Published in final edited form as:
J Natl Cancer Inst. 2004 December 15; 96(24): 1840–1850. doi:10.1093/jnci/djh333.

Accuracy of Screening Mammography Interpretation by Characteristics of Radiologists

William E. Barlow, **Chen Chi**, **Patricia A. Carney**, **Stephen H. Taplin**, **Carl D'Orsi**, **Gary Cutter**, **R. Edward Hendrick**, and **Joann G. Elmore**
Cancer Research and Biostatistics, Seattle, WA (WEB); Center for Health Studies, Group Health Cooperative, Seattle, WA (WEB, SHT); Department of Biostatistics, University of Washington, Seattle, WA (WEB); Fred Hutchinson Cancer Research Center, Seattle, WA (CC); Office of Medical Education, Dartmouth University, Hanover, NH (PAC); Department of Radiology, Emory University, Atlanta, GA (CD); Center for Research Design and Statistical Methods, University of Nevada, Reno, NV (GC); Lynn Sage Comprehensive Breast Center, Department of Radiology, Northwestern University Feinberg School of Medicine, Chicago, IL (REH); Department of Internal Medicine, University of Washington School of Medicine, Seattle, WA (JGE)

Abstract

Background—Radiologists differ in their ability to interpret screening mammograms accurately. We investigated the relationship of radiologist characteristics to actual performance from 1996 to 2001.

Methods—Screening mammograms (n = 469 512) interpreted by 124 radiologists were linked to cancer outcome data. The radiologists completed a survey that included questions on demographics, malpractice concerns, years of experience interpreting mammograms, and the number of mammograms read annually. We used receiver operating characteristics (ROC) analysis to analyze variables associated with sensitivity, specificity, and the combination of the two, adjusting for patient variables that affect performance. All *P* values are two-sided.

Results—Within 1 year of the mammogram, 2402 breast cancers were identified. Relative to low annual interpretive volume (≤1000 mammograms), greater interpretive volume was associated with higher sensitivity (*P* = .001; odds ratio [OR] for moderate volume [1001–2000] = 1.68, 95% CI = 1.18 to 2.39; OR for high volume [>2000] = 1.89, 95% CI = 1.36 to 2.63). Specificity decreased with volume (OR for 1001–2000 = 0.65, 95% CI = 0.52 to 0.83; OR for more than 2000 = 0.76, 95% CI = 0.60 to 0.96), compared with 1000 or less (*P* = .002). Greater number of years of experience interpreting mammograms was associated with lower sensitivity (*P* = .001), but higher specificity (*P* = .003). ROC analysis using the ordinal BI-RADS interpretation showed an association between accuracy and both previous mammographic history (*P* = .012) and breast density (*P*<.001). No association was observed between accuracy and years interpreting mammograms (*P* = .34) or mammography volume (*P* = .94), after adjusting for variables that affect the threshold for calling a mammogram positive.

Conclusions—We found no evidence that greater volume or experience at interpreting mammograms is associated with better performance. However, they may affect sensitivity and specificity, possibly by determining the threshold for calling a mammogram positive. Increasing volume requirements is unlikely to improve overall mammography performance.

Correspondence to: William Barlow, PhD, Cancer Research and Biostatistics, 1730 Minor Ave, Ste. 1900, Seattle WA 98101 (williamb@crab.org).

A meta-analysis of randomized trials of screening mammography has demonstrated that screening mammography can reduce mortality from breast cancer by approximately 15% in women 40 to 74 years of age (1). Consequently, the United States Preventive Services Task Force recommends screening mammography for women age 40 years and above every 1 to 2 years (2). However, for the expected benefits of screening to be obtained, the elements in the screening process must perform as expected (3). In particular the accuracy of screening mammography interpretation in community radiology practice must be maximized. In addition, many factors that affect accuracy must be considered, including characteristics of the radiologist as well as those of the woman having the mammogram.

The accuracy of mammographic interpretation among individual radiologists varies widely (4,5). One study showed a 40% disparity among radiologist screening sensitivity and a 45% range in the rates at which women without breast cancer are recommended for biopsy (5). As indicated by receiver operating characteristic (ROC) analyses, the ability of radiologists to detect cancer varies by as much as 11% (5).

The source of radiologist variability has not been completely explained. Two studies (6,7) showed that experience and training of the radiologist could affect the accuracy of mammography; however, this conclusion was based on small samples in test settings. Younger and more recently trained radiologists had higher false-positive rates (8). The number of mammograms performed, i.e., mammographic volume, has also been cited as a putative agent affecting accuracy. Esserman et al. (9) found an increase in both sensitivity and specificity with higher volumes, although Beam et al. (10) did not. McKee et al. (11) also found no association between positive biopsy rates and radiologist volume.

Because patient variation can affect the performance measures for a particular radiologist, it is necessary to adjust for important determinants of accuracy in the patient population before assessing radiologist accuracy. Three important determinants of accuracy for patients include age, breast density, and whether the mammogram is a first or subsequent mammogram. Sensitivity increases with age, but specificity is only moderately associated with age (12–14). Denser breasts may obscure breast tumors and make interpretation more difficult, resulting in decreased sensitivity and specificity (15). Fajardo (16) has demonstrated that a radiologist's certainty of interpretation of a mammogram is inversely related to breast density and complexity of the image. This uncertainty may lead to a greater number of "abnormal" readings and increase the false positive rate. Having had a prior mammogram is also associated with increased specificity but decreased sensitivity (17). Because tumors may be smaller on repeat examination than on initial screens (17a), sensitivity may decrease even when earlier films are available. However, in the absence of a tumor, evaluating previous screening films may help the radiologist rule out tumors (17).

We conducted a study to assess the relationship between self-reported volume, experience, and other radiologist characteristics and their recorded performance in community settings after accounting for patient factors. To do this, we measured variability in radiologist performance using recall rate, sensitivity, specificity, and ROC area under the curve with adjustment for patient variation.

Subjects and Methods

Study Population

Seven mammography registries participate in the Breast Cancer Surveillance Consortium (BCSC) (http://breastscreening.cancer.gov) funded by the National Cancer Institute (18). Three of the registries received additional funding to conduct a survey of radiologists to link the survey responses to actual mammography performance as recorded in the BCSC

database. Two registries are geographically based—the New Hampshire Mammography Network captures approximately 90% of women undergoing mammography in New Hampshire (19), and the Colorado Mammography Program includes approximately 50% of the women undergoing mammography in the Denver metropolitan area. The third mammography registry, the Breast Cancer Surveillance Project at Group Health Cooperative, is based on a nonprofit health maintenance organization that has a defined population in the Pacific Northwest. The three registries reflect diversity in geography, mode of delivery of health care, and size, and, therefore, represent a broad spectrum of mammography in the United States.

Mammography outcome data were available from the BCSC for the period January 1996 through December 2001. Radiologists who interpreted mammograms anytime in that calendar period were included. A radiologist must read 960 mammograms over a 2-year period to be considered eligible to interpret mammograms at an accredited mammography facility in the United States (20). Hence, we included only radiologists who interpreted at least 480 mammograms (the required annual average) during the study period. We recruited radiologists already participating in the BCSC by mail, with telephone follow-up for non-responders. Radiologists were informed that their survey responses would be linked to their actual performance results by an encrypted linkage variable and that their identity would remain anonymous. De-identified mammographic data and radiologist survey data were sent to the BCSC's Statistical Coordinating Center for analysis. All study activities were approved by the institutional review boards of Group Health Cooperative (Washington), Dartmouth College (New Hampshire), the Cooper Institute (Colorado), and the University of Washington School of Medicine. Patients included in the New Hampshire Mammography Network have signed informed consent statements. The two remaining sites allow patients to exclude their data from research, but the institutional review boards do not require informed consent for primarily clinical activity. All radiologists gave informed consent for the survey and the linkage to their performance data.

Measured Covariates and Outcomes

The radiologist survey included questions on demographic and clinical variables. The survey was extensively pilot tested (21) before its use in this study. Data from the mammography registry included patient demographics and risk factors, indication for the mammogram (screening versus diagnostic), screening interpretation and recommendations, and breast cancer outcome.

The covariates considered in the analysis included both radiologist covariates suspected to influence screening interpretations and patient factors known to affect breast cancer risk and screening accuracy. The radiologist factors considered were demographic characteristics (age, sex), clinical practice characteristics (number of years in practice, number of mammograms interpreted per year, percentage of time working in breast imaging, percentage of mammograms that were screening, and whether affiliated with an academic medical center), malpractice history (ever had a mammography-related claim, type of medical malpractice insurance), and concerns on malpractice (whether concerned about malpractice when interpreting a mammogram and if malpractice influences recommendation in assessment). Data from the survey were collected in defined categories and analyzed by each category unless small sample sizes required collapsing adjacent categories. All categorization was performed prior to analysis.

We included mammograms from women 40 years of age and older who did not have a personal history of breast cancer. The mammogram had to be designated as bilateral screening by the radiologist and needed to have occurred at least 9 months after any preceding breast imaging to avoid misclassifying a diagnostic examination as screening.

Each screening mammogram was classified into one of six assessment categories according to the initial BI-RADS interpretation code: 0) need additional imaging, 1) negative, 2) benign finding, 3) probably benign finding, 4) suspicious abnormality, and 5) highly suggestive of malignancy (22). The action taken after each of these assessments can vary from biopsy, surgical evaluation, additional imaging, or short-term follow-up. For the purpose of assessing recall, sensitivity, and specificity, we classified code 0, 4, and 5 assessments as positive interpretations and code 1 and 2 assessments as negative interpretations. Code 3 assessment was classified as positive if immediate work-up was recommended and negative if it was not. BI-RADS suggest short-term follow-up screening (e.g., 6 months later) after an assessment of "probably benign," but in practice it is common to continue work-up immediately with additional imaging or other procedures (23). The ROC analyses used the assessments codes as an ordinal response ordered as follows: 1, 2, 3 with no immediate work-up; 3 with immediate work-up; and 0, 4, and 5 reflecting the increasing likelihood of cancer.

Patient characteristics we considered were breast density, previous mammography, and age at the time of the screening mammogram. Age was categorized as 40–49, 50–59, 60–69, 70–79, and 80 years or greater corresponding to age deciles typically used in analyses of mammography. Breast density was determined by the radiologist using four categories based on the BI-RADS coding system: 1) almost entirely fat, 2) scattered fibroglandular densities, 3) heterogeneously dense, and 4) extremely dense (22). Previous mammography was based on self-report and/or data in the mammography registry indicating a prior examination.

The follow-up interval for cancer assessment was 1 year after the screening mammogram or until the next screening examination if more than 9 months after the preceding screening examination, whichever occurred first. (Rosenberg et al. (9) show the effect of alternative definitions of the follow-up period on sensitivity and specificity). Breast cancer case patients were identified through the cancer registry in their geographic area or pathology data and included patients with both invasive carcinoma and ductal carcinoma *in situ* (DCIS). Lobular carcinoma *in situ* (LCIS) was not included as a breast carcinoma.

Statistical Analysis

The analysis evaluated sensitivity, specificity, and recall rate as functions of both the patients' and radiologist's characteristics. Sensitivity, also called the true positive rate, was defined as the proportion of screening examinations judged positive among all patients with a diagnosis of breast cancer within the 1-year follow-up interval. Specificity was defined as the proportion of screening examinations judged negative among all patients who did not have a diagnosis of breast cancer in the follow-up period. Therefore, the false positive rate = 1 – specificity. Overall accuracy of mammography was assessed using a ROC curve, which plots true positive rate against false positive rate and allows alternative definitions of a positive or negative examination at each BI-RADS assessment code. The area under the ROC curve was interpreted as the probability that the radiologist will correctly choose the mammogram that contains cancer when presented with two mammograms, one with cancer and the other without. Recall rate was defined as the proportion of all screening mammograms judged positive.

The data were hierarchical such that mammograms read by an individual radiologist are considered "nested" under that particular radiologist. In turn, each radiologist works for a mammography facility in one of the three mammography registries. Screening outcomes for the same radiologist may therefore not be statistically independent. Naive analysis assuming independence gives valid parameter estimates but possibly incorrect (i.e., underestimated) standard errors. We considered two approaches to account for the correlation of

mammographic outcomes within radiologist. A generalized estimating equations (GEE) approach is based on a marginal model assuming a working correlation structure to adjust the standard errors for the correlation (24) and provides valid inference without requiring that the correlation structure be correctly specified. We have used this method previously to adjust for correlation of mammographic outcomes within radiologists when radiologist variability was not being directly analyzed (25). Alternatively, a random-effects approach assumes that the conditional mean of the outcome depends on a radiologist-specific effect, the common sharing of which introduces correlation among the outcomes of that radiologist. In addition to providing valid inference on random-effect-specific estimates, a random-effects model can also help clarify the source of variation and allows estimation of the variability of the random effects. It is necessary to assume, however, that mammographic outcomes within the radiologist's practice are conditionally independent given the random effect and radiologist-specific covariates. We chose the random effects approach instead of a GEE approach because we were most interested in quantifying the impact of the radiologist on the outcome. All *P* values are two-sided. Categorical variables were tested by assessing the joint statistical significance against the referent unless a test for trend was specifically mentioned. Statistical significance was defined as $P<.05$. Because of the number of statistical comparisons being performed, *P* values should be interpreted cautiously.

Our aim was to identify radiologist factors that contribute to inter-radiologist variability in outcomes. For sensitivity, specificity, and recall, we used a logistic regression model to link the mammographic outcome to patient and radiologist covariates. We first used a random-effects model with each radiologist having a separate intercept, and we subsequently fitted a mixed-effects model incorporating patient and radiologist covariates as fixed effects as well as a radiologist-specific random intercept. Model fitting was performed using SAS (version 8; Cary, NC) procedure NLMIXED, which allows estimation of a normally distributed random effect in a logistic regression model (26). The software provides Wald tests of the covariate effects as well as the estimated variance of the random effect. We then examined the radiologist factors that might explain the variation among radiologists, estimated the variability before and after incorporating radiologist fixed effects, and statistically determined how they affected the variability. Because our main focus was to examine the source of variability at the radiologist level, we included the three mammography registries as fixed effects in the model but did not account for additional variation due to different facilities within each of the three registries. Hierarchical modeling of facility, radiologist, and mammogram could allow greater separation of the individual sources of variation but would require complex models that are more difficult to interpret.

ROC Analysis

One way to increase sensitivity is to lower the threshold for calling a mammogram positive. This shift in the threshold will decrease specificity but does not reflect an increased ability to discriminate between mammograms with cancer and those without. ROC analysis can be used to separately estimate effects due to threshold values from effects on accuracy (27). Each point on the ROC curve represents a true positive rate (sensitivity) and false positive rate (1 – specificity) pair obtained by applying a particular threshold (the value above which a mammogram is classified as positive). Changing the threshold without changing the overall accuracy will move the radiologist along the same ROC curve, trading off specificity for sensitivity. A true change in accuracy will lead to a different ROC curve altogether. The measure of interest is the area under the ROC curve (AUC), with larger areas indicating greater accuracy. The maximum value of AUC is 1.0 and indicates a perfect test; an AUC of 0.50 indicates a test that performs no better than chance. The advantage of using ROC analysis in evaluating test accuracy compared with separate analyses of sensitivity and

specificity is that ROC provides an index of overall test accuracy that does not require the selection of a single cutoff and can separate threshold from discrimination (accuracy).

ROC models can be fit using ordinal regression models that accommodate covariates that affect either the threshold or the location of the ROC curve (28). Without covariates, the model reduces to the usual binormal model (27,29). The ROC model was extended to fit random effects (30–33). Two random effects were fit for each radiologist—one adjusted for their likelihood of calling the mammogram positive (threshold) and the second measured their ability to discriminate disease from nondisease (accuracy). The nonlinear mixed ordinal regression model was fit using the SAS procedure NLMIXED (26). Likelihood ratio tests were used to test effects of covariates on accuracy in the ROC curves. We allowed the scale parameter to differ between patients with and without breast cancer.

Results

Of 181 eligible radiologists contacted, 139 (77%) completed the survey with consent to link their responses to performance measures already recorded in the BCSC data. We found no statistically significant differences between responders and non-responders with regard to sensitivity, specificity, or recall rate. Among the 139 participating radiologists, 124 (89%) had interpreted 480 or more screening mammograms within BCSC facilities between 1996 and 2001 and thus were included here.

The 124 radiologists interpreted mammograms at 81 facilities in the three registries. A total of 308 634 women obtained 557 143 screening mammograms during the study period at these registries. However, some mammograms did not have breast density reported on the four-category scale and were eliminated (n = 87 631 or 15.7%) leaving 469 512 screening mammograms. Breast cancer was diagnosed within 1 year in 2402 women, a rate of 5.12 per 1000 screening mammograms (95% CI 4.92 to 5.32).

The distribution of patient characteristics and BI-RADS assessment for screening mammograms by a subsequent diagnosis of breast cancer in the follow-up period is shown in Table 1. Recall rates generally increased with breast density and decreased with age and previous mammography. The probability of cancer increased monotonically with the ordered BI-RADS assessments, demonstrating that the radiologists use the BI-RADS scale appropriately. Using the distribution of BI-RADS assessments, sensitivity and specificity can be computed at any choice of cutpoint. For example, the definition used in this article yielded an overall specificity of 90.1% and sensitivity of 81.6%. However, increasing the threshold for calling the mammogram positive yielded lower sensitivity and greater specificity by definition.

Characteristics of Radiologists and Their Associated Recall Rate

Radiologist demographic and clinical practice characteristics are shown in Table 2. The ages of radiologists ranged from 35 to 79 years. Most were male (77.4 %), worked full-time (74.0 %), were not affiliated with an academic medical center (83.9 %), had more than 10 years of experience interpreting mammograms (77.2 %), and spent less than 40% of their time working in breast imaging (87.7 %). The reported number of mammograms interpreted in the year before the survey ranged from 500 to more than 5000.

The recall rates for the radiologists are shown in Table 2 by response to the survey items. Each screening mammogram had equal weight in the analysis, so a radiologist who had interpreted more screening examinations would have a higher weight. The crude recall rates were statistically significantly higher among radiologists who were younger, had fewer years of experience interpreting mammograms, and had higher annual volumes of screening

mammograms. For example, there was a statistically significant difference ($P = .002$) comparing recall rate over the different levels of volume. The odds ratios indicated that radiologists who read 1001–2000 mammograms in the past year had a higher recall rate compared with those who read 1000 or fewer (referent) (OR = 1.51, CI = 1.12 to 1.82), than radiologists with volumes of 2001 or more (OR = 1.29, CI = 0.97 to 1.35). In subsequent analyses, we adjusted for patient characteristics to determine if these apparent differences were maintained.

Radiologist Performance

Individual recall rates for the 124 radiologists varied considerably, ranging from 1.8% to 26.2% (data not shown). Higher sensitivity was often associated with a higher false-positive rate for the 124 radiologists (Fig. 1). The estimate of sensitivity was based on small numbers of cancers per radiologist, so it was highly variable. Specificity varied from 74% to 98%.

For modeling sensitivity, specificity, and recall rate, we adjusted for the patient's age, breast density, previous mammography experience, and mammography registry as fixed effects before assessing variability of the radiologist. Analysis (data not shown) showed statistically significant variability among radiologists for sensitivity ($P = .002$), specificity ($P<.001$), and recall rate ($P<.001$) after adjusting for patient covariates.

To explain the statistically significant radiologist variability, we modeled the additional impact of each of the radiologist variables separately. The odds ratios, 95% confidence intervals (CI), and overall *P* values for analyses of sensitivity and specificity after adding each radiologist characteristic individually are shown in Table 3. Increasing age of the radiologist was associated with decreased sensitivity ($P = .001$) and increased specificity ($P = .005$). Radiologists who worked full-time had higher sensitivity ($P = .002$) than those who worked part-time but the same specificity. Increasing number of years interpreting mammograms was associated with decreased sensitivity ($P = .003$) and increased specificity ($P = .007$). There was a statistically significant ($P = .018$) but inconsistent, nonlinear relationship of percentage of time spent working in breast imaging with specificity. Statistically significant associations of both sensitivity and specificity with mammographic volume (number of mammograms interpreted within the past year) were also found. Sensitivity increased with volume ($P = .004$) and specificity decreased with volume ($P = .002$). Sex, academic affiliation of the radiologist, and percentage of examinations interpreted by each radiologist that were screening mammograms were not associated with either outcome. The radiologist variables were not adjusted for each other, only for patient variation. Simultaneous consideration of radiologist factors is considered below.

We also considered the effects of malpractice history, insurance, concern about malpractice, and attitudes toward mammography (Table 4). Only concern about malpractice was statistically significantly associated with sensitivity ($P = .03$), but there was no ordinal relationship between level of concern and sensitivity. There were no other associations with sensitivity and no malpractice variable was related to specificity.

Statistically significant radiologist factors were then tested together using mixed-effects models. Age of radiologist, percentage of time spent working in breast imaging, and concern about malpractice were no longer statistically significant after adjustment for other radiologist variables. The final model thus included the two radiologist factors of experience —i.e., number of years in mammography practice and number of mammograms interpreted per year (Table 5). Sensitivity decreased with number of years interpreting mammograms and increased with number of mammograms interpreted per year ($P = .001$ for both). Conversely, specificity increased with number of years interpreting mammograms and decreased with number of mammograms interpreted per year ($P = .003$ and .002,

respectively). Residual radiologist variability was no longer statistically significant for sensitivity ($P = .34$) but remained highly statistically significant for specificity ($P<.001$).

Results From the ROC Models

Figures 2 and 3 show the empirical ROC curves for the number of mammograms interpreted annually and the number of years of experience. Although there appears to be a difference in accuracy by number of mammograms interpreted, the comparison may be confounded by patient variation. Radiologists with lower volumes interpreted mammograms from women that were younger and were more likely to be having their first mammogram. Consequently, it was necessary to statistically adjust for confounders before assessing the effect of radiologist level covariates. The model for the ROC analysis included the statistically significant variables in the analyses of sensitivity and specificity. Table 6 shows the statistical significance of the covariate effects on estimation of threshold for calling the mammogram positive ($P_{threshold}$) and the estimated difference in ROC curves ($P_{accuracy}$). After adjustment for mammography registry, statistically significant effects on the threshold included the patient's age ($P<.001$), breast density ($P<.001$), previous mammography history ($P<.001$), and the radiologist's number of years of experience ($P = .012$). Radiologist volume had no effect ($P = .088$) on threshold. Statistically significant factors for accuracy included breast density ($P<.001$) and previous mammography history ($P = .012$) but not age ($P = .59$). Neither mammographic volume ($P = .94$) nor years of experience in mammography ($P = .34$) showed statistically significant associations with accuracy.

Modeled areas under the ROC curve (AUC) that measure overall accuracy are also shown in Table 6. (Adjusted AUC values were obtained by weighing the distribution of all other covariate combinations in the model except for the one of interest; this approach internally standardizes the AUC.) The modeled ROC curves for breast density adjusted for age, prior mammography, and mammography registry are shown in Fig. 4. Accuracy was greatest for women with the least dense breasts, and it decreased as breast density increased.

We also performed a GEE analysis of the BI-RADS scores using an ordered probit model (data not shown). This analysis also showed no statistically significant effect of number of mammograms interpreted ($P = 1.00$) or years of experience ($P = .12$) when using this method for analysis of ROC data. A simpler method is to dichotomize the BI-RADS scale into a positive or negative screening examination and to perform a logistic regression analysis on the screening result using disease status, covariates, and the interaction of each covariate with disease status. This simpler GEE model (adjusted for clustering of mammograms within radiologist) showed no effect of mammography volume ($P = .057$) but a statistically significant effect of years of experience ($P = .021$) on the joint model of sensitivity and specificity.

Discussion

We investigated how a radiologist's self-reported clinical experience was associated with actual mammography performance. Radiologist's age, gender, malpractice experience, and malpractice concerns did not seem to be associated with performance. Radiologist's years of experience had the strongest association with performance, such that radiologists with fewer years in practice had higher sensitivity but lower specificity. It is reasonable to hypothesize that those with fewer years in practice are most recently trained; they missed fewer cancers than did radiologists who were in practice longer. However, it appeared that the effect of experience was on the threshold for calling the mammogram positive, rather than in a true difference in ability to detect cancers. For radiologists reporting fewer than 5 years of experience, we inspected how their recall rates changed over calendar time and observed little variation (data not shown), suggesting that early experience is not impacting

subsequent decision making. We hypothesize that training prior to starting practice is the most important component of accuracy, but this hypothesis requires further examination of the individual radiologist's performance over time.

The U.S. Mammography Quality Standards Act requires an average of 480 mammogram interpretations per year, which may suggest that higher interpretive volumes will translate into greater experience and hopefully better accuracy. The literature has been mixed on this point, with one study showing improved performance with higher volumes (9) and another not supporting a role of volume on performance (10). In our study, radiologists with higher volumes did show higher recall rates and higher sensitivity but lower specificity. However, unless there is adequate feedback regarding cancer outcomes and discriminative skills, the effect of volume may be to simply encourage more positive calls. In general, a radiologist will see only 5.12 cancers per 1000 screening mammograms. With a sensitivity of 81.6%, the radiologist will have a false negative rate of 0.92 per 1000 mammograms. If the radiologist reads 500 screening mammograms each year, one cancer would be missed every 2 years. It is also possible that the radiologist would not receive any information about that missed cancer unless a malpractice claim was filed. Therefore, in the absence of feedback about performance it may be unlikely that a radiologist would change his or her threshold over time.

ROC analysis separates factors that affect threshold on the same ROC curve from factors that generate different curves. For example, the parameter estimates for number of years in practice indicates that the threshold used by radiologists with more years of experience is higher than that used by those with fewer years of experience. However, there was no statistically significant effect on accuracy after adjustment for the difference in threshold. These results imply that the same ROC curve applies to radiologists with different levels of years of experience, but because the thresholds are different, sensitivity and specificity will be affected. Similarly, mammographic volume in the past year appears to influence threshold but not overall accuracy. Therefore, for a fixed level of specificity, there is no evidence that greater volume will improve mammography performance.

We primarily used one statistical approach, a random effects model for the analysis of ROC curves though there are many other approaches. Both the random effects and GEE models showed no statistically significant effects of mammography volume on accuracy when the entire BI-RADS scale was used in the analysis. However, when the model was simplified to a dichotomous outcome, marginally statistically significant effects of these variables were observed. However, this simpler model assumes, incorrectly, that all positive mammograms have a common work-up evaluation and an equal likelihood of being confirmed as cancer. The data in Table 1 show that radiologists accurately use the ordinal BI-RADS scale so that using the full range of values will be more efficient than dichotomizing BI-RADS scores as positive or negative.

A potential weakness of this study is that the surveyed radiologists were not a random sample of all radiologists in the United States but only a sample participating in the national Breast Cancer Surveillance Consortium in three distinct locations. It is therefore possible that these radiologists may not be completely representative of all United States radiologists. However, the demographic characteristics of the radiologists in our study are similar to those of a randomly chosen national sample of radiologists (10). Thus, we believe that the radiologists surveyed represent a broad and reasonable cross-section of radiologists interpreting screening mammograms in the United States.

Another potential limitation of the study is that reported mammographic volume may have been estimated inaccurately by the radiologists when they responded to the survey. We

cannot directly compare reported volume to observed volume, for several reasons. First, a radiologist may practice at several facilities, some of which may not be included in the BCSC database. Second, the radiologist is reporting all mammograms in the survey, not just screening mammograms. Finally, the volume is reported in discrete categories used in the survey, rather than the actual numbers. However, we did assess actual screening volume per month from BCSC data and classified the radiologists into three groups: low, medium, or high volume. The same pattern of results was obtained in the ROC analysis when we used actual screening volume in place of reported total mammographic volume (data not shown). Thus, no difference in accuracy after comparing radiologists by actual screening mammography volume after adjustment for patient variables was observed. So, although the survey responses may not always be accurate, there is no evidence that the findings are misleading or that the results would be different using actual volume. Because all other radiologist variables were based on self-reported data we opted to be consistent.

This study was intended primarily to identify characteristics of radiologists that could affect performance. We found few measurable characteristics that had a statistically significantly association with performance, apart from those already identified in the literature. Another study using these data found that radiologists' degree of comfort with uncertainty also was not associated with performance (21). A second study (34) also reported no statistically significant association of malpractice fears or claims with recall rate. Thus, it may be important to look for other sources of radiologist variability than those included in this survey.

The study survey allowed us to gather data from the radiologist population pertinent to our hypothesis. The response rate of 77% is higher than that in many physician surveys, and no difference was noted between responders versus non-responders in patient outcomes. Furthermore, these radiologists are community-based and therefore are more representative of current practice than academically based radiologists, who are more involved in training and research.

The study design had several strengths. The study took advantage of a large population of women who underwent screening mammography at regular intervals and whose data was collected in a standardized manner. The mammograms were linked to cancer outcomes so that the full performance of screening mammography could be evaluated. Data on individual patients will allow adjustment for patient variation that has not been addressed in other studies. For example, it may be that radiologists with high volumes are associated with large urban screening programs and therefore their performance measures may differ due to factors other than volume.

How can radiologist performance be improved? It should be noted that mammography performance as measured by AUC is already high and therefore difficult to improve dramatically. Although radiologists differ in performance, accuracy does not appear to be simply attributable to years of experience or number of mammograms interpreted. Direct feedback of performance characteristics coupled with training (35–37) may be more helpful than experience without feedback. The most instructive exercise may be to have an open discussion of misjudged mammograms, but concern about malpractice claims may prevent this opportunity from occurring. Peer review of missed cancers is more commonly practiced in other countries, where legal action may be less likely than in the United States. Nonetheless, slight improvements in interpreting screening mammograms may have large consequences when weighted by the enormous number of screening mammograms performed annually.

Although Esserman et al. (9) argued that volume increases performance, we found no evidence in this data set that performance would be improved by increasing the required number of mammograms to be interpreted by each radiologist. Such an increase may even be counterproductive because it may decrease the number of radiologists available to interpret mammograms. In particular, raising the volume requirement could have a negative impact in sparsely populated rural areas where an individual radiologist may already have difficulty meeting the current requirement in the Mammography Quality Standards Act. A Government Accounting Office report suggests there is currently enough mammography capacity to meet demand in most areas but that delays in obtaining screening mammograms are already occurring in some urban areas (38). A recent analysis suggests that setting a marginally higher criterion for performance could eliminate 50% of the radiologists currently practicing mammography (39). Given the substantial burden on mammographers already, increasing volume requirements, without evidence of efficacy, would seem unwise.

Acknowledgments

This project was supported by public health service grant HS-10591 (J. Elmore) from the Agency for Healthcare Research and Quality, and the National Cancer Institute, National Institutes of Health and the Department of Health and Human Services and National Cancer Institute, Surveillance Grant #s CA63731 (S. Taplin), CA86082 (P. Carney), CA63736 (G. Cutter), and CA86076 (W. Barlow).

While NCI funded this work, all opinions are those of the co-authors and do not imply agreement or endorsement by the Federal Government or the National Cancer Institute.

We appreciate the dedication of the participating radiologists and project support staff. We also thank the reviewers for many helpful comments that improved the manuscript considerably.

References

1. Humphrey LL, Helfand M, Chan B, Woolf SH. Breast Cancer Screening: A Summary of the Evidence for the United States Preventive Services Task Force. Ann Intern Med. 2002; 137:347–60. [PubMed: 12204020]
2. United States Preventive Services Task Force. Summaries for patients. Screening for breast cancer: recommendations from the United States Preventive Services Task Force. Ann Intern Med. 2002; 137:I47. [PubMed: 12204048]
3. Zapka JG, Taplin SH, Solberg LI, Manos MM. A framework for improving the quality of cancer care: the case of breast and cervical cancer screening. Cancer Epidemiol Biomarkers Prev. 2003; 12:4–13. [PubMed: 12540497]
4. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med. 1994; 331:1493–9. [PubMed: 7969300]
5. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med. 1996; 156:209–13. [PubMed: 8546556]
6. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? J Women's Health. 1998; 7:443–9.
7. Nodine CF, Kundel HL, Mello-Thoms C, Weinstein SP, Orel SG, Sullivan DC, et al. How experience and training influence mammography expertise. Acad Radiol. 1999; 6:575–85. [PubMed: 10516859]
8. Elmore JG, Miglioretti DL, Reish LM, Barton MB, Kreuter W, Christiansen CL, et al. Screening mammograms by community radiologists: variability in false positive rates. J Natl Cancer Inst. 2002; 94:1373–80. [PubMed: 12237283]
9. Esserman L, Cowley H, Eberle C, Kirkpatrick A, Chang S, Berbaum K, et al. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst. 2002; 94:369–75. [PubMed: 11880475]

10. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst. 2003; 95:282–90. [PubMed: 12591984]
11. McKee MD, Cropp MD, Hyland A, Watroba N, McKinley B, Edge SB. Provider case volume and outcome in the evaluation and treatment of patients with mammogram-detected breast carcinoma. Cancer. 2002; 95:704–12. [PubMed: 12209712]
12. Rosenberg RD, Hunt WC, Williamson MR, Gilliland FD, Wiest PW, Kelsey CA, et al. Effects of age, breast density, ethnicity, and estrogen replacement therapy on screening mammographic sensitivity and cancer stage at diagnosis: review of 183 134 screening mammograms in Albuquerque, New Mexico. Radiology. 1998; 209:511–8. [PubMed: 9807581]
13. Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med. 2003; 138:168–75. [PubMed: 12558355]
14. Houssami N, Ciatto S, Irwig L, Simpson JM, Macaskill P. The comparative sensitivity of mammography and ultrasound in women with breast symptoms: an age-specific analysis. Breast. 2002; 11:125–30. [PubMed: 14965658]
15. Laya MB, Larson EB, Taplin SH, White E. Effect of estrogen replacement therapy on the specificity and sensitivity of screening mammography. J Natl Cancer Inst. 1996; 88:643–9. [PubMed: 8627640]
16. Fajardo LL, Hillman BJ, Frey C. Correlation between breast parenchyma patterns and mammographer's certainty of diagnosis. Invest Radiol. 1988; 23:505–8. [PubMed: 3170137]
17. Rosenberg RD, Yankaskas BC, Hunt WC, Ballard-Barbash R, Urban N, Ernster VL, et al. Effect of variations in operational definitions on performance estimates for screening mammography. Acad Radiol. 2000; 7:1058–68. [PubMed: 11131050]

17a. Frankel SD, Sickles EA, Curpen BN, Sollitto RA, Ominsky SH, Galvin HB. Initial versus subsequent screening mammography: comparison of findings and their prognostic significance. Am J Roentgenol. 1995; 164:1107–9. [PubMed: 7717214]

18. 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. AJR Am J Roentgenol. 1997; 169:1001–8. [PubMed: 9308451]
19. Carney P, Poplack S, Wells W, Littenberg B. Development and Design of a Population-Based Mammography Registry: The New Hampshire Mammography Network. Am J Roentgenol. 1996; 167:367–372. [PubMed: 8686606]
20. Houn F, Elliot ML, McCrohan JL. The mammography quality standards act of 1992. Radiology Clinical North America. 1995; 33:1059–1065.
21. Carney PA, Elmore JG, Abraham LA, Gerrity MS, Hendrick RE, Taplin SH, et al. Radiologist uncertainty and the interpretation of screening mammography. Med Decision Making. 2004; 24:225–64.
22. American College of Radiology. Illustrated Breast Imaging Reporting and Data System (BI-RADS), 3rd ed. Reston (VA): American College of Radiology; 1998.
23. 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; 222:529–35. [PubMed: 11818624]
24. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 42:121–30. [PubMed: 3719049]
25. Barlow WE, Lehman CD, Zheng Y, Ballard-Barbash R, Yankaskas BC, Cutter GR, et al. Performance of diagnostic mammography for women with signs or symptoms of breast cancer. J Natl Cancer Inst. 2002; 94:1151–9. [PubMed: 12165640]
26. SAS Institute Inc. SAS/STAT user's guide, version 8. Cary (NC): SAS Institute; 1999.
27. Dorfman DD, Alf E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals – rating-method data. J Math Psychol. 1969; 6:487–96.

28. Tosteson AN, Begg CB. A general regression methodology for ROC curve estimation. Med Decis Making. 1988; 8:204–15. [PubMed: 3294553]
29. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol. 1989; 24:234–45. [PubMed: 2753640]
30. Toledano AY, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. Stat Med. 1996; 15:1807–26. [PubMed: 8870162]
31. Peng FC, Hall WJ. Bayesian analysis of ROC curves using Markovchain Monte Carlo methods. Med Decis Making. 1996; 16:404–11. [PubMed: 8912302]
32. Daniels MJ, Gatsonis C. Hierarchical polytomous regression models with applications to health services research. Stat Med. 1997; 16:2311–25. [PubMed: 9351167]
33. Hellmich M, Abrams KR, Jones DR, Lambert PC. A Bayesian approach to a general regression model for ROC curves. Med Decis Making. 1998; 18:436–43. [PubMed: 10372587]
34. Elmore JG, Taplin SH, Barlow WE, Cutter GR, D'Orsi C, Hendrick RE, et al. Does litigation influence medical practice? The influence of community radiologists' medical malpractice perceptions and experience on screening mammography. Radiology. In press.
35. Pisano ED, Burns CB, Washburn D. Educational outreach to mammography facility staff to assist with compliance with the Mammography Quality Standards Act in rural North Carolina. Acad Radiol. 1998; 5:485–90. [PubMed: 9653465]
36. Beam CA, Conant EF, Sickles EA. Factors affecting radiologist inconsistency in screening mammography. Acad Radiol. 2002; 9:531–40. [PubMed: 12458879]
37. Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. Radiographics. 2003; 23:881–95. [PubMed: 12853663]
38. Government Accounting Office (GAO). Mammography: Capacity Generally Exists to Deliver Services. GAO Report 02–532. April. 2002 (www.fda.gov/cdrh/mammography/pubs/d02532.pdf)
39. Beam CA, Conant EF, Sickles EA, Weinstein SP. Evaluation of proscriptive health care policy implementation in screening mammography. Radiology. 2003; 229:534–40. [PubMed: 14595152]

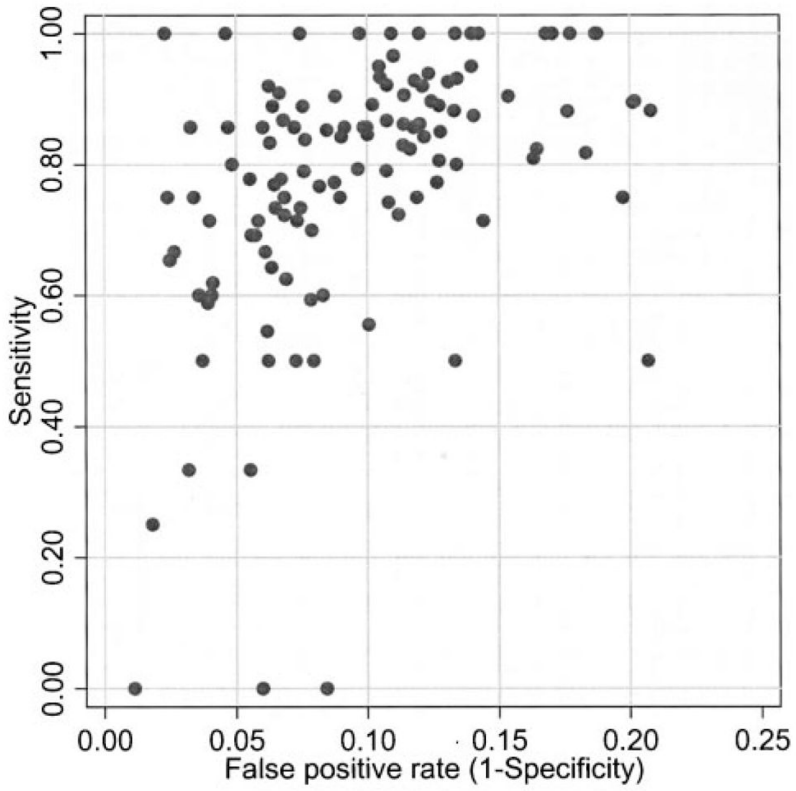

Fig. 1.
True positive rate (sensitivity) of the 124 radiologists versus the false positive rate (1 – specificity). Rates not adjusted for patient variables.

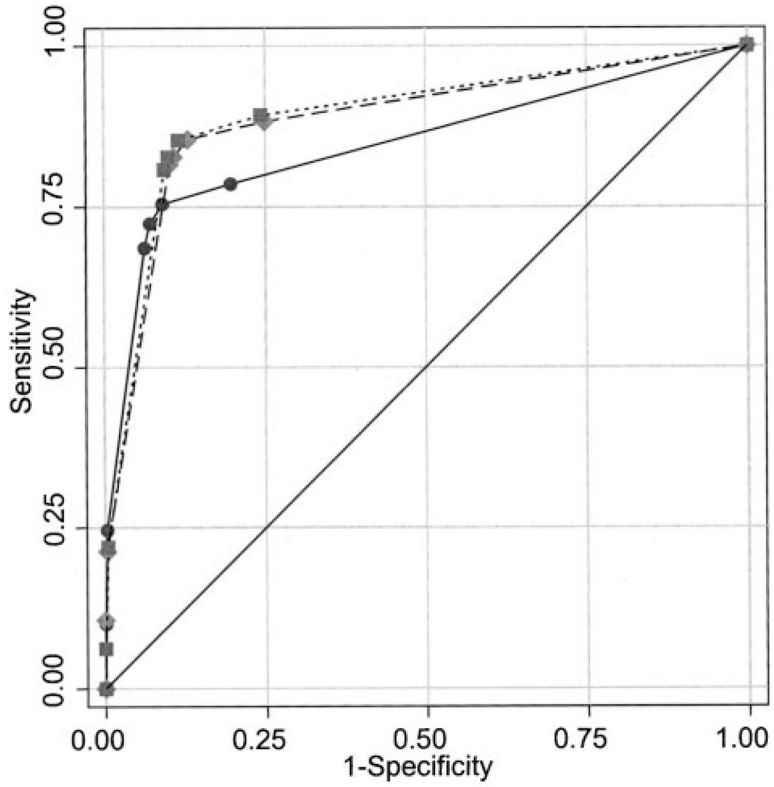

Fig. 2.

Empirical receiver operating characteristic curves for mammograms of radiologists who reported interpreting ≤1000 (**solid circles**), 1001–2000 (**solid diamonds**), or >2000 (**solid squares**) mammograms per year, not adjusted for patient variables.

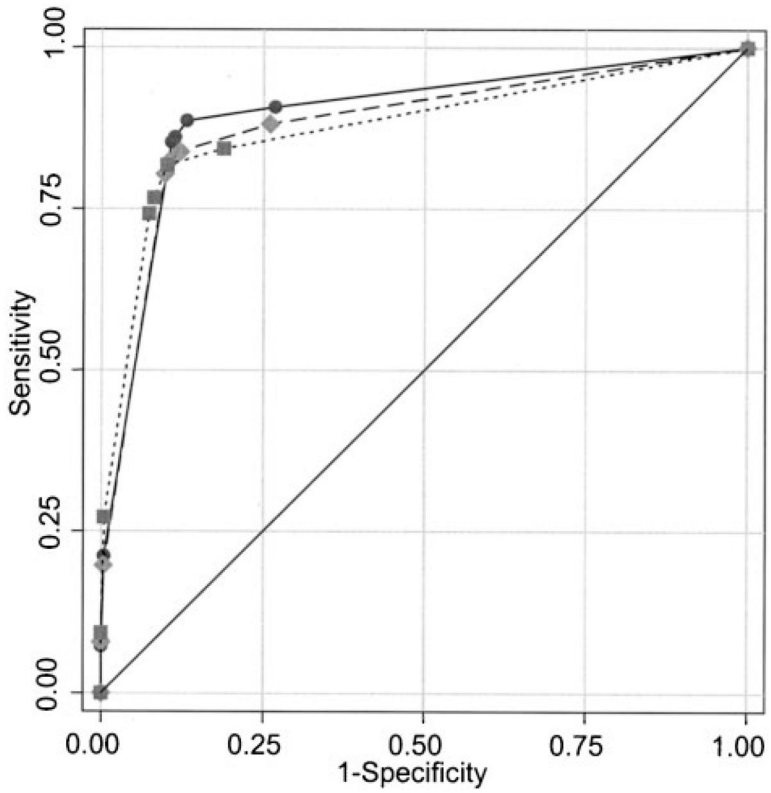

Fig. 3.

Empirical receiver operating characteristic curves for mammograms of radiologists who reported <10 years (**solid circles**), 10–19 years (**solid diamonds**), or ≥20 years (**solid squares**) of experience interpreting mammograms, not adjusted for patient variables.

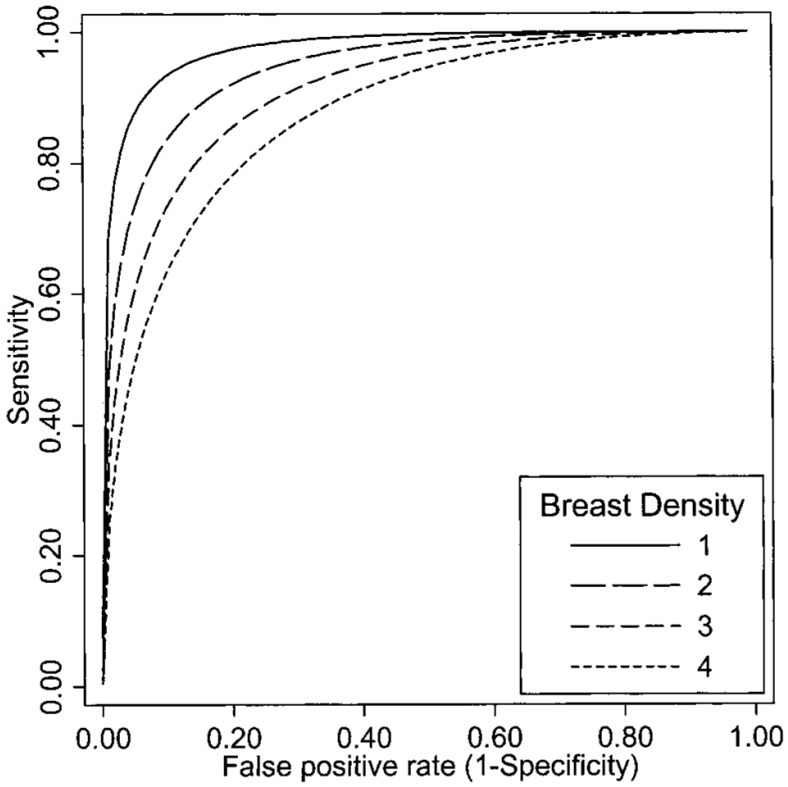

Fig. 4.

Modeled receiver operating characteristic curves by breast density adjusted for age, prior mammography, and mammography registry. Category 1 = almost entirely fatty, category 2 = scattered fibroglandular densities, category 3 = heterogeneously dense, and category 4 = extremely dense.

Table 1

Description of patient characteristics and BI-RADS assessment by breast cancer status and associated cancer and recall rates

Characteristic	Total no. of mammograms	Mammograms in women without breast cancer (%)	Mammograms in women with breast cancer (%)	Cancer rate per 1,000 screening mammograms	% recall rate*
Total	469 512	467 110	2,402	5.12	
Age at time of mammogram, y					
40–49	154 761	154 303 (33.0)	458 (19.1)	2.96	11.7
50–59	145 680	144 999 (31.0)	681 (28.4)	4.67	10.6
60–69	86 099	85 518 (18.3)	581 (24.2)	6.75	9.3
70–79	64 246	63 733 (13.6)	513 (21.4)	7.98	8.0
≥80	18 726	18 557 (4.0)	169 (7.0)	9.02	7.7
Breast density					
Almost entirely fat	51 345	51 197 (11.0)	148 (6.2)	2.88	4.6
Scattered fibroglandular tissue	203 821	202 862 (43.4)	959 (39.9)	4.71	9.1
Heterogeneously dense	175 155	174 097 (37.3)	1058 (44.0)	6.04	12.9
Extremely dense	39 191	38 954 (8.3)	237 (9.9)	6.05	12.0
Previous mammography					
No	20 806	20 686 (4.4)	120 (5.0)	5.77	17.2
Yes	448 706	446 424 (95.6)	2282 (95.0)	5.09	9.9
BI-RADS assessment					
1: Normal	356 030	355 734 (76.2)	296 (12.3)	0.83	
2: Benign finding	56 614	56 533 (12.1)	81 (3.4)	1.43	
3: Probably benign with a recommendation for Normal or Short-term follow-up	8692	8627 (1.8)	65 (2.7)	7.48	
3+: Probably benign with a recommendation for immediate work-up	3094	3049 (0.7)	45 (1.9)	14.54	
0: Need additional imaging evaluation	42 823	41 442 (8.9)	1381 (57.5)	32.25	
4: Suspicious abnormality, biopsy should be considered	2022	1687 (0.4)	335 (13.9)	165.68	
5: Highly suggestive of malignancy	237	38 (0.0)	199 (8.3)	839.66	

*Recall rate percentage = proportion of patients having a positive mammogram. Recall rate is the proportion of all mammograms called positive (BI-RADS value of 3+, 0, 4, or 5). For example, 11.7% refers to the percentage of the 154 761 called positive. It is blank for the BI-RADS codes because this is the source of the outcome variable (positive/negative) and would just be 0% or 100% by definition.

Table 2

Radiologist demographic and clinical practice characteristics and recall rate (N = 124 radiologists)*

Characteristic	No.	%	Recall rate	Recall OR	95% CI	*P*
Radiologist demographics						
Age, y						
35–44	37	29.8	12.1	1.00	(referent)	.001
45–54	48	38.7	10.6	0.88	0.86 to 0.9	
≥55	39	31.5	8.5	0.78	0.76 to 0.8	
Sex						
Male	96	77.4	9.8	1.00	(referent)	.47
Female	28	22.6	11.4	1.09	0.86 to 1.37	
Practice type						
Work full-time						
No	32	26.0	10.9	1.00	(referent)	.11
Yes	91	74.0	9.9	0.95	0.93 to 1.04	
Affiliation with an academic medical center						
Yes	20	16.1	9.8	1.00	(referent)	.42
No	104	83.9	10.3	1.13	0.83 to 1.54	
General experience in breast imaging						
Years of mammography interpretation						
<10	28	22.7	11.8	1.00	(referent)	.005
10–19	57	46.3	10.8	0.81	0.64 to 1.02	
≥20	38	30.9	8.6	0.66	0.51 to 0.85	
% of time spent working in breast imaging						
<20	55	45.1	9.2	1.0	(referent)	.015
20–39	52	42.6	11.3	1.30	1.06 to 1.61	
≥40	15	12.3	9.5	0.92	0.68 to 1.44	
No. of mammograms interpreted in the past year						
≤1000	31	25.2	7.6	1.0	(referent)	.002
1001–2000	46	37.4	11.1	1.51	1.12 to 1.82	
>2000	46	37.4	10.4	1.29	0.97 to 1.35	
% of mammograms interpreted in the past year that were screening mammograms						

J Natl Cancer Inst. Author manuscript; available in PMC 2011 July 25.

Characteristic	No.	%	Recall rate	Recall OR	95% CI	*P*
≤50	10	8.1	9.5	1.0	(referent)	.77
51–75	52	42.3	10.0	1.06	0.73 to 1.55	
76–100	61	49.6	10.7	0.99	0.68 to 1.44	

* Frequencies may not add up to 124 due to missing responses. CI = confidence interval. OR = odds ratio. Recall rate is the rate of calling a mammogram positive. Odds ratios, confidence intervals, and omnibus *P* values (two-sided) were calculated using logistic regression adjusting for patient age, breast density, and prior mammography.

Table 3

Mixed effects modeling of sensitivity and specificity by each radiologist characteristic adjusting for patient characteristics*

	Sensitivity			Specificity		
Characteristic	**OR**	**95% CI**	***P***	**OR**	**95% CI**	***P***
Radiologist demographics						
Age, y						
35–44	1.00	(referent)	.001	1.00	(referent)	.005
45–54	0.79	(0.58 to 1.10)		1.13	(0.90 to 1.42)	
≥55	0.52	(0.37 to 0.74)		1.48	(1.16 to 1.88)	
Sex						
Male	1.00		.45	1.00		.45
Female	0.89	(0.66 to 1.20)		0.92	(0.72 to 1.16)	
Practice type						
Work full-time						
Yes	1.00	(referent)	.002	1.00	(referent)	.82
No	0.60	(0.44 to 0.82)		1.03	(0.81 to 1.30)	
Affiliation with an academic medical center						
Yes	1.00	(referent)	.31	1.00	(referent)	.42
No	0.82	(0.56 to 1.20)		0.88	(0.64 to 1.20)	
General experience in breast imaging						
Years of mammography interpretation						
<10	1.00	(referent)	.003	1.00	(referent)	.007
10–19	0.69	(0.48 to 0.98)		1.23	(0.98 to 1.56)	
≥20	0.50	(0.34 to 0.74)		1.51	(1.17 to 1.95)	
% of time spent working in breast imaging						
<20	1.0	(referent)	.87	1.0	(referent)	.018
20–39	1.07	(0.78 to 1.46)		0.77	(0.62 to 0.95)	
≥40	0.98	(0.65 to 1.47)		1.10	(0.81 to 1.48)	
No. of mammograms interpreted in the past year						
≤1000	1.0	(referent)	.004	1.0	(referent)	.002
1001–2000	1.57	(1.09 to 2.27)		0.67	(0.52 to 0.86)	

	Sensitivity			Specificity		
Characteristic	**OR**	**95% CI**	***P***	**OR**	**95% CI**	***P***
>2000	1.80	(1.27 to 2.54)		0.75	(0.59 to 0.96)	
% of mammograms interpreted in the past year that were screening mammograms						
≤50	1.0	(referent)	.56	1.0	(referent)	.77
51–75	1.30	(0.79 to 2.16)		0.93	(0.63 to 1.37)	
76–100	1.30	(0.79 to 2.13)		1.00	(0.69 to 1.46)	

* Odds ratio (OR), 95% confidence intervals (CI), and omnibus Wald test *P* values (two-sided) were calculated by using logistic regression, adjusting for patient age, breast density, and prior mammography. Additionally, random effects for radiologists were estimated and tested with a likelihood ratio test.

Table 4

Mixed effects modeling of sensitivity and specificity by each response about malpractice claims and concerns adjusting for patient characteristics*

	Sensitivity			Specificity		
Variable	**OR**	**95% CI**	***P***	**OR**	**95% CI**	***P***
Medical malpractice insurance						
Self pay/other	1.00	(referent)	.93	1.00	(referent)	.81
Facility pays	0.98	(0.60 to 1.59)		0.96	(0.70 to 1.32)	
Ever had a malpractice claim						
No claims	1.00	(referent)	.28	1.00	(referent)	.25
Non-mammogram related	0.79	(0.59 to 1.06)		1.19	(0.97 to 1.47)	
Mammogram related	0.86	(0.60 to 1.22)		1.11	(0.84 to 1.46)	
Concerned about malpractice						
Disagree	1.00	(referent)	.03	1.00	(referent)	.14
Neutral	0.77	(0.43 to 1.35)		1.11	(0.73 to 1.69)	
Agree	1.21	(0.73 to 2.01)		0.87	(0.60 to 1.25)	
Malpractice influences recommendation for ultrasound						
Not changed	1.00	(referent)	.52	1.00	(referent)	.46
Moderately increased	1.18	(0.87 to 1.59)		0.92	(0.74 to 1.15)	
Greatly increased	1.19	(0.77 to 1.82)		0.81	0.58 to 1.13)	
Interpreting mammograms is tedious						
Disagree	1.00	(referent)	.56	1.00	(referent)	.70
Neutral	0.82	(0.57 to 1.18)		0.95	(0.72 to 1.25)	
Agree	0.94	(0.70 to 1.28)		0.91	(0.73 to 1.13)	
Worry when not sure of a mammogram						
Disagree	1.00	(referent)	.52	1.0	(referent)	.25
Agree	0.90	(0.67 to 1.23)		0.88	(0.70 to 1.10)	

*Odds ratio (OR), 95% confidence intervals (CI), and omnibus Wald test *P* values (two-sided) were calculated using logistic regression, adjusting for patient age, breast density, and prior mammography. Additionally, random effects for radiologist were estimated with a likelihood ratio test.

J Natl Cancer Inst. Author manuscript; available in PMC 2011 July 25.

J Natl Cancer Inst. Author manuscript; available in PMC 2011 July 25.

Table 5

Mixed effects modeling of sensitivity and specificity by statistically significant radiologist characteristics adjusting for patient characteristics*

	Sensitivity			Specificity		
Variable	**OR**	**(95% CI)**	***P***	**OR**	**(95% CI)**	***P***
Patient variables						
Age, y						
40–49	1.00	(referent)	.19	1.00	(referent)	< .001
50–59	1.38	(1.01 to 1.89)		1.03	(1.00 to 1.05)	
60–69	1.26	(0.90 to 1.75)		1.12	(1.08 to 1.15)	
70–79	1.05	(0.74 to 1.48)		1.29	(1.24 to 1.34)	
≥80	1.44	(0.86 to 2.44)		1.44	(1.35 to 1.53)	
Breast density						
1. Almost entirely fat	4.76	(2.55 to 8.87)	< .001	2.29	(2.17 to 2.43)	< .001
2. Scattered fibroglandular densities	3.39	(2.38 to 4.84)		1.22	(1.18 to 1.27)	
3. Heterogeneously dense	1.94	(1.40 to 2.68)		0.88	(0.85 to 0.91)	
4. Extremely dense	1.00	(referent)		1.00	(referent)	
Previous mammography						
No	1.00	(referent)	.028	1.00	(referent)	< .001
Yes	0.51	(0.28 to 0.93)		1.78	(1.71 to 1.86)	
Radiologist variables						
Number of years interpreting mammograms						
<10	1.00	(referent)	.001	1.00	(referent)	.003
10–19	0.72	(0.52 to 1.01)		1.25	(1.00 to 1.57)	
≥20	0.51	(0.36 to 0.74)		1.55	(1.21 to 1.99)	
No. of mammograms interpreted in the past year						
≤1000	1.00	(referent)	.001	1.00	(referent)	.002
1001–2000	1.68	(1.18 to 2.39)		0.65	(0.52 to 0.83)	
>2000	1.89	(1.36 to 2.63)		0.76	(0.60 to 0.96)	
Residual random variation among radiologists			.34			< .001

*Odds ratio (OR), 95% confidence intervals (CI), and omnibus Wald test *P* values (two-sided) were calculated by using logistic regression, adjusting for all variables in the joint model including a random effect for radiologist.

Table 6

Modeling of ROC curves: *P*-values for threshold and accuracy and area under the ROC curve*

Factors	$P_{threshold}$	$P_{accuracy}$	AUC
Patient variables			
Age, y	< .001	.59	
40–49			0.921
50–59			0.925
60–69			0.922
70–79			0.918
≥80			0.934
Breast density	< .001	< .001	
1. Almost entirely fat			0.975
2. Scattered fibroglandular densities			0.943
3. Heterogeneously dense			0.909
4. Extremely Dense			0.872
Previous mammogram	< .001	.012	
No			0.946
Yes			0.921
Radiologist variables			
No. of years interpreting mammograms	.012	.34	
<10			0.920
10–19			0.919
≥20			0.931
No. of mammograms interpreted in the past year	.088	.94	
≤1000			0.921
1001–2000			0.924
>2000			0.922

* Threshold is movement along a single ROC curve, and accuracy is the difference among ROC curves. AUC = area under the curve. Two-sided *P* values were computed from the ROC analysis using likelihood ratio tests on the covariate main effects ($P_{threshold}$) or interactions with disease status ($P_{accuracy}$). All variables were included in the joint model, including radiologist random effects for both threshold and accuracy.