. Rather, they average performance for each radiologist across her/his case mix. For example, radiologists who have more *difficult* case mix will likely have lower specificity and sensitivity.

Hence, we now define an appropriate standard against which to compare an individual radiologist's $\overline{sp}$ and $\overline{se}$. The standard is derived by averaging $\tau_i^{(0)}$ and $\tau_i^{(1)}$ against their priors as follows.

$$\overline{sp}_i^{(std)}(\beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)}) = \int \overline{sp}_i(\beta^{(0)}, \tau_i^{(0)})[\tau_i^{(0)}|\mathbf{W}_i, \gamma^{(0)}, \sigma^{2(0)}]d\tau_i^{(0)}$$

$$\overline{se}_i^{(std)}(\beta^{(1)}, \gamma^{(1)}, \sigma^{2(1)}) = \int \overline{se}_i(\beta^{(1)}, \tau_i^{(1)})[\tau_i^{(1)}|\mathbf{W}_i, \gamma^{(1)}, \sigma^{2(1)}]d\tau_i^{(1)}$$

In practice, instead of exact integration we use Monte Carlo integration, sampling from the conditional normal distributions for the τ's. Posterior samples of $\overline{sp}_i^{(std)}(\beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)})$ and $\overline{se}_i^{(std)}(\beta^{(1)}, \gamma^{(1)}, \sigma^{2(1)})$ are obtained through posterior samples of $(\gamma^{(0)}, \sigma^{2(0)})$ and $(\gamma^{(1)}, \sigma^{2(1)})$, respectively. However, for performance assessments we will want to sample the posteriors $[\overline{sp}_i(\beta^{(0)}, \tau_i^{(0)}) - \overline{sp}_i^{(std)}(\beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)})|\mathcal{D}^{(0)}]$ and $[\overline{se}_i(\beta^{(1)}, \tau_i^{(1)}) - \overline{se}_i^{(std)}(\beta^{(1)}, \gamma^{(1)}, \sigma^{2(1)})|\mathcal{D}^{(1)}]$. Through significant departure from 0, these latter posteriors will indicate significant under-performance or over-performance. We find such posteriors more informative than the ad-hoc "2/3" rules suggested in Normand et al. [9], as is discussed in section 1.

We now turn to comparison across radiologists. One can predict the sensitivity or specificity of each radiologist on a hypothetical mammography exam with a particular combination of patient risk factors $\mathbf{X}_0$. These predictions correspond to estimating the posterior distributions $[sp(\mathbf{X}_0, \beta^{(0)}, \tau^{(0)})|\mathcal{D}^{(0)}]$ and $[se(\mathbf{X}_0, \beta^{(1)}, \tau^{(1)})|\mathcal{D}^{(1)}]$. In fact, we can introduce an ensemble of risk factor combinations, each representing patients who are, for example, "typical" or "high risk."

In considering these various performance measures the practical differences may be

summarized as follows. Pairwise comparison between radiologists may be done using predicted specificity or sensitivity on a specified combination of patient risk factors. This measure accounts for differences among radiologists in case mix but does not account for differences in radiologist attributes. Performance relative to the standard, however, does take into account the radiologist attributes. The radiologist is compared to a standard appropriate for his/her set of attributes. Lastly, note that while radiologists are evidently concerned with implementing accurate screening mammography, it is also clear that, in processing their case mixes, they are not attempting to optimize their performance with regard to any specific criterion.

5. COMPUTATION AND RESULTS

In subsection 5.1 we look at the results of fitting the separate sensitivity and specificity models. In subsection 5.2 we turn to the results for the performance metrics described in the previous section.

5.1. Sensitivity and specificity models

Gibbs sampling was used to fit the models for sensitivity and specificity. Since the outcome is binary-valued, closed-form conditional posterior distributions do not exist for the patient regression coefficients or the radiologist random effects. For this reason we use adaptive rejection sampling (ARS) (see [15] for a description) to sample from the conditional posterior distributions of these parameters. We use an ARS implementation developed by Giovanni Petris at the University of Arkansas. All of the computation is implemented in the R statistical package (see www.r-project.org).

Table III shows the resulting coefficient estimates for the sensitivity and specificity models. For each patient risk factor and radiologist attribute the table gives coefficient point estimates and 95% credible intervals. We find that breast density, mammographic history, menopause/hormone replacement therapy status, the presence of breast cancer symptoms, and a history of breast biopsy or surgery are important factors in the probability of recall for women with no cancer (important factors in specificity). These results are consistent with Paliwal et al. [13], who find that these patient risk factors are strong determinants of the probability of recall for all women.

In addition to the patient risk factors, the radiologist attributes are also correlated with specificity. Radiologists who have been in practice for a longer period of time and those who are associated with an academic center have higher specificity. Radiologists with a medium case load have lower specificity than those with either a large or small case load.

Fewer patient risk factors and radiologist attributes are shown to affect sensitivity than specificity. This is reasonable given that the sensitivity model is fit on many fewer cases (those with cancer) than the specificity model (those without). We find that a woman's breast density, her mammographic history, and possibly her age affect the probability that a woman's cancer is found. Radiologists who have been trained more recently and those with a larger case volume have slightly higher sensitivity.

Although our results are theoretically dependent on our choice of prior distribution for the model parameters, we find that different vague priors give very similar coefficient estimates

and intervals. For instance, increasing the prior variance of the coefficients does not have a large effect on Table III. This suggests that the Bayesian inferences about the coefficients are dominated by the information in the data. This is true even in the sensitivity model, which is fit on fewer cases than the specificity model. Indeed, our results regarding the effect of the patient and radiologist attributes on sensitivity and specificity are much like results from similar frequentist models [2].

5.2. Performance analysis

Using the Bayesian models given by (1)-(4) we estimate several performance metrics for each radiologist. For each performance metric we obtain a point estimate and an interval estimate; the latter captures the uncertainty about the radiologist's performance. One metric described in section 4 predicts each radiologist's sensitivity or specificity on a hypothetical mammography exam with a particular combination of patient risk factors. This metric allows comparison of the predicted sensitivity or specificity of two radiologists. We first choose "typical" values of the patient risk factors, meaning that we take a hypothetical patient in registry A with no history of breast biopsy or surgery or breast cancer in the family, who is aged 50 to 59 but post-menopausal, not using hormone replacement therapy, who has had a previous mammogram, who has breast density classification 2, and no symptoms of breast cancer. Figure 1 compares the predicted sensitivity and specificity for such a patient using a randomly selected subset of fifteen radiologists.

The point estimates in the figure are the posterior means of the radiologists' sensitivity or specificity. The expected sensitivity ranged from 58.9% to 91.8%, while the expected specificity

ranged from 83.5% to 96.9%. This suggests that sensitivity varies more among radiologists than specificity, though this is mitigated somewhat due to the greater uncertainty associated with sensitivity. In Table IV we show the range of expected radiologist sensitivities and specificities, along with interval estimates to show uncertainty. Examining robustness to the prior mean for σ^2 we found that large changes did change the width of the interval estimates for sensitivity, but only slightly.

One can compare the predicted performance of two radiologists by looking at whether or not their intervals overlap. Comparisons done in this way show many pairs of radiologists who have significantly different predicted specificity. For the case of sensitivity, the intervals are much wider (the scales differ for sensitivity and specificity in the figure), so that almost all of the sensitivity intervals overlap.

One can also employ a high-risk or a low-risk combination of patient risk factors to see how changing the patient risk factors changes a radiologist's predicted sensitivity and specificity. For sensitivity, the lowest-risk hypothetical mammogram (the mammogram with the lowest risk of a cancer being missed) differs from the "typical" mammogram by being in registry B and for a woman aged 50-59, with no previous mammogram, and with breast density classification 1. For sensitivity the highest-risk hypothetical mammogram differs from the "typical" mammogram by being for a woman with extremely dense breasts. For specificity the lowest-risk hypothetical mammogram differs from the "typical" mammogram by being for a woman who is post-menopausal with hormone therapy status unknown and breast density classification 1. For specificity the highest-risk hypothetical mammogram differs from the "typical" mammogram

by being for a woman who has breast cancer symptoms and a history of biopsy or surgery, with no previous mammogram and heterogeneously dense breasts. The range of predicted sensitivity and specificity on these patient profiles are shown in Table IV.

There is a large spread in the expected radiologist sensitivity on the high-risk combination of patient risk factors, as shown in Figure 2. In contrast, there is a much smaller spread of expected radiologist sensitivity on the low-risk combination of patient risk factors, which is also shown. Therefore variation among radiologists in sensitivity is more of a concern for high-risk patients.

We also estimated radiologist performance relative to an appropriate standard, as described in section 4. Figure 3 shows performance relative to a standard for the same set of fifteen radiologists chosen for Figure 1. The interval estimates for all of the radiologists in the data sets contain zero. This means that there is no evidence that any of the radiologists have lower or higher sensitivity than their standard. This may be a result of the small sample sizes that are available to learn about sensitivity. Larger sample sizes are available to study specificity, and indeed narrower posterior intervals are witnessed. This is also shown in Figure 3; note that the scales are different for sensitivity and specificity. Many of the intervals for the specificity differences are completely above or below zero; 12 out of 120 (10.0%) of the radiologists have specificity differences significantly above zero, and 10 (8.3%) have specificity differences significantly below zero. A specificity difference above zero indicates that a radiologist is performing better than is expected given her/his patient mix and radiologist attributes. A specificity difference below zero indicates that the radiologist is underperforming compared to a hypothetical standard radiologist with the same case mix and attributes. Overall, the estimated

sensitivity differences range from -11.3% to 6.1%. The estimated specificity differences range from -8.8% to 5.5%.

6. A PREDICTION PROBLEM

In this section we address the question of assessing the performance of a new radiologist not originally included in our analysis. That is, if we are provided with a new set of data $\mathcal{D}_0$, consisting of case mix $\mathbf{X}_0$, outcomes $\{C_{0j}\}$ and $\{R_{0j}\}$ and radiologist attributes $\mathbf{W}_0$, can we infer the performance metrics for this new individual?

To address this problem, we could go back and refit the entire model adding in the new radiologist data. That is, we would merely add another product term at $i = 0$ to the likelihood in (5). Illustrating for specificity, an alternative way to learn about the new $\tau_0^{(0)}$ that does not require refitting the model is as follows. Employ the posterior draws from $[\beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)}|\mathcal{D}^{(0)}]$ and simply draw $\tau_0^{(0)}$ from $[\tau_0^{(0)}|\mathbf{X}_0, \beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)}, \mathcal{D}_0] \propto \left[\prod_{j=1}^{n_0} Pr(R_{0j}|C_{0j} = 0, \beta^{(0)}, \tau_0^{(0)})\right] \mathcal{N}(\tau_0^{(0)}; \mathbf{W}_0'\gamma^{(0)}, \sigma^{2(0)})$. This approximation presumes that $[\beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)}|\mathcal{D}^{(0)}] \approx [\beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)}|\mathcal{D}^{(0)}, \mathcal{D}_0]$.

To sample $[\beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)}|\mathcal{D}^{(0)}, \mathcal{D}_0]$ exactly, one can use importance sampling. Straightforward calculation reveals that the sampling weights are calculated with only a univariate integration as follows.

$$w(\beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)}) \propto \int \left[\prod_{j=1}^{n_0} Pr(R_{0j}|C_{0j} = 0, \beta^{(0)}, \tau_0^{(0)})\right] \mathcal{N}(\tau_0^{(0)}; \mathbf{W}_0'\gamma^{(0)}, \sigma^{2(0)}) d\tau_0^{(0)} \quad (11)$$

With either the approximate or exact posterior samples of $\beta^{(0)}, \gamma^{(0)}, \sigma^{2(0)}$, and $\tau_0^{(0)}$ sampled

as above, we can estimate any of the performance metrics developed in section 4 for the new radiologist.

7. MULTIPLE PERFORMANCE RATES

In the screening setting there may be interest in a broader scope of performance rates than just sensitivity and specificity. In the introduction we alluded to several possibilities; here we take this idea a bit further. We also briefly comment on the use of utility ideas to summarize and compare these multiple performance rates.

In some settings a positive response to a screening procedure arises by obtaining a level that exceeds a specified threshold. In the mammography context, the output of the screening mammogram is actually a BI-RADS^{TM} score from 0 to 5 (see [1]) and we have defined a positive screen as one attaining a suitably high BI-RADS^{TM} score. One could redefine a positive screen by changing the required score. In a different setting, the threshold might be chosen along a continuous scale. Then the performance metrics of section 4 would change according to this choice, leading to a continuum of performance metrics, one for each threshold. Plotting the value of the metric versus the threshold value yields a performance curve of, for example, case mix averaged sensitivity relative to a standard. In this manner we could obtain a curve for each radiologist. Such curves may be compared with simple overlays, and such visual comparison may be more satisfying than reduction to a single number. If we seek the latter in order to do the performance comparison, these curves could be reduced to a summary performance metric. For example, specificity can be integrated with respect to threshold z. Perhaps the integration is over $z \in Z$, where Z is a bounded set, or perhaps integration is with respect to a proper

distribution for z. Such integration would be most easily carried out through a Monte Carlo approximation.

In fact, for an ROC setting, for each radiologist we would have defined a relationship between sensitivity and specificity through, say, the parametric functions of z, $\overline{se}(z)$ and $\overline{sp}(z)$. Computing the customary area under the ROC curve [10] provides a new performance metric for a particular radiologist, one that balances both sensitivity and specificity. Such areas will, again, be most easily computed using Monte Carlo integration.

The area under the ROC curve illustrates a particular choice of utility function. A simpler utility function can be used in the special case when considering only one threshold for sensitivity and specificity. This could be, for instance, some arithmetic combination of that sensitivity and specificity, such as a suitable weighted average.

8. CONCLUSIONS AND FUTURE WORK

Building on methodology developed by Normand et al. [9], we have provided an approach for assessing the performance of physicians who interpret screening tests. Evidently, this is an objective that is distinct from assessing the effectiveness of screening test procedures themselves, though the former certainly informs about the latter. We have shown how performance can be addressed individually as well as comparatively.

In particular, we assess radiologists interpreting screening mammography. Previous studies have established that there is significant variability among radiologists in empirical sensitivity

and specificity, even after accounting for differences among radiologists in patient mix. We assess the performance of individual radiologists using patient-level models for sensitivity and specificity. We also attempt to explain the performance differences among the radiologists by incorporating information on radiologist attributes into the models. This radiologist information comes from survey data.

Using these models, we estimate the sensitivity and specificity of each radiologist for a hypothetical mammogram that has a particular combination of patient risk factors. One can use such predictions to compare one radiologist to another. Radiologists are also compared individually against a standard which is appropriate for them. Many radiologists are identified who have specificity significantly above or below their standard. No radiologists are found to have sensitivity significantly above or below their standard. This may be due to the much smaller sample sizes used to estimate sensitivity than specificity.

Such performance evaluation can be used to give feedback to radiologists, and, in future work, we plan to do so through a web-based continuing medical education program. Future work will also pursue an analogous investigation of diagnostic mammography. We also will consider profiling associated with other screening procedures, for example cardiac (ECHO) screening, clinical breast examination, MRI and ultrasound. In addition, we will analyze the effect on performance of the technique of "double reading" of mammograms, in which two radiologists separately analyze each mammogram.

On a more technical level, here we are modeling sensitivity and specificity separately.

However, one might suspect that these are correlated because more cautious radiologists will have higher sensitivity and lower specificity. To investigate this one could jointly model screening test result and disease outcome and then calculate sensitivity and specificity as resultant parametric functions under this model. An alternative error specification to the normal distribution in the modeling of the radiologist random effects could also be considered.

Finally, since section 7 opens up the possibility of different utility functions in assessing performance, an attractive exercise would be to attempt to elicit a suitable utility function. How would a radiologist envision assessing her/his performance? How could this be converted to a valid utility? Perhaps it would be a single metric combining sensitivity and specificity, or combining recall rate and positive predictive value. A patient might have a different utility function, as would a regulator. Resulting posterior utilities could be studied and compared for the radiologists.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. Patricia Carney, Dr. Stephen Taplin and the participating radiologists. The authors also appreciate the comments of the three referees, which have resulted in an improved presentation of the material.

REFERENCES

1. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS^{TM}). Reston, VA, 1998.
2. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi CJ, Cutter G, Hendrick RE, Elmore JG. Accuracy of

screening mammography interpretation by characteristics of radiologists. *Journal of the National Cancer Institute* 2004; **96**:1840–1850.

3. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammography interpretations in community-based programs. *Journal of the National Cancer Institute* 2003; **95**:1384–1393.
4. Mammography Quality Standards Act. Publication L, October 1992; pp. 102–539.
5. Landon BE, Normand ST. Physician clinical performance assessment: Prospects and barriers. *Journal of the American Medical Association* 2003; **290**:1183–1212.
6. Iezzoni LI, Greenberg LG. Widespread assessment of risk-adjusted outcomes: lessons from local initiatives. *Joint Commission Journal on Quality Improvement* 1994; **20**:305–316.
7. Landon BE, Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. *Inquiry* 1996; **33**:155–166.
8. Salem-Schatz S, Morre G, Rucker M, Pearson SD. The case for case-mix adjustment in practice profiling: when good apples look bad. *Journal of the American Medical Association* 1994; **272**:871–874.
9. Normand ST, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association* 1997; **92**:803–814.
10. Rosner B. *Fundamentals of Biostatistics* (5th edn). Duxbury, 2000; pp. 63–65.
11. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, Barlow WE, Geller K, Kerlikowske BM, Edwards BK, Lynch CF, Urban N, Chrvala CA, Key CR, Poplack SP, Worden JK, Kessler LG. Breast Cancer Surveillance Consortium: A national mammography screening and outcomes database. *American Journal of Roentgenology* 1997; **169**:1001–1008.
12. Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, Geller BM, Abraham LA, Taplin SH, Dignan M, Cutter G, Ballard-Barbash R. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Annals of Internal Medicine* 2003; **138**:168–175.
13. Paliwal P, Gelfand AE, Abraham L, Barlow WE, Elmore JG. Examining accuracy of screening mammography using an event order model. *Statistics in Medicine* 2006; **25**:267–283.
14. Gelfand AE, Smith AF. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**:398–409.
15. Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* 1992; **41**:337–348.

APPENDIX

Full-conditional distributions

As noted in the text, the prior distributions of the parameters are identical for the two models and are as follows (suppressing the model indicators). All parameters have independent prior distributions unless noted otherwise. Here we use the bracket notation of [14] to denote densities.

$$[\beta] = \mathcal{N}(\ \beta_0,\ \sigma_\beta^2 \cdot \mathbf{I}_p) \tag{12}$$

$$[\tau_i|\sigma^2] = \mathcal{N}(\ \mathbf{W}_i'\gamma,\ \sigma^2) \tag{13}$$

$$[\gamma] = \mathcal{N}(\ \gamma_0,\ \sigma_\gamma^2 \cdot \mathbf{I}_q) \tag{14}$$

$$[\sigma^2] = \mathcal{IG}(a, b) \tag{15}$$

Here σ_β^2 and σ_γ^2 are constants, p is the length of the vector β, q is the length of the vector γ, and $\mathbf{I}$ is the identity matrix. Using these prior distributions for the parameters, the full-conditional posterior distributions take the following form.

$$[\beta|\tau,\gamma,\sigma^2,\mathcal{D}] \quad \propto \quad \mathcal{N}(\ \beta_0,\ \sigma_\beta^2 \cdot \mathbf{I}_p) \times \textstyle\prod_{i=1}^{I} \left[\prod_{j=1}^{n_i} Pr(R_{ij}|C_{ij},\beta,\tau_i)\right] \tag{16}$$

$$[\tau_i|\beta,\gamma,\sigma^2,\mathcal{D}] \quad \propto \quad \mathcal{N}(\ \mathbf{W}_i'\gamma,\ \sigma^2) \times \textstyle\prod_{j=1}^{n_i} Pr(R_{ij}|C_{ij},\beta,\tau_i) \tag{17}$$

$$[\gamma|\beta,\tau,\sigma^2,\mathcal{D}] \quad = \quad \mathcal{N}(\ \mathbf{V}(\tfrac{\gamma_0}{\sigma_\gamma^2} + \tfrac{\mathbf{W}'\tau}{\sigma^2}),\ \mathbf{V}) \tag{18}$$

$$[\sigma^2|\beta,\tau,\gamma,\mathcal{D}] \quad = \quad \mathcal{IG}(\ a + \tfrac{I}{2},\ b + \tfrac{1}{2}\textstyle\sum_{i=1}^{I}(\tau_i - \mathbf{W_i'}\gamma)^2) \tag{19}$$

Here $\mathcal{D}$ indicates the data ($\{R_{ij}\}$ and $\{C_{ij}\}$), and $\mathbf{V} = (\frac{\mathbf{I}_q}{\sigma_\gamma^2} + \frac{\mathbf{W}'\mathbf{W}}{\sigma^2})^{-1}$. Note that the τ_i are conditionally independent from each other a posteriori.

Table I. Summary of the radiologist attributes. Sensitivity and specificity are both calculated with equal weight given to each of the mammograms, so that radiologists with more mammograms contribute more to the performance rates.

Radiologist Attribute	# Radiologists	Sensitivity (%)	Specificity (%)
NO. YEARS INTERPRETING MAMMOGRAMS			
< 10	26	86.0	88.8
10-19	56	82.8	89.4
20+	38	78.8	91.6
NO. OF MAMMOGRAMS INTERPRETED			
≤ 1000	31	73.3	92.6
1001-2000	45	82.7	89.1
> 2000	44	84.2	89.9
ACADEMIC AFFILIATION			
Primary appointment	6	78.0	93.9
Adjunct appointment	12	83.6	89.4
None	102	82.5	89.8
ALL RADIOLOGISTS	120	82.4	90.0

Table II. Summary of patient risk factors in the screening mammography data

Patient Risk Factor	# Patients	# with Cancer	# with Positive Screen	Sensitivity (%)	Specificity (%)
AGE					
40-49	128591	388	15326	74.7	88.3
50-59	123169	572	13314	84.8	89.5
60-69	75316	513	7091	84.2	91.1
70+	71938	570	5712	83.5	92.7
FAMILY HISTORY OF BREAST CANCER					
No	337578	1638	35130	82.5	89.9
Yes	61436	405	6313	81.7	90.2
HISTORY OF BREAST BIOPSY/SURGERY					
No	325142	1524	32321	82.4	90.4
Yes	73872	519	9122	82.3	88.1
BREAST CANCER SYMPTOMS					
No symptom	385205	1860	39186	83.0	90.2
Lump/nipple discharge/other	13809	183	2257	76.5	84.5
MENOPAUSE/HT* STATUS					
Pre-menopausal	121067	431	14719	77.5	88.1
Post-menopausal, no HT	127760	729	11001	85.5	91.8
Post-menopausal, HT	124751	739	13648	82.9	89.5
Post-menopausal, unknown HT	25436	144	2075	78.5	92.2
PREVIOUS MAMMOGRAM					
No	17542	107	3053	87.9	83.0
Yes	381472	1936	38390	82.1	90.3
BREAST DENSITY					
1. Almost entirely fatty	43696	127	2040	91.3	95.6
2. Scattered fibroglandular tissue	174763	818	16073	86.9	91.2
3. Heterogeneously dense	147575	899	19322	80.6	87.3
4. Extremely dense	32980	199	4008	65.8	88.2
ALL MAMMOGRAMS	399014	2043	41443	82.4	90.0

* HT= Hormone Therapy

Table III. Mixed effects modeling of sensitivity and specificity by patient and radiologist attributes. Both point and 95% credible interval estimates are shown. Odds ratios are estimated jointly for all the patient and radiologist attributes. A "+" indicates that the interval for the odds ratio is entirely above one. A "-" indicates that the interval for the odds ratio is entirely below one, while a "0" indicates that the interval for the odds ratio contains one.

	Sensitivity OR of mean	(LCL**-UCL**)		**Specificity** OR of mean	(LCL**-UCL**)	
Patient Risk Factors						
Age						
40-49	referent			referent		
50-59	**1.69**	**(1.10 - 2.62)**	+	0.96	(0.86 - 1.08)	0
60-69	1.46	(0.88 - 2.40)	0	1.03	(0.89 - 1.19)	0
70+	1.23	(0.75 - 1.99)	0	1.17	(1.00 - 1.36)	0
Family History						
No	referent			referent		
Yes	0.91	(0.67 - 1.24)	0	1.02	(0.91 - 1.14)	0
History of Biopsy/Surgery						
No	referent			referent		
Yes	1.12	(0.84 - 1.48)	0	**0.74**	**(0.67 - 0.82)**	-
Breast Symptoms						
No symptom	referent			referent		
Lump/nipple discharge	0.68	(0.46 - 1.02)	0	**0.61**	**(0.52 - 0.73)**	-
Menopause/HT* Status						
Pre-menopausal	0.99	(0.62 - 1.55)	0	0.98	(0.87 - 1.11)	0
Post-menop., no HT	1.04	(0.76 - 1.43)	0	**1.15**	**(1.04 - 1.28)**	+
Post-menop., HT	referent			referent		
Post-menop., unknown HT	0.81	(0.50 - 1.31)	0	**1.35**	**(1.12 - 1.62)**	+
Previous Mammogram						
No	**1.87**	**(1.01 - 3.65)**	+	**0.55**	**(0.47 - 0.65)**	-
Yes	referent			referent		
Breast Density						
1	**1.93**	**(1.01 - 3.95)**	+	**2.20**	**(1.83 - 2.64)**	+
2	referent			referent		
3	**0.56**	**(0.42 - 0.75)**	-	**0.82**	**(0.75 - 0.89)**	-
4	**0.26**	**(0.18 - 0.39)**	-	0.96	(0.82 - 1.11)	0
Radiologist Attributes						
No. Yrs Interpreting Mammograms						
< 10	**1.67**	**(1.07 - 2.65)**	+	**0.74**	**(0.59 - 0.91)**	-
10-19	referent			referent		
20+	0.76	(0.52 - 1.10)	0	**1.30**	**(1.07 - 1.58)**	+
No. of Mammograms Interpreted						
≤ 1000	**0.60**	**(0.37 - 0.97)**	-	**1.38**	**(1.11 - 1.72)**	+
1001 - 2000	referent			referent		
> 2000	1.17	(0.83 - 1.65)	0	**1.24**	**(1.03 - 1.50)**	+
Academic Affiliation						
Primary appointment	0.48	(0.24 - 1.00)	0	**2.22**	**(1.44 - 3.44)**	+
Adjunct appointment	0.93	(0.57 - 1.54)	0	0.86	(0.65 - 1.14)	0
None	referent			referent		

* HT= Hormone Therapy

** LCL= Lower Confidence Limit, UCL = Upper Confidence Limit

Table IV. Range of predicted radiologist sensitivity and specificity on several patient profiles. For each patient profile the range of expected sensitivity over the radiologists is shown. The expected sensitivity and 95% credible interval for the radiologist with the lowest expected sensitivity are shown, as are those for the radiologist with the highest expected sensitivity. The range of expected specificity is shown similarly.

	Sensitivity(%)		**Specificity(%)**	
	Expected	(LCL*-UCL*)	Expected	(LCL*-UCL*)
"TYPICAL" PATIENT				
Lowest expected value	58.9	(35.9-79.8)	83.5	(78.8-87.8)
Highest expected value	91.8	(81.9-97.3)	96.9	(95.0-98.3)
"HIGH-RISK" PATIENT				
Lowest expected value	28.1	(11.7-50.7)	47.6	(37.9-57.7)
Highest expected value	74.8	(52.7-90.7)	84.7	(76.4-91.3)
"LOW-RISK" PATIENT				
Lowest expected value	91.6	(78.1-98.1)	92.8	(89.9-95.1)
Highest expected value	98.9	(96.8-99.8)	98.7	(97.9-99.3)

* LCL= Lower Confidence Limit, UCL = Upper Confidence Limit

Figure 1. Predicted radiologist sensitivity (top) and specificity (bottom) with "typical" patient risk factors, by radiologist. A randomly selected subset of fifteen radiologists are displayed. Dots are the expected values, while the vertical bars show 95% credible intervals. Note that the scales are different for the two plots.

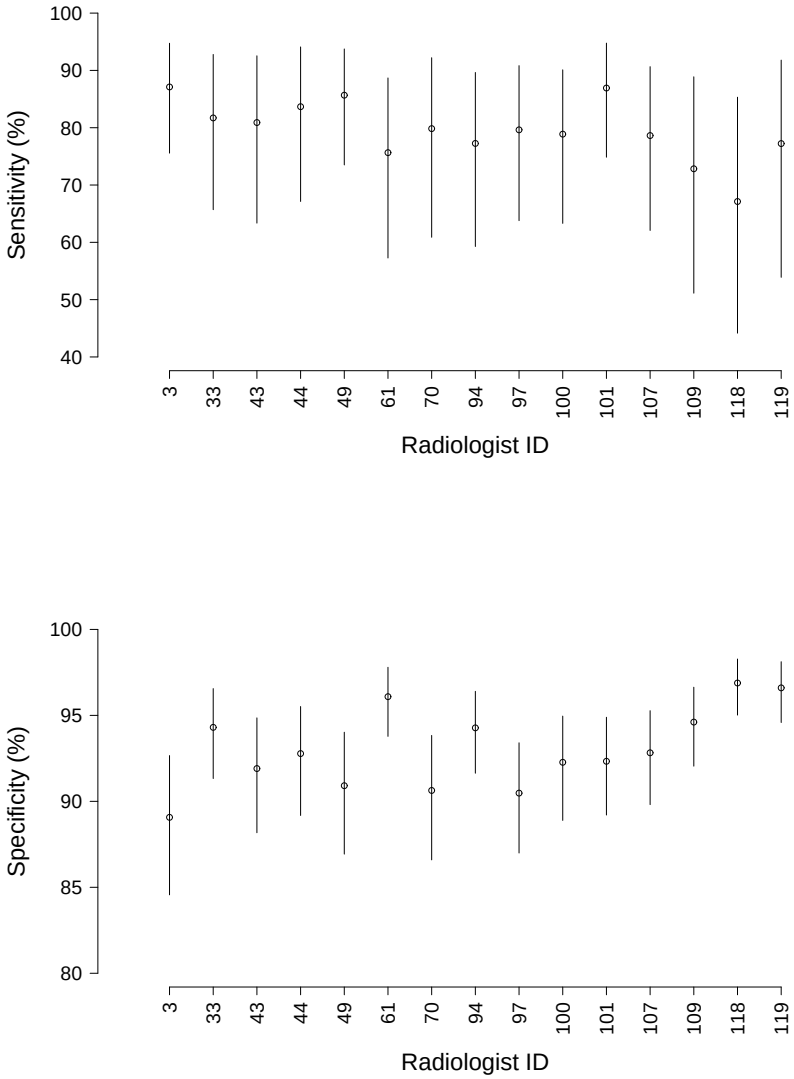
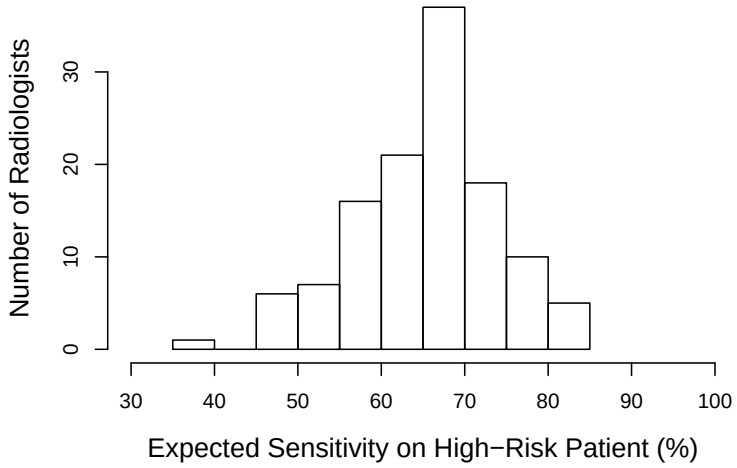


Figure 2. Predicted radiologist sensitivity on "high-risk" (top) and "low-risk" (bottom) patients.



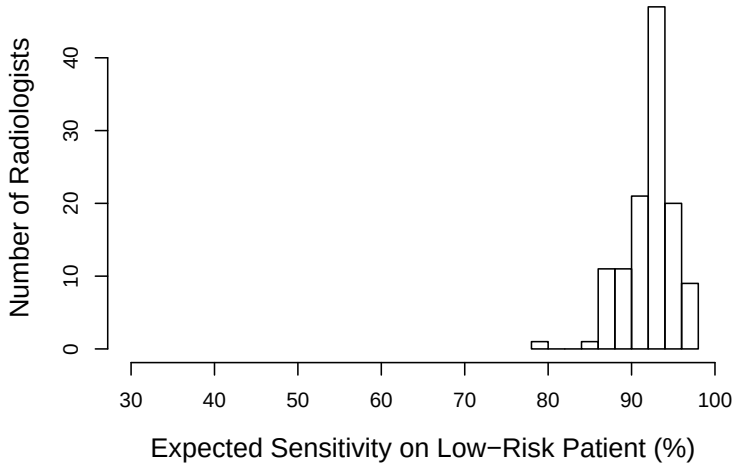


Figure 3. Difference between radiologist sensitivity/specificity and a standard. This performance metric is shown for the same subset of radiologists as in Figure 1. The top graph shows sensitivity differences, while the bottom graph shows specificity differences. Dots show the radiologists' estimated performance, while the vertical bars give the 95% credible interval for the performance of the radiologist. Note that the scales are different for the two plots.