Radiologist Characteristics Associated With Interpretive Performance of Diagnostic Mammography

Diana L. Miglioretti, Rebecca Smith-Bindman, Linn Abraham, R. James Brenner, Patricia A. Carney, Erin J. Aiello Bowles, Diana S. M. Buist, Joann G. Elmore

Background Extensive variability has been noted in the interpretive performance of screening mammography; however, less is known about variability in diagnostic mammography performance.

Methods We examined the performance of 123 radiologists who interpreted 35895 diagnostic mammography examinations that were obtained to evaluate a breast problem from January 1, 1996, through December 31, 2003, at 72 facilities that contribute data to the Breast Cancer Surveillance Consortium. We modeled the influence of radiologist characteristics on the sensitivity and false-positive rate of diagnostic mammography, adjusting for patient characteristics by use of a Bayesian hierarchical logistic regression model.

Results The median sensitivity was 79% (range = 27%–100%) and the median false-positive rate was 4.3% (range = 0%–16%). Radiologists in academic medical centers, compared with other radiologists, had higher sensitivity (88%, 95% confidence interval [CI] = 77% to 94%, versus 76%, 95% CI = 72% to 79%; odds ratio [OR] = 5.41, 95% Bayesian posterior credible interval [BPCI] = 1.55 to 21.51) with a smaller increase in their false-positive rates (7.8%, 95% CI = 4.8% to 12.7%, versus 4.2%, 95% CI = 3.8% to 4.7%; OR = 1.73, 95% BPCI = 1.05 to 2.67) and a borderline statistically significant improvement in accuracy (OR = 3.01, 95% BPCI = 0.97 to 12.15). Radiologists spending 20% or more of their time on breast imaging had statistically significantly higher sensitivity than those spending less time on breast imaging (80%, 95% CI = 76% to 83%, versus 70%, 95% CI = 64% to 75%; OR = 1.60, 95% BPCI = 1.05 to 2.44) with non–statistically significant increased false-positive rates (4.6%, 95% CI = 4.0% to 5.3%, versus 3.9%, 95% CI = 3.3% to 4.6%; OR = 1.17, 95% BPCI = 0.92 to 1.51). More recent training in mammography and more experience performing breast biopsy examinations were associated with a decreased threshold for recalling patients, resulting in similar statistically significant increases in both sensitivity and false-positive rates.

Conclusions We found considerable variation in the interpretive performance of diagnostic mammography across radiologists that was not explained by the characteristics of the patients whose mammograms were interpreted. This variability is concerning and likely affects many women with and without breast cancer.

J Natl Cancer Inst 2007;99:1854–63

Radiologists' interpretive performance of mammography has remained highly variable despite improvements in the technical quality of mammography since the implementation of the Mammography Quality Standards Act of 1992 (1). Most studies on variability in mammography interpretation have focused on screening examinations; however, understanding variability in the performance of diagnostic mammography is equally important because the prevalence of breast cancer is approximately 10-fold higher and the stage of disease more advanced in women receiving diagnostic mammography than in those receiving screening mammography (2). Sickles et al. (2) evaluated the performance of diagnostic mammography for 646 radiologists in the United States and found that the abnormal interpretation rate for mammograms performed for the evaluation of a palpable breast lump ranged

Affiliations of authors: Group Health Center for Health Studies, Group Health Cooperative, Seattle, WA (DLM, LA, EJAB, DSMB); Department of Biostatistics, University of Washington School of Public Health and Community Medicine, Seattle, WA (DLM); Departments of Radiology (RSB, RJB) and Epidemiology and Biostatistics (RSB), University of California, San Francisco, CA; Departments of Family Medicine and Public Health and Preventive Medicine, Oregon Health & Science University, Portland, OR (PAC); Department of Medicine, University of Washington School of Medicine, Harborview Medical Center, Seattle, WA (JGE).

Correspondence to: Diana L. Miglioretti, PhD, Group Health Center for Health Studies, Group Health Cooperative, 1730 Minor Ave, Ste 1600, Seattle, WA 98101 (e-mail: miglioretti.d@ghc.org).

See "Funding" and "Notes" following "References."

DOI: 10.1093/jnci/djm238

from 0% to 33% and that the positive-predictive value among women for whom a biopsy examination was recommended ranged from 6% to 92% across radiologists. That study indicated profound variation in the interpretive performance of diagnostic mammography across radiologists; however, some of this variation may be due to differences in patient or radiologist characteristics.

Radiologists who specialize in breast imaging and facilities with at least one radiologist who has high interpretive volume have been shown to have better interpretative performance of diagnostic mammography (3,4). Better performance has also been associated with such system-level variables as availability of information about clinical symptoms (5), comparison with previous examinations (6), and performance of additional imaging workup (7). Except for the study by Jensen et al. (4), which included all facilities in Denmark, these studies were limited to a few radiologists (i.e., 10 or fewer) from a single facility. In addition, these studies did not adjust for important patient characteristics other than age (8).

High sensitivity is critical for mammography performed to evaluate a breast problem because it can reduce delays in diagnosis among women with later-stage, clinically manifested disease. However, false-positive rates tend to rise along with sensitivity, and it is important to control the number of false-positive examinations because these lead to potentially unnecessary biopsy examinations that are associated with patient morbidity and high financial cost. Achieving high sensitivity for detecting breast cancer with acceptable false-positive rates of biopsy examination is an important performance goal in diagnostic breast imaging. The purpose of this study was to evaluate whether radiologist experience and practice characteristics are associated with the interpretive performance of diagnostic mammography.

Subjects and Methods

Study Population

The following three mammography registries that are part of the National Cancer Institute–funded Breast Cancer Surveillance Consortium (9) (http://breastscreening.cancer.gov) contributed data for this study: Group Health, a nonprofit integrated health care organization with headquarters in Seattle, WA; the New Hampshire Mammography Network, which captures approximately 90% of mammograms performed in New Hampshire; and the Colorado Mammography Program, which captures approximately 50% of mammograms performed in the Denver metropolitan area. These registries collect patient demographic and clinical information each time a woman receives a mammography examination at a participating facility. This information was then linked to regional cancer registries and pathology databases to determine cancer outcomes. Data from the registries were pooled at a single location for analysis.

Radiologists who interpreted mammograms at a facility contributing to any of the three registries were invited to participate in a mailed survey in early 2002 by use of survey methods described previously (10) (questionnaire is available online, Supplementary Fig. 1). Of the 139 radiologists who responded to the survey (77% response rate), we excluded three who did not interpret diagnostic mammograms performed to evaluate a breast problem during the study period; four who self-reported interpreting fewer than 500 mammograms annually because requirements of the Mammography Quality Standards Act require radiologists to interpret approximately this volume per year; and nine who were missing information on one or more of the key variables of interest (years of experience, affiliation with an academic medical center, percentage of time spent on breast imaging, number of mammograms interpreted, percentage of mammograms interpreted that were diagnostic, number of breast biopsy examinations performed, and breast density on all the mammograms that they interpreted). The final study sample included 123 radiologists from 72 facilities. The Institutional Review Boards associated with the study sites approved all study activities.

We limited this study to 35 895 diagnostic mammograms indicated by the radiologist as being performed for the evaluation of a breast problem (clinical sign or symptom of breast cancer) from January 1, 1996, through December 31, 2003. We did not include mammograms performed for additional diagnostic evaluation after a screening examination or for short-interval follow-up of a probably benign finding. We excluded mammograms performed on women with breast augmentation, reduction, or reconstruction and on women younger than 18 years. Mammograms obtained after 2003 were excluded to ensure adequate time for ascertainment of cancers diagnosed within 365 days of the mammogram (11).

Measurements

The radiologist survey included questions about demographic characteristics (age and sex), experience (years of mammography interpretation, fellowship training in breast imaging, percentage

CONTEXT AND CAVEATS

Prior knowledge

Although high variability has been reported in the interpretive performance of screening mammography, less is known about variability in diagnostic mammography.

Study design

Multifacility retrospective study of the performance of 123 radiologists who interpreted more than 35 000 diagnostic mammographic examinations. The influence of radiologist characteristics on sensitivity and the false-positive rate were modeled by use of Bayesian hierarchical logistic regression.

Contribution

Considerable variation in interpretive performance of diagnostic mammography was found across radiologists that was not explained by characteristics of the patients whose mammograms were interpreted.

Implications

The variability in performance of diagnostic mammography is concerning and likely affects many women with and without breast cancer. Ways to improve the interpretive performance of diagnostic mammography should be investigated.

Limitations

This study represented a small percentage of radiologists working in breast imaging and of mammography facilities in the United States, which may limit the generalizability of its results.

of time working in breast imaging in the prior year, and total number of mammography examinations interpreted in the prior year), and clinical practice characteristics in the prior year (affiliation with academic medical center, percentage of mammograms interpreted that were diagnostic, and the number of breast biopsy examinations performed). Survey responses were independently entered into the database by two individuals at each mammography registry. Radiologists' survey responses were linked to their interpretive performance obtained from their respective mammography registries by use of an encrypted study identifier. Information on patient characteristics collected at the time of the mammography examination included patient age, Breast Imaging Reporting and Data System (BI-RADS) mammographic breast density (12), time since last mammography examination, and self-reported presence of a breast lump (possibly found by patient self-examination or by a clinician during a breast examination). We also obtained the radiologists' BI-RADS assessment and recommendation (12) and information on whether invasive breast cancer or ductal carcinoma in situ was diagnosed within 1 year of the mammogram.

Mammography examinations given a final BI-RADS assessment of 4 (i.e., suspicious abnormality) or 5 (i.e., highly suggestive of cancer) at the end of the imaging workup were considered to be positive (12). A final BI-RADS assessment of 0 (i.e., needs additional imaging evaluation) with a recommendation for biopsy examination, fine-needle aspiration, or surgical consultation was also considered to be positive. We classified as negative those mammograms given a final BI-RADS assessment of 1 (i.e., negative), 2 (i.e., benign finding), 3 (i.e., probably benign finding), or 0 without a recommendation for biopsy examination, fine-needle aspiration, or surgical consultation. Women were considered to have breast cancer if invasive carcinoma or ductal carcinoma in situ was diagnosed within 1 year of the mammography examination. Sensitivity was defined as the percentage of positive examinations among women diagnosed with breast cancer. The false-positive rate was defined as the percentage of positive examinations among women without a breast cancer diagnosis.

Statistical Analysis

We calculated unadjusted sensitivity and false-positive rates separately for each radiologist and by the characteristics of the women and radiologists. We calculated the median, range, and interquartile range (IQR) of the unadjusted rates across radiologists, restricting to radiologists who interpreted at least 10 mammograms, so that we could obtain reasonably precise estimates. Performance measures were also calculated by radiologist characteristics with adjustment for patient age, BI-RADS mammographic breast density (categories: almost entirely fatty, scattered fibroglandular tissue, heterogeneously dense, or extremely dense), time since last mammogram (<1 year, 1 to <3 years, ≥3 years, or no previous mammography), self-reported presence of a breast lump, and mammography registry. We calculated 95% confidence intervals (CIs) by use of generalized estimating equations (13) with an exchangeable correlation structure to account for correlation among multiple mammograms interpreted by the same radiologist. Because only four radiologists had fellowship training, we did not report performance separately for this group.

The association between the performance measures and radiologist characteristics were examined in multivariable models that adjusted for patient age, mammographic breast density, time since last mammogram, self-reported presence of a breast lump, and mammography registry. Because of the collinearity between radiologist age and years of mammography interpretation, only the latter was included in the multivariable models.

To take into account the trade-off between sensitivity and false-positive rate, we jointly modeled the sensitivity (the probability of a positive mammogram among women with cancer) and the false-positive rate (the probability of a positive mammogram among women without cancer) as a function of radiologist characteristics with a binary hierarchical receiver operating characteristic (ROC) model (14) with a logit (logarithm of the odds) link (15), adjusting for patient age, mammographic breast density, time since last mammography examination, and mammography registry. We included separate radiologist-specific random effects for sensitivity and false-positive rate. These random effects represent latent radiologist-level effects that account for residual differences in each radiologist's sensitivity and false-positive rate after adjusting for all covariates in the model. We defined accuracy as the coefficient corresponding to the interaction between cancer status and the covariate under study, which is a measure of whether the effect of the covariate on sensitivity is different from the effect on the false-positive rate, on the logit scale (14,16). A group of radiologists was considered to be more accurate if they had a higher sensitivity without a corresponding increase in false-positive rate of the same magnitude (on the logit scale) or if they had a lower false-positive rate without a corresponding decrease in sensitivity of the same magnitude.

We were interested in whether the variability in performance among radiologists is associated with any characteristics of radiologists (e.g., if more experienced radiologists are less variable in their interpretations than less experienced ones). In addition, regression coefficients from hierarchical logistic regression models may be biased if the variation of the random effects depends on covariates (17,18). Therefore, we modeled the standard deviation of the radiologist-specific random effects distributions as a function of radiologist characteristics by use of a log link (15).

A fully Bayesian approach was taken, and models were fit by use of WinBUGS software (19). We used vague prior distributions in which we assumed that the ROC model coefficients were normally distributed with a mean of zero and a prior variance of 100 000 and that the coefficients in the model for the standard deviation were uniformly distributed between zero and 10. For each model, we ran three chains (i.e., separate simulations) with different starting values to ensure convergence to same values. We ran each chain for 20 000 samples and discarded the first 5000 samples for burn-in (to allow the sampler to reach convergence). Along with estimates of the odds ratios (ORs) that were based on the posterior modes from the samplers, we report Bayesian 95% highest posterior density credible intervals (BPCIs), which provide a measure of precision similar to traditional (or frequentist) confidence intervals (20).

We display the results of the hierarchical modeling for sensitivity and false-positive rates by plotting the observed (unadjusted) rates for each radiologist (model 1) and the rates after each of the following three different levels of adjustment: in model 2, adjustments were made for registry and correlation within radiologists; in model 3, adjustments were made for patient characteristics (age, breast density, time since last mammography examination, and self-reported presence of a lump) in addition to the adjustments made in model 2; and in model 4, adjustments were also made for radiologist characteristics in addition to the adjustments made in model 3. Thus, each model was built on the prior model. All rates were adjusted to the overall distribution of covariates in the study population. The rates for each radiologist were connected by a line to illustrate how the individual rates change after these different levels of adjustment. One radiologist who only interpreted one mammogram was excluded from these figures because his or her false-positive rate was not meaningful.

All tests of statistical significance are two-sided. An □level of .05 was used to determine statistical significance.

Results

Our study included 123 radiologists from 72 facilities who interpreted 35 895 diagnostic mammography examinations performed on 32 587 women for the evaluation of a breast problem from January 1, 1996, through December 31, 2003. Of these examinations, 1424 (40 per 1000 mammography examinations) were associated with a breast cancer diagnosis within 1 year. The number of radiologists per facility ranged from 1 to 23, with a median of seven; the number of mammograms per facility ranged from 1 to 2828, with a median of 210. Over the study period, radiologists interpreted a median of 208 diagnostic mammography examinations (range = 1–1249; IQR = 90–445), of which a median of eight (range = 0–77; IQR = 3–17) were associated with a breast cancer diagnosis within 1 year.

Characteristics for the 123 radiologists in this study are shown in Table 1. The median age of radiologists was 49 years (range = 34–70 years), most (96 radiologists or 78%) were male, and most (94 radiologists or 76%) had been interpreting mammography for at least 10 years. Only four (3%) of the radiologists had fellowship training in breast imaging, all of whom had less than 10 years of experience interpreting mammography. Most radiologists (107 or 87%) spent less than 40% of their time working in breast imaging, and 32 (26%) of the 123 radiologists interpreted 1000 mammograms or fewer in the year before the survey. The primary appointment of seven radiologists (6%) was at an academic medical center. For 114 (93%) radiologists, less than half of the mammograms that

Table 1. Characteristics of 123 radiologists included in the study

Radiologist characteristic	No. (%) of radiologists	No. (%) of diagnostic mammograms	
		With cancer	Without cancer
Demographics			
Radiologist's age, y			
34–44	39 (31.7)	381 (26.8)	8813 (25.6)
45–54	47 (38.2)	690 (48.5)	16 202 (47.0)
≥55	37 (30.1)	353 (24.8)	9456 (27.4)
Radiologist's sex			
Male	96 (78.0)	1014 (71.2)	25 602 (74.3)
Female	27 (22.0)	410 (28.8)	8869 (25.7)
Experience			
In mammography interpretation, y			
<10	29 (23.6)	190 (13.3)	4678 (13.6)
10–19	57 (46.3)	877 (61.6)	19 944 (57.9)
≥20	37 (30.1)	357 (25.1)	9849 (28.6)
Fellowship training in breast imaging			
No	119 (96.7)	1398 (98.2)	33 862 (98.2)
Yes	4 (3.3)	26 (1.8)	609 (1.8)
% of time working in breast imaging			
<20	55 (44.7)	417 (29.3)	11 531 (33.5)
20–39	52 (42.3)	818 (57.4)	18 816 (54.6)
≥40	16 (13.0)	189 (13.3)	4124 (12.0)
No. of mammograms interpreted in prior year			
500–1000	32 (26.0)	171 (12.0)	4320 (12.5)
1001–2000	46 (37.4)	529 (37.1)	12 008 (34.8)
>2000	45 (36.6)	724 (50.8)	18 143 (52.6)
Practice characteristics			
Primary affiliation with an academic medical center			
No	116 (94.3)	1373 (96.4)	33 377 (96.8)
Yes	7 (5.7)	51 (3.6)	1094 (3.2)
% of mammograms interpreted that were diagnostic			
0–24	61 (49.6)	679 (47.7)	17 145 (49.7)
25–49	53 (43.1)	667 (46.8)	15 052 (43.7)
50–100	9 (7.3)	78 (5.5)	2274 (6.6)
Performed breast biopsy examination in prior year			
No	33 (26.8)	258 (18.1)	7101 (20.6)
Yes	90 (73.2)	1166 (81.9)	27 370 (79.4)

Table 2. Characteristics of patients included in this study and associated observed (unadjusted) sensitivity and false-positive rates of diagnostic mammography examinations*

	No. (%) of diagnostic mammograms			
Patient characteristic	**With cancer**	**Without cancer**	**Observed sensitivity, % (95% CI)**	**Observed false-positive rate, % (95% CI)**
Patient's age, y				
<40	122 (8.6)	7476 (21.7)	72.3 (63.2 to 79.8)	4.9 (4.2 to 5.8)
40–49	348 (24.4)	12063 (35.0)	75.6 (70.1 to 80.3)	4.2 (3.7 to 4.9)
50–59	345 (24.2)	7393 (21.4)	76.3 (71.1 to 80.8)	3.7 (3.2 to 4.3)
60–69	242 (17.0)	3768 (10.9)	76.9 (71.3 to 81.7)	3.9 (3.2 to 4.8)
≥70	367 (25.8)	3771 (10.9)	77.6 (72.6 to 81.9)	5.4 (4.6 to 6.3)
Time since woman's last mammography examination				
<1 y	405 (28.4)	9541 (27.7)	67.6 (62.6 to 72.2)	4.9 (4.3 to 5.6)
1 to <3 y	556 (39.0)	15065 (43.7)	74.7 (69.5 to 79.3)	3.4 (2.9 to 3.9)
≥3 y	227 (15.9)	4108 (11.9)	82.8 (76.5 to 87.7)	5.2 (4.5 to 6.0)
No previous mammography	236 (16.6)	5757 (16.7)	86.0 (80.8 to 90.0)	5.5 (4.8 to 6.4)
Mammographic breast density				
Almost entirely fat	72 (5.1)	2247 (6.5)	86.4 (77.5 to 92.2)	3.2 (2.5 to 4.0)
Scattered fibroglandular densities	462 (32.4)	11353 (32.9)	77.1 (72.4 to 81.3)	3.5 (3.1 to 4.0)
Heterogeneously dense	690 (48.5)	15215 (44.1)	75.6 (71.8 to 79.1)	5.0 (4.5 to 5.6)
Extremely dense	200 (14.0)	5656 (16.4)	71.3 (63.3 to 78.1)	4.8 (4.1 to 5.8)
Reported presence of a breast lump				
Yes	937 (65.8)	16546 (48.0)	80.3 (77.0 to 83.2)	5.4 (4.8 to 6.0)
No	487 (34.2)	17925 (52.0)	68.4 (63.4 to 73.1)	3.4 (3.0 to 3.8)

* In this study, 35895 diagnostic mammography examinations were performed on 32587 women. CI = confidence interval.

they interpreted were diagnostic. Most radiologists (90 or 73%) performed breast biopsy examinations.

Patient Characteristics and Performance of Diagnostic Mammography

Characteristics of women included in this study and the associated sensitivity and false-positive rates among these women are shown in Table 2. Most women were younger than 60 years and had a prior mammography examination within the previous 2 years. A breast lump was reported by 937 (66%) women with breast cancer compared with 16546 (48%) women without breast cancer.

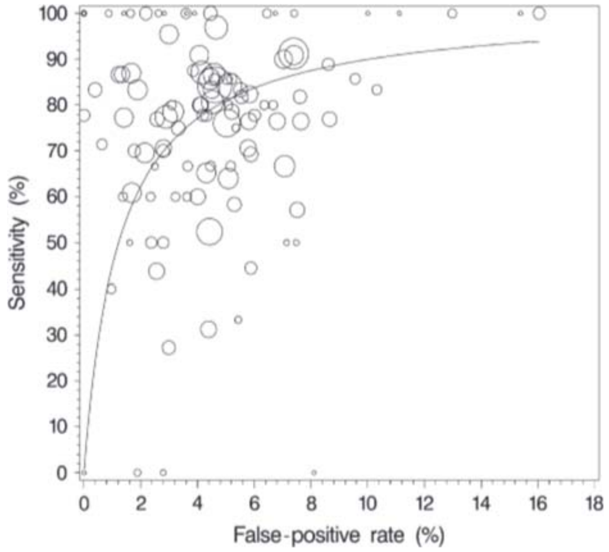


Fig. 1. Observed (unadjusted) radiologist-specific sensitivity versus false-positive rate and the corresponding receiver operating characteristic curve within the observed range of false-positive rates. The area of a **circle** is proportional to the number of mammograms from patients with a diagnosis of breast cancer that were interpreted by that radiologist (range = 1–77 mammograms).

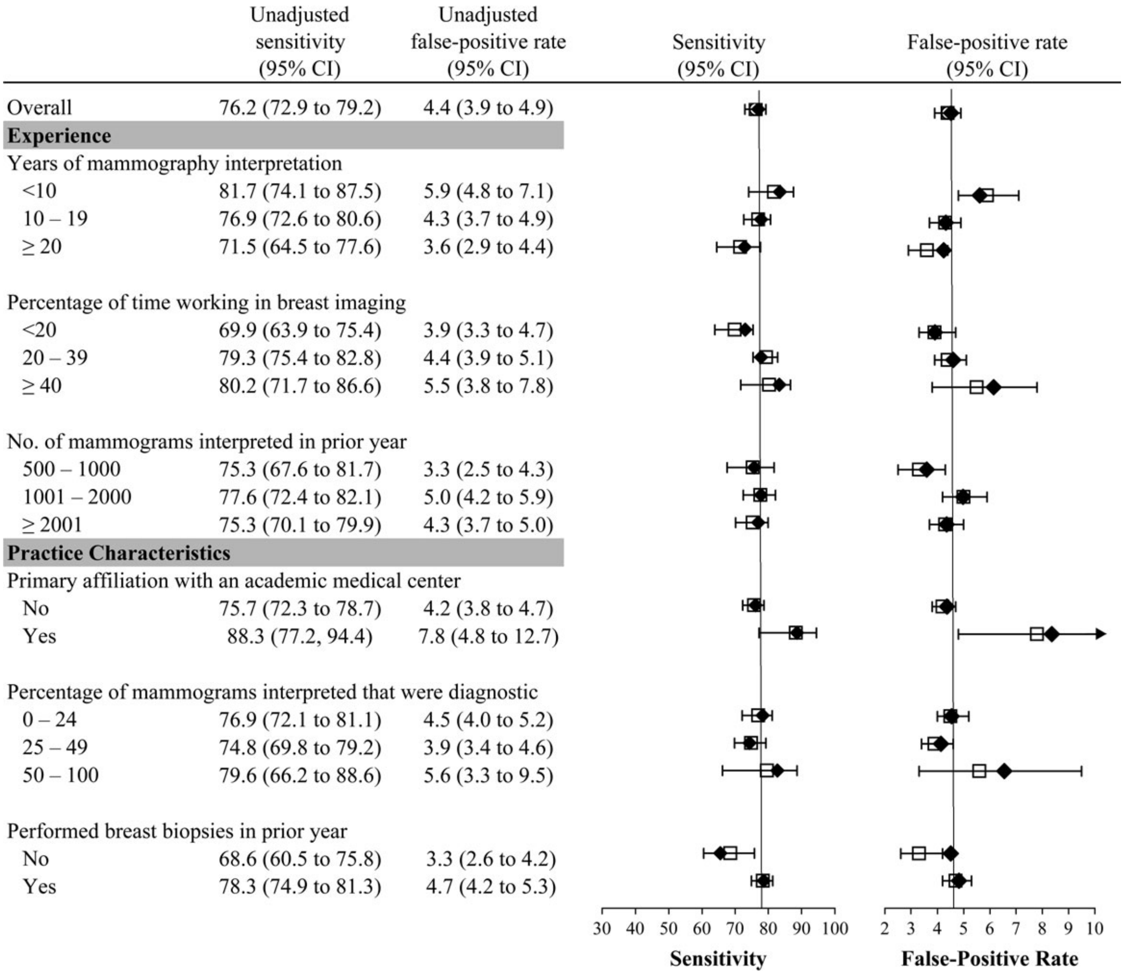
The sensitivity of diagnostic mammography increased with increasing age and decreasing mammographic breast density and was higher among women who had not received a mammography examination within 3 years and among women with a self-reported breast lump. The false-positive rate was higher among women with denser breasts and among women with a self-reported breast lump. Neither patient age nor time since last mammography examination showed any consistent trends with the false-positive rate.

Variability in Mammography Accuracy Among Radiologists

Among the 54 radiologists who interpreted at least 10 mammograms that were associated with a cancer diagnosis, the median sensitivity was 79% (range = 27%–100%; IQR = 70%–86%). Among the 118 radiologists who interpreted at least 10 mammograms that were not associated with a cancer diagnosis, the median false-positive rate was 4.3% (range = 0%–16%; IQR = 2.4%–5.8%). Some of the observed variability may likely be attributed to radiologists who used different thresholds for recommending a biopsy examination because sensitivity generally increases with increasing false-positive rate. However, sensitivity varied widely even among radiologists with similar false-positive rates (Fig. 1). The normalized partial area under the summary ROC curve was 0.80 across the observed range of false-positive rates when we assumed a constant accuracy among radiologists and varied the threshold for recall (21,22).

Radiologist Characteristics and Diagnostic Mammography Performance

The sensitivity and false-positive rates by radiologist characteristics, both unadjusted and adjusted for patient age, mammographic



	Unadjusted sensitivity (95% CI)	Unadjusted false-positive rate (95% CI)
Overall	76.2 (72.9 to 79.2)	4.4 (3.9 to 4.9)
Experience		
Years of mammography interpretation		
<10	81.7 (74.1 to 87.5)	5.9 (4.8 to 7.1)
10 – 19	76.9 (72.6 to 80.6)	4.3 (3.7 to 4.9)
≥ 20	71.5 (64.5 to 77.6)	3.6 (2.9 to 4.4)
Percentage of time working in breast imaging		
<20	69.9 (63.9 to 75.4)	3.9 (3.3 to 4.7)
20 – 39	79.3 (75.4 to 82.8)	4.4 (3.9 to 5.1)
≥ 40	80.2 (71.7 to 86.6)	5.5 (3.8 to 7.8)
No. of mammograms interpreted in prior year		
500 – 1000	75.3 (67.6 to 81.7)	3.3 (2.5 to 4.3)
1001 – 2000	77.6 (72.4 to 82.1)	5.0 (4.2 to 5.9)
≥ 2001	75.3 (70.1 to 79.9)	4.3 (3.7 to 5.0)
Practice Characteristics		
Primary affiliation with an academic medical center		
No	75.7 (72.3 to 78.7)	4.2 (3.8 to 4.7)
Yes	88.3 (77.2, 94.4)	7.8 (4.8 to 12.7)
Percentage of mammograms interpreted that were diagnostic		
0 – 24	76.9 (72.1 to 81.1)	4.5 (4.0 to 5.2)
25 – 49	74.8 (69.8 to 79.2)	3.9 (3.4 to 4.6)
50 – 100	79.6 (66.2 to 88.6)	5.6 (3.3 to 9.5)
Performed breast biopsies in prior year		
No	68.6 (60.5 to 75.8)	3.3 (2.6 to 4.2)
Yes	78.3 (74.9 to 81.3)	4.7 (4.2 to 5.3)

Fig. 2. Unadjusted and adjusted sensitivity and false-positive rates for diagnostic mammography by radiologist characteristics. Rates were adjusted for patient age, time since last mammogram, self-report of lump, breast density, and mammography registry. **Open squares** = unadjusted values; **solid diamonds** = adjusted values; CI = confidence interval.

breast density, time since last mammogram, reported presence of a breast lump, and mammography registry, are shown in Fig. 2. In general, most of the radiologist characteristics that were examined were associated with a change in threshold for interpreting an examination as abnormal. This change in threshold is shown in Fig. 2 as a shift in the same direction for both sensitivity and false-positive rate. For example, increasing years of experience interpreting mammography examinations were associated with a higher threshold for calling an examination abnormal, resulting in lower sensitivity (for <10 years of mammography interpretation, 82%, 95% CI = 74% to 88%; for 10–19 years, 77%, 95% CI = 73% to 81%; and for ≥20 years, 72%, 95% CI = 65% to 78%) as well as a lower false-positive rate (for <10 years of mammography interpretation, 5.9%, 95% CI = 4.8% to 7.1%; for 10–19 years, 4.3%, 95% CI = 3.7% to 4.9%; and for ≥20 years, 3.6%, 95% CI = 2.9% to 4.4%). In contrast, radiologists who had a primary affiliation with an academic medical center had a lower threshold for calling an examination abnormal, resulting in a higher sensitivity (88%, 95% CI = 77% to 94%, versus 76%, 95% CI = 72% to 79%) and a higher false-positive rate (7.8%, 95% CI = 4.8% to 12.7%, versus 4.2%, 95% CI = 3.8% to 4.7%). Radiologists spending 20% or more of their time on breast imaging had higher sensitivity (80%, 95% CI = 76% to 83%, versus 70%, 95% CI = 64% to 75%) with increased false-positive rates (4.6%, 95% CI = 4.0% to 5.3%, versus 3.9%, 95% CI = 3.3% to 4.6%). Similarly, radiologists who performed breast biopsy examinations, compared with those who did not, had higher sensitivity (78%, 95% CI = 75% to 81%, versus 69%, 95% CI = 60% to 76%) and higher false-positive rates (4.7%, 95% CI = 4.2% to 5.3%, versus 3.3%, 95% CI = 2.6% to 4.2%). Neither the total number of mammography examinations interpreted in the prior year nor the percentage of mammograms that were diagnostic was associated with sensitivity or false-positive rate.

Associations between radiologist characteristics and measures of mammography performance, after adjusting for the other radiologist characteristics, patient characteristics, and mammography registry, are shown in Fig. 3. Radiologists who had interpreted mammography for less than 10 years had a lower threshold for recalling women than those with more experience, which resulted in statistically significantly higher false-positive rates (OR = 1.27, 95% BPCI = 1.00 to 1.56) with a similar but not statistically significant increase in sensitivity (OR = 1.26, 95% BPCI = 0.77 to 2.16) and no difference in accuracy (OR = 1.00, 95% BPCI = 0.62 to 1.72).

	Sensitivity OR (95% BPCI)	False-positive rate OR (95% BPCI)	Accuracy OR (95% BPCI)
Experience			
Years of mammography interpretation			
<10	1.26 (0.77 to 2.16)	1.27 (1.00 to 1.56)	1.00 (0.62 to 1.72)
≥10	1.00 (referent)	1.00 (referent)	1.00 (referent)
Percentage of time working in breast imaging			
<20	1.00 (referent)	1.00 (referent)	1.00 (referent)
≥20	1.60 (1.05 to 2.44)	1.17 (0.92 to 1.51)	1.37 (0.89 to 2.07)
No. of mammograms interpreted in prior year			
500 – 1000	1.25 (0.73 to 1.96)	1.35 (0.98 to 1.89)	0.89 (0.52 to 1.53)
1001 – 2000	0.92 (0.56 to 1.53)	1.15 (0.79 to 1.58)	0.82 (0.48 to 1.40)
≥ 2001	1.00 (referent)	1.00 (referent)	1.00 (referent)
Practice Characteristics			
Primary affiliation with an academic medical center			
No	1.00 (referent)	1.00 (referent)	1.00 (referent)
Yes	5.41 (1.55 to 21.51)	1.73 (1.05 to 2.67)	3.01 (0.97 to 12.15)
Percentage of mammograms interpreted that were diagnostic			
0 – 24	1.00 (referent)	1.00 (referent)	1.00 (referent)
25 – 49	1.30 (0.52 to 2.92)	1.13 (0.74 to 1.76)	1.03 (0.45 to 2.69)
50 – 100	0.83 (0.37 to 1.92)	0.97 (0.65 to 1.45)	0.85 (0.37 to 2.08)
Performed breast biopsies in prior year			
No	1.00 (referent)	1.00 (referent)	1.00 (referent)
Yes	1.62 (1.05 to 2.78)	1.43 (1.11 to 1.87)	1.16 (0.74 to 1.88)

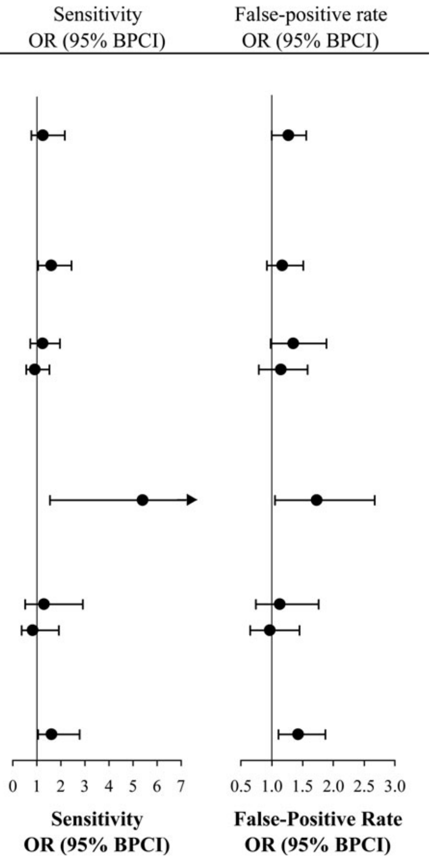


Fig. 3. Association between radiologist characteristics and a true-positive (sensitivity) and false-positive mammogram. In addition to the radiologist characteristics indicated, data were also adjusted for patient age, time since last mammogram, report of breast lump, mammographic breast density, and mammography registry. OR = odds ratio. BPCI = Bayesian posterior credible interval.

Radiologists who spent 20% or more of their time in breast imaging had a statistically significantly higher sensitivity than those spending less than 20% of their time in breast imaging (OR = 1.60, 95% BPCI = 1.05 to 2.44), with a smaller and not statistically significant increase in false-positive rate (OR = 1.17, 95% BPCI = 0.92 to 1.51) and a non–statistically significant increased accuracy (OR = 1.37, 95% BPCI = 0.89 to 2.07). Radiologists with a primary appointment at an academic medical center were statistically significantly more likely to detect breast cancer when it was present (i.e., had higher sensitivity, OR = 5.41, 95% BPCI = 1.55 to 21.51) than other radiologists, with a smaller but statistically significantly increased false-positive rate (OR = 1.73, 95% BPCI = 1.05 to 2.67) and a borderline statistically significant improvement in accuracy (OR = 3.01, 95% BPCI = 0.97 to 12.15). Radiologists who performed breast biopsy examinations had a lower threshold for recalling patients than those who did not perform breast biopsy examinations, which resulted in a statistically significantly higher sensitivity (OR = 1.62, 95% BPCI = 1.05 to 2.78), a statistically significantly higher false-positive rate (OR = 1.43, 95% BPCI = 1.11 to 1.87) and no difference in accuracy (OR = 1.16, 95% BPCI = 0.74 to 1.88). Neither annual interpretive volume nor the percentage of mammograms that were diagnostic was statistically significantly associated with sensitivity or false-positive rate.

Radiologists who worked at least 20% of their time in breast imaging showed less variability than those who spent less time in breast imaging in their false-positive rates (ratio of standard deviations of the radiologist-specific effects = 0.50, 95% BPCI = 0.26 to 0.93). None of the other radiologist characteristics were statistically significantly associated with the variability of the radiologist-specific effects for either sensitivity or false-positive rate.

Fig. 4 shows the observed sensitivity and false-positive rates of the study radiologists and the sensitivity and false-positive rates after three levels of adjustment. Substantial variation among radiologists remained even after full adjustment, with adjusted sensitivity ranging from 61.2% to 80.5% and adjusted false-positive rates ranging from 2.6% to 8.3%.

Discussion

We found considerable variation in the interpretive performance of diagnostic mammography that was not explained by the characteristics of the patients whose mammograms were interpreted. Because the rate of breast cancer is 10-fold higher among diagnostic mammograms than among screening mammograms (2,8) and the majority of women with breast cancer have a physical sign or symptom at the time of diagnosis (23–30), this variability in interpretive performance is concerning and likely affects many women both with and without breast cancer.

When examining the interpretive performance of mammography, it is important to distinguish differences in accuracy from differences in interpretive performance that result from the threshold that radiologists tend to use to consider an examination to be abnormal. For example, radiologists could increase their detection of cancer (sensitivity) by lowering their threshold for considering an examination to be abnormal, which would result in more false-positive examinations in addition to more true-positive examinations. Alternatively, radiologists could raise their thresholds for considering an examination to be abnormal, which would result in fewer false-positive and true-positive examinations. It is not inherently clear which of these strategies is better because this decision would depend on the extent to which detection of cancer is valued over limiting false-positive examinations. In contrast, a more accurate radiologist who is more skilled at mammography interpretation would be more successful in distinguishing cancer from the absence of cancer. Such a radiologist would have either higher sensitivity without a corresponding higher false-positive rate or a lower false-positive rate without a corresponding lower sensitivity. Therefore, both differences in threshold and differences in accuracy may be important. For diagnostic mammography, we suggest that it is particularly important to maximize sensitivity even at the expense of higher false-positive rates because the pretest probability of cancer is higher in diagnostic mammography than in screening mammography (3).

To distinguish changes in threshold from changes in accuracy, we used a binary ROC approach that modeled accuracy as the interaction between cancer status and covariates, on the logit scale. This model is equivalent to the hierarchical approach proposed by Rutter and Gatsonis (21) with the scale parameter set to 1.0. This constraint results in a symmetric radiologist-specific ROC curve (when both radiologist-specific effects equal zero) and a constant accuracy effect across the range of false-positive rates. Because cancer is uncommon, the statistical power for detecting changes in false-positive rate will always be higher than the power for detecting changes in sensitivity and accuracy (which involves an interaction with cancer status). Thus, we believe that it is important to evaluate the point estimates and absolute differences in sensitivity and false-positive rates in addition to the statistical significance because even large accuracy effects that are clinically relevant may fail to reach statistical significance.

In our study, the strongest predictor of improved accuracy of diagnostic mammography interpretation was having a primary affiliation with an academic medical center. Radiologists with a primary academic affiliation had much higher sensitivity, with a smaller increase in the false-positive rate and, as a result, a borderline statistically significant improvement in our overall accuracy. Some caution must be taken when interpreting these results, however, given that only seven radiologists in our study had a primary academic affiliation. In addition, it is possible that academic radiologists may see a different patient population with a different pretest probability of having breast cancer. For example, they may be more likely to receive referrals and second opinions. Although our study indicates that academic radiologists may be better at detecting cancer, they represent a small proportion of radiologists in the United States who interpret mammograms. Academic radiologists interpret only 6.5% of mammograms across the nation (31). We also found some evidence that a concentration in breast imaging may lead to improved performance, with those working at least 20% of their time in breast imaging and those who perform breast biopsy examinations having higher sensitivity than other radiologists. These results are consistent with two smaller studies that found breast-imaging specialists had higher cancer detection rates for diagnostic mammography (3,32). We hypothesize that working in a teaching environment and actively participating in the follow-up of their studies might enable radiologists to better learn whether specific lesions reflect cancer.

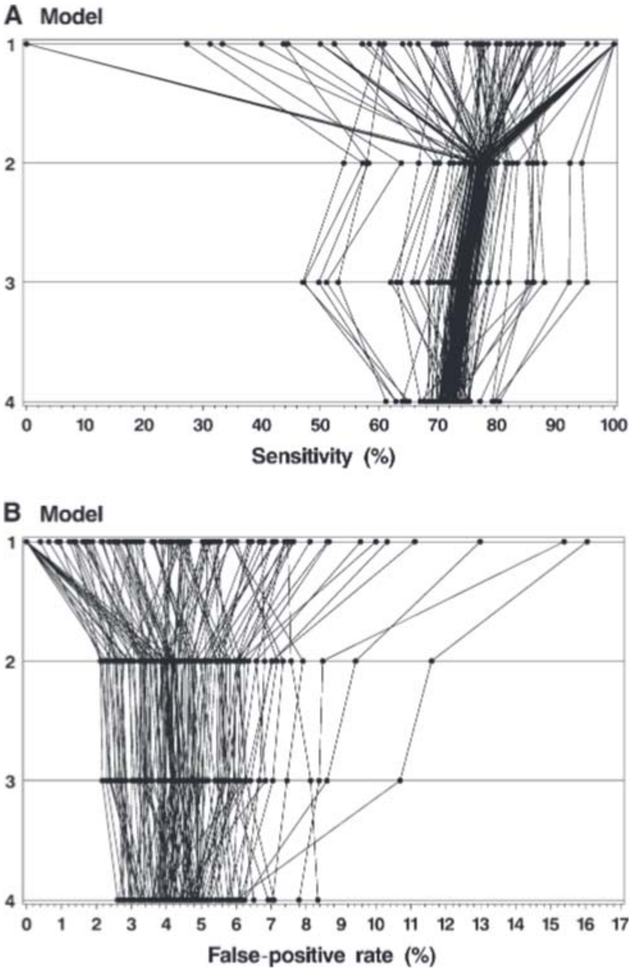


Fig. 4. Observed and adjusted sensitivity (**A**) and false-positive rate (**B**) for each radiologist. Model 1 = observed (unadjusted) rates; model 2 = adjusted for registry and correlation within radiologists; model 3 = additionally adjusted for patient characteristics (age, breast density, time since last mammography examination, and self-reported presence of a breast lump); model 4 = additionally adjusted for radiologist characteristics. Data for each radiologist have been connected with a **line**.

Consistent with several previous studies of screening mammography (10,16,33,34), we found that radiologists who had been interpreting mammography for many years tended to have a higher threshold for considering a diagnostic examination to be abnormal (i.e., lower false-positive rates but also lower sensitivity) than less experienced radiologists. Perhaps radiologists with more experience tend to recognize and avoid recall of the types of lesions that they have found to be benign in the past. Alternatively, more recent training may emphasize efforts to increase sensitivity at the cost of increased false-positive rates.

We found it interesting that two measures of experience– years of practice and percentage of time spent on breast imaging– had different impacts on the threshold for recommending a biopsy. More years of interpreting mammography examinations was associated with a higher threshold, whereas more time spent in breast imaging was associated with a lower threshold. Defining radiologist experience is complex, and it is difficult to isolate a single characteristic as most important.

We found that interpretive volume was not associated with either the sensitivity or false-positive rate of diagnostic mammography; however, we used a self-report measure of interpretive volume with only three categories. Jensen et al. (4) found higher sensitivity for diagnostic mammography among facilities in Denmark that had at least one high-volume radiologist than other facilities. They did not find a consistent trend between the overall facility volume and performance, but they did not look at the association between radiologist volume and their individual performance. The influence of interpretive volume on screening mammography performance has recently received much attention (1,10,16,35–38), but conflicting study findings have defied consensus on this issue. Studies differed in the methods used for measuring interpretive volume as well as the performance indices and statistical methods used. More research is needed to understand the implications of the different methodologies used. In addition, we need to develop more accurate measures of both short- and long-term interpretive volume.

Our study has several possible limitations. First, although we studied many radiologists (n = 123) from many facilities (n = 72) in three geographically diverse regions, our study population represents a small percentage of radiologists working in breast imaging and of mammography facilities in the United States. As a result, some of our subgroups were small, possibly limiting the generalizability of the results. For example, only seven radiologists had their primary affiliation with an academic medical center and only four radiologists had fellowship training in breast imaging. Second, we relied on self-reported measures of clinical experience and practice from radiologists. Although, for most of these measures, self-reports should be accurate, some measures, such as interpretive volume, should be validated in future studies. Third, mammography examinations may not be designated as being diagnostic in a standardized way across radiologists and use of BI-RADS among radiologists differs (39–44); however, we used standardized definitions developed and approved by the national Breast Cancer Surveillance Consortium.

Our study also has several strengths. We were able to examine interpretive performance in clinical practice without having to rely on test sets, in which measures of performance might not reflect actual clinical performance. We obtained detailed standardized clinical information for each patient, so that we were able to adjust radiologist measures of interpretive performance for differences in patient characteristics. Last, breast cancer outcomes were obtained on essentially all women, allowing us to identify the false-negative examinations accurately and therefore to calculate sensitivity.

Interest has been expressed in creating specialized regional breast imaging centers of excellence in the United States, in which experienced and high-volume breast-imaging specialists could provide multidisciplinary and coordinated breast cancer care (1). Although our study supports this idea in that it suggests that such specialists might have relatively high sensitivity and low false-positive rates, it may not be a feasible option in rural areas in which community hospitals or imaging facilities must provide all radiology services, including mammography, and cannot afford to specialize in breast imaging. In fact, the vast majority of mammograms in the United States are interpreted by general radiologists who interpret mammograms as a small percentage of their practice (31). To realistically improve mammography interpretation for women across the United States, we need to identify ways to improve accuracy and reduce variability among all radiologists who interpret mammography. Research on specific continuing medical education strategies in mammography, such as the American College of Radiology's Mammography Interpretive Skills Assessment (45), have not been rigorously evaluated, and so it is not known whether these programs are associated with improved interpretive performance. Academic detailing, in which experts in the field spend time with clinicians one-on-one or in small groups to review complex cases and improve techniques, has been successfully used in many areas other than mammography to change the behavior of physicians in practice (46–48). Future research should focus on the impact of different educational interventions, such academic detailing, interactive case-based education, double reading, and direct feedback on radiologists' interpretive performance via audit data, so that efforts can be made to improve the overall accuracy of mammography interpretation.

References

(1) Nass SJ, Ball J, editors. Improving breast imaging quality standards. Washington (DC): The National Academies Press; 2005.

(2) Sickles EA, Miglioretti DL, Ballard-Barbash R, Geller BM, Leung JW, Rosenberg RD, et al. Performance benchmarks for diagnostic mammography. Radiology 2005;235:775–90.

(3) Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002;224:861–9.

(4) Jensen A, Vejborg I, Severinsen N, Nielsen S, Rank F, Mikkelsen GJ, et al. Performance of clinical mammography: a nationwide study from Denmark. Int J Cancer 2006;119:183–91.

(5) Houssami N, Irwig L, Simpson JM, McKessar M, Blome S, Noakes J. The influence of clinical information on the accuracy of diagnostic mammography. Breast Cancer Res Treat 2004;85:223–8.

(6) Burnside ES, Sickles EA, Sohlich RE, Dee KE. Differential value of comparison with previous examinations in diagnostic versus screening mammography. AJR Am J Roentgenol 2002;179:1173–7.

(7) Houssami N, Irwig L, Simpson JM, McKessar M, Blome S, Noakes J. The contribution of work-up or additional views to the accuracy of diagnostic mammography. Breast 2003;12:270–5.

(8) 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.

(9) 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.

(10) Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst 2004;96:1840–50.

(11) Ernster VL, Ballard-Barbash R, Barlow WE, Zheng Y, Weaver DL, Cutter G, et al. Detection of ductal carcinoma in situ in women undergoing screening mammography. J Natl Cancer Inst 2002;94:1546–54.

(12) American College of Radiology. American College of Radiology (ACR) Breast Imaging Reporting and Data System atlas (Bi-RADS atlas). Reston (VA): American College of Radiology; 2003.

(13) Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.

(14) Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time varying covariates. Biostatistics 2004;5:381–98.

(15) McCullagh P, Nelder JA. Generalized linear models 2nd ed. London: Chapman & Hall; 1989.

(16) Smith-Bindman R, Chu P, Miglioretti DL, Quale C, Rosenberg RD, Cutter G, et al. Physician predictors of mammographic accuracy. J Natl Cancer Inst 2005;97:358–67.

(17) Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika 2001;88:973–85.

(18) Neuhaus JM, McCulloch CE. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. J R Statist Soc B 2006;68:859–72.

(19) Spiegelhalter D, Thomas A, Best N editors. Win-BUGS version 1.2 user manual ed. Cambridge: Medical Research Council Biostatistics Unit; 1999.

(20) Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. London: Chapman & Hall; 1995.

(21) Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20: 2865–84.

(22) Pepe MS, editor. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press; 2003.

(23) Reeves MJ, Newcomb PA, Remington PL, Marcus PM. Determinants of breast cancer detection among Wisconsin (United States) women, 1988–90. Cancer Causes Control 1995;6:103–11.

(24) Rosenberg A, Burke L, Vox P, Liles D. Self-exam is the most common method of breast cancer identification. J Clin Oncol 2007;25:1543.

(25) Breen N, Yabroff KR, Meissner HI. What proportion of breast cancers are detected by mammography in the United States? Cancer Detect Prev 2007;31:220–4.

(26) Reeves MJ, Newcomb PA, Remington PL, Marcus PM, MacKenzie WR. Body mass and breast cancer. Relationship between method of detection and stage of disease. Cancer 1996;77:301–7.

(27) Maibenco D, Daoud Y, Phillips E, Saxe A. Relationship between method of detection of breast cancer and stage of disease, method of treatment, and survival in women aged 40 to 49 years. Am Surg 1999;65:1061–6.

(28) McPherson CP, Swenson KK, Jolitz G, Murray CL. Survival of women ages 40-49 years with breast carcinoma according to method of detection. Cancer 1997;79:1923–32.

(29) Coates RJ, Uhler RJ, Brogan DJ, Gammon MD, Malone KE, Swanson CA, et al. Patterns and predictors of the breast cancer detection methods in women under 45 years of age (United States). Cancer Causes Control 2001;12:431–42.

(30) Newcomer LM, Newcomb PA, Trentham-Dietz A, Storer BE, Yasui Y, Daling JR, et al. Detection method and breast carcinoma histology. Cancer 2002;95:470–7.

(31) Lewis RS, Sunshine JH, Bhargavan M. A portrait of breast imaging specialists and of the interpretation of mammography in the United States. AJR Am J Roentgenol 2006;187:W456–68.

(32) Leung JW, Margolin FR, Dee KE, Jacobs RP, Denny SR, Schrumpf JD. Performance parameters for screening and diagnostic mammography in a community practice: are there differences between specialists and general radiologists? AJR Am J Roentgenol 2007;188:236–41.

(33) Tan A, Freeman DH Jr, Goodwin JS, Freeman JL. Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment. Breast Cancer Res Treat 2006;100:309–18.

(34) Elmore JG, Miglioretti DL, Reisch 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.

(35) Theberge I, Hebert-Croteau N, Langlois A, Major D, Brisson J. Volume of screening mammography and performance in the Quebec population-based Breast Cancer Screening Program. CMAJ 2005;172:195–9.

(36) Kan L, Olivotto IA, Warren Burhenne LJ, Sickles EA, Coldman AJ. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. Radiology 2000;215:563–7.

(37) Coldman AJ, Major D, Doyle GP, D'yachkova Y, Phillips N, Onysko J, et al. Organized breast screening programs in Canada: effect of radiologist reading volumes on outcomes. Radiology 2006;238:809–15.

(38) Rickard M, Taylor R, Page A, Estoesta J. Cancer detection and mammogram volume of radiologists in a population-based screening programme. Breast 2006;15:39–43.

(39) Geller BM, Ichikawa LE, Buist DS, Sickles EA, Carney PA, Yankaskas BC, et al. Improving the concordance of mammography assessment and management recommendations. Radiology 2006;241:67–75.

(40) Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value. Radiology 2006;239:385–91.

(41) Ciatto S, Houssami N, Apruzzese A, Bassetti E, Brancato B, Carozzi F, et al. Reader variability in reporting breast imaging according to BI-RADS assessment categories (the Florence experience). Breast 2006;15:44–51.

(42) 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.

(43) Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol 2000;174:1769–77.

(44) Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles EA, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst 1998;90:1801–9.

(45) Sickles EA. The American College of Radiology's Mammography Interpretive Skills Assessment (MISA) examination. Semin Breast Dis 2003;6:133–39.

(46) Broadhurst NA, Barton CA, Rowett D, Yelland L, Matin DK, Gialamas A, et al. A before and after study of the impact of academic detailing on the use of diagnostic imaging for shoulder complaints in general practice. BMC Fam Pract 2007;8:12.

(47) Simon SR, Rodriguez HP, Majumdar SR, Kleinman K, Warner C, Salem-Schatz S, et al. Economic analysis of a randomized trial of academic detailing interventions to improve use of antihypertensive medications. J Clin Hypertens (Greenwich) 2007;9:15–20.

(48) Schuster RJ, Terwoord NA, Tasosa J. Changing physician practice behavior to measure and improve clinical outcomes. Am J Med Qual 2006;21: 394–400.

Funding

Agency for Healthcare Research and Quality (HS10591); National Cancer Institute (1R01 CA107623 and 1K05 CA104699); Breast Cancer Surveillance Consortium (U01 CA63731, U01 CA86082, U01 CA63736, and U01 CA86076).

Notes

The authors had full responsibility in the design of the study, the collection of the data, the analysis and interpretation of the data, the decision to submit the manuscript for publication, and the writing of the manuscript.

Manuscript received July 12, 2007; revised October 9, 2007; accepted October 25, 2007.