Physician Predictors of Mammographic Accuracy

Rebecca Smith-Bindman, Philip Chu, Diana L. Miglioretti, Chris Quale, Robert D. Rosenberg, Gary Cutter, Berta Geller, Peter Bacchetti, Edward A. Sickles, Karla Kerlikowske

Background: **The association between physician experience and the accuracy of screening mammography in community practice is not well studied. We identified characteristics of U.S. physicians associated with the accuracy of screening mammography.** ***Methods:*** **Data were obtained from the Breast Cancer Surveillance Consortium and the American Medical Association Master File. Unadjusted mammography sensitivity and specificity were calculated according to physician characteristics. We modeled mammography sensitivity and specificity by multivariable logistic regression as a function of patient and physician characteristics. All statistical tests were two-sided.** ***Results:*** **We studied 209 physicians who interpreted 1 220 046 screening mammograms from January 1, 1995, through December 31, 2000, of which 7143 (5.9 per 1000 mammograms) were associated with breast cancer within 12 months of screening. Each physician interpreted a mean of 6011 screening mammograms (95% confidence interval [CI] = 4998 to 6677), including a mean of 34 (95% CI = 28 to 40) from women diagnosed with breast cancer. The mean sensitivity was 77% (range = 29%–97%), and the mean false-positive rate was 10% (range = 1%–29%). After adjustment for the patient characteristics of those whose mammograms they interpreted, physician characteristics were strongly associated with specificity. Higher specificity was associated with at least 25 years (versus less than 10 years) since receipt of a medical degree (for physicians practicing for 25–29 years, odds ratio [OR] = 1.54, 95% CI = 1.14 to 2.08; *P* = .006), interpretation of 2500–4000 (versus 481–750) screening mammograms annually (OR = 1.30, 95% CI = 1.06 to 1.59; *P* = .011) and a high focus on screening mammography compared with diagnostic mammography (OR = 1.59, 95% CI = 1.37 to 1.82; *P*<.001). Higher overall accuracy was associated with more experience and with a higher focus on screening mammography. Compared with physicians who interpret 481–750 mammograms annually and had a low screening focus, physicians who interpret 2500–4000 mammograms annually and had a high screening focus had approximately 50% fewer false-positive examinations and detected a few less cancers.** ***Conclusion:*** **Raising the annual volume requirements in the Mammography Quality Standards Act might improve the overall quality of screening mammography in the United States. [J Natl Cancer Inst 2005;97:358–67]**

Screening mammography is a nonspecific test for breast cancer, because only 5%–10% of screening mammograms that are interpreted as abnormal harbor cancer *(1–5)*. Although patient characteristics such as age and breast density contribute to variations in reported mammographic accuracy *(1,6,7)*, it is not clear how physician characteristics affect variability in accuracy.

A growing body of evidence has shown that physicians with greater experience in performing procedures, such as cardiac angioplasty *(8)*, have a higher proportion of patients with good outcomes *(9)*. Physician training in mammographic interpretation has been associated with improved accuracy *(10,11)*. The few studies that have evaluated the relationship between annual volume of mammographic interpretation and accuracy, however, have obtained conflicting results. Some studies have reported that volume is of prime importance *(12,13)*, whereas others have reported that accuracy is associated with the interplay of many interrelated factors involving physician experience but that volume itself is not important *(14,15)*. However, all of these studies *(12–15)* used practice sets of mammograms that were greatly enriched with mammograms showing cancer; some of these practice sets contained up to 100 times more cancer-associated mammograms than generally encountered in actual practice, which raises concerns about context bias *(16,17)*. Two studies evaluated the association between mammographic volume and accuracy with the prospective interpretation of clinical mammograms by a small number of physicians *(18,19)* and found that physicians who read higher volumes of mammograms tended to have improved accuracy. No large study has evaluated the association between physicians' volume and accuracy by use of prospectively collected clinical data in the United States on a broad sample of physicians.

In the United States, the Mammography Quality Standards Act of 1992 requires physicians to interpret at least 960 mammographic examinations within a 2-year period to qualify to interpret mammograms *(20)*. This minimum is 10-fold lower than the number required by the United Kingdom National Health Service Breast Screening Program *(21)* and reflects a minimum volume of approximately 10 mammograms per week. Although it seems reasonable to assume that increasing experience will improve the accuracy of mammographic interpretation, the values chosen by the Mammography Quality Standards Act and the National Health Service Breast Screening Program were arbitrary

Affiliations of authors: Departments of Radiology (RS-B, PC, CQ, EAS) and Epidemiology and Biostatistics (RS-B, PB, KK), University of California, San Francisco, CA; Center for Health Studies, Group Health Cooperative and Department of Biostatistics, University of Washington, Seattle, WA (DLM); Department of Radiology, University of New Mexico, Albuquerque, NM (RDR); Center for Research Design and Statistical Methods, University of Nevada School of Medicine, Applied Research Facility, Reno, NV (GC); Health Promotion Research, University of Vermont, College of Medicine, Burlington, VT (BG); General Internal Medicine Section, Department of Veterans Affairs, University of California, San Francisco, CA (KK).

Correspondence to: Rebecca Smith-Bindman, MD, Department of Radiology, University of California, San Francisco, 1600 Divisadero St., San Francisco, CA 94115 (e-mail: Rebecca.Smith-Bindman@Radiology.UCSF.Edu).

See "Notes" following "References."

DOI: 10.1093/jnci/dji060

minima derived primarily from perceptions about the supply of physicians able to interpret mammograms rather than from actual data to ensure adequate practice and skill *(22)*. The purpose of this study was to evaluate physician predictors associated with accuracy of screening mammographic interpretation in community practice in the United States.

Patients and Methods

Data Sources

We obtained data on mammographic interpretations, volume and cancer outcomes from mammography registries that participate in the Breast Cancer Surveillance Consortium *(1,23,24)*, a National Cancer Institute–funded consortium that collects patient demographic and clinical information *(25)*, mammographic interpretation, and cancer diagnoses from participating facilities in seven states. Four registries—Colorado (Colorado Mammography Project), New Mexico (New Mexico Mammography Project), San Francisco (San Francisco Mammography Registry), and Vermont (Vermont Breast Cancer Surveillance System)—contributed data to this study. Details of data collection have been reported previously *(1,26–30)*. The Breast Cancer Surveillance Consortium links data within registries from patient surveys and radiologist reports and ascertains cancer outcomes through linkage with state tumor registries (Colorado and Vermont), Surveillance Epidemiology and End Results (SEER[1]) tumor registries (San Francisco and New Mexico), and pathology databases (Vermont and New Mexico).

Physician characteristics (age and years since receipt of medical degree) were obtained from the American Medical Association Physician Profile Service *(31)*. Linkage with the Breast Cancer Surveillance Consortium data was done in a way that maintained physician confidentiality. Institutional Review Boards of all collaborating institutions approved the study.

Subjects

The study subjects were physicians who interpreted screening mammograms between January 1, 1995, and December 31, 2000. Overall, 95% of physicians who practice at facilities that participate in the Breast Cancer Surveillance Consortium were included in the analysis. We excluded screening examinations that occurred after December 31, 2000, to ensure at least 12 months follow-up for a cancer diagnosis after a normal or abnormal screening result and an additional 18 months for the cancer to be reported to the tumor registries, which would provide a cancer ascertainment that was at least 94.3% complete *(26)*. We assumed that all physicians interpreted an average of at least 480 mammograms per year, the minimum number required by Mammography Quality Standards Act guidelines, although a particular mammography registry may not capture all interpretations. Consequently, we excluded 45 physicians who appeared to interpret less than an average of 480 mammograms annually or during each year of the study period, because the volume of mammographic interpretations estimated for these physicians is likely to be inaccurate. The mean annual volume of the 45 excluded physicians was 388 mammographic interpretations (95% confidence interval [CI] = 372 to 405 mammographic interpretations). For any physician, we also excluded any calendar year during which that physician interpreted less than 300 mammograms. For example, a physician who read 1200, 1100, 200, and 1300 mammograms in each year of the 4-year study would be included, but his or her accuracy and annual volume would not be assessed during the third year.

Among the 209 physicians, the mean age (± standard deviation) was 52.2 ± 9.6 years, the mean number of years since receipt of a medical degree was 24.5 ± 10.6 years, and 46 were female (Table 1).

Mammographic Volume and Screening Focus

We calculated each physician's mean annual volume of mammographic interpretations (including both screening and diagnostic examinations) over the study period and then stratified annual volume into groups that had been used by others *(13,18)*, and we roughly balanced the number of physicians in each group when possible. The mean annual volume of mammographic interpretations was 1572, and the mean ranged from 1397 to 1928 across the four registries ($P = .01$). The median annual volume was 1054, and the median ranged from 835 to 1682 across the four registries ($P = .01$). Of the 209 physicians, 63 (30.1%) interpreted 481–750 mammograms annually, for a total of 123 789 (10.2%) of all 1 220 046 screening mammograms in this study. An additional 32 physicians (15.3%) interpreted 751–1000 mammograms annually, for a total of 91 801 (7.5%) of all 1 220 046 screening mammograms. Thus, 95 (45.4%) of the physicians interpreted fewer than 1001 mammograms annually, and these physicians interpreted 17.7% of all screening mammograms.

We assessed each physician's relative focus on screening as opposed to diagnostic mammography as their ratio of screening to diagnostic mammograms interpreted. The median ratio of screening to diagnostic mammographic examinations was 5.6 (interquartile range = 4.2–7.6), and this ratio was comparable across the four registries. We dichotomized this ratio at 5 (<5

Table 1. Characteristics of physicians included in this study

Characteristic	No. (%)
Sex	
Male	163
Female	46 (22.0)
Physician age	
<40 y	22 (11.3)
40–49 y	60 (30.8)
50–59 y	73 (37.4)
60–69 y	33 (16.9)
≥70 y	7 (3.6)
Time since receipt of medical degree	
<10 y	16 (07.7)
10–14 y	26 (12.4)
15–19 y	37 (17.7)
20–24 y	27 (12.9)
25–29 y	24 (11.5)
30–34 y	43 (20.6)
>34 y	35 (17.2)
Average annual volume of mammogram interpretation	
481–750 mammograms	63 (30.1)
751–1000 mammograms	32 (15.3)
1001–1500 mammograms	41 (19.6)
1501–2500 mammograms	43 (20.6)
2501–4000 mammograms	16 (7.7)
>4000 mammograms	14 (6.7)
Ratio of screening to diagnostic mammograms	
<5	81 (0.3)
≥5	128 (0.6)

vs. ≥5) as a round cut point that approximately balanced the numbers of physicians in these two groups.

Screening Mammography Accuracy

We calculated annual volume and screening focus from all of a physician's interpretations but restricted the analysis of mammography accuracy to screening examinations. We considered mammograms to be diagnostic whenever the woman reported a breast symptom [consistent with the American College of Radiology Breast Imaging Reporting and Data Systems (BI-RADS) *(32)*] or the mammogram occurred within 9 months of a previous screening examination. Women could have more than one screening examination included as long as the interval between examinations was more than 9 months.

A screening mammogram was classified as positive *(32)* if the initial assessment was incomplete or suspicious for cancer (BI-RADS interpretations 0, 4, or 5; n = 92 439 or 7.6% of total screening mammograms) or if the initial assessment was "probably benign" (BI-RADS interpretation 3) but had a recommendation for immediate further assessment (n = 27 753 or 2.3% of total screening mammograms). The remaining mammograms were classified as negative. Mammograms without a BI-RADS assessment were excluded from the analyses (0.10% of total screening mammograms). Women were considered to have breast cancer if reports from a breast pathology database, SEER program, or state tumor registry showed invasive carcinoma or ductal carcinoma in situ within 12 months of the index mammogram.

If breast cancer was diagnosed within 12 months of a positive screening mammogram, the mammogram was considered a true positive. If breast cancer was diagnosed within 12 months of a negative screening mammogram, the mammogram was considered a false negative. If no breast cancer was diagnosed within 12 months of a negative screening mammogram, the mammogram was considered a true negative. If no breast cancer was diagnosed within 12 months of a positive screening mammogram, the mammogram was considered a false positive.

To adjust each physician's accuracy according to the characteristics of his or her patients, we included patient age, physician-reported assessment of breast density, and a classification of mammographic examination as a first or a subsequent examination in our multivariable models. Breast density was classified as almost entirely fat, scattered fibroglandular densities, heterogeneously dense, or extremely dense. A mammogram was considered a patient's "first" mammogram if there was no registry record of a prior mammogram within 4 years and if the patient reported no prior mammogram within 4 years. Remaining mammograms were considered subsequent.

Statistical Analysis

We calculated the overall sensitivity and specificity of screening mammography for each physician. Whenever the value in any cell was equal to zero, we added 0.5 to the value in all cells to obtain a less extreme value. Unadjusted mammographic sensitivity and specificity were calculated according to patient characteristics (age, breast density, and whether examination was a first or a subsequent) and physician characteristics (age, years since receipt of medical degree, average annual volume of mammogram interpretations, and ratio of screening to diagnostic mammographic interpretations). We plotted the sensitivity against the false-positive rate of screening mammography, with each physician contributing a single point to this graph. We then graphed the sensitivity and false-positive rate of screening mammograms stratified by physician characteristics, with each mammogram weighed equally.

We modeled sensitivity and specificity as a function of patient and physician characteristics by use of multivariable logistic regression. Because of the colinearity of physician age and time since receipt of medical degree, only the latter was included in the multivariable analysis. To determine whether patient and physician characteristics influence the threshold at which a physician operates (which results in a tradeoff between sensitivity and specificity) or the accuracy of mammographic interpretation (additional probability of a positive mammogram if a woman has cancer), we jointly modeled the false-positive rate (1 minus the specificity) and true-positive rate (sensitivity) in a single receiver operator characteristic (ROC)–type logistic regression model. This model included main effects for each covariate and cancer status plus interactions of each covariate with cancer status *(33)*. Specifically,

$$\text{logit}[p(y_i = 1 \mid x_i, d_i)] = x_i\boldsymbol{\beta} + x_i d_i \boldsymbol{\delta},$$

where y_i is the mammography outcome (1 if positive, 0 if negative) for the ith woman, $\boldsymbol{x}i$ is a vector of her covariate values including an intercept term, and d_i is an indicator of whether or not she had cancer diagnosed during the 1-year follow-up period. By use of this notation, the false-positive rate for the covariate combination × is defined as $p(y = 1|\boldsymbol{x}, d = 0)$, which is equal to the inverse logit of $\boldsymbol{x\beta}$. Sensitivity is $p(y = 1|\boldsymbol{x}, d = 1)$, which is equal to the inverse logit of $\boldsymbol{x}(\boldsymbol{\beta} + \boldsymbol{\delta})$. Thus, the β coefficients measure the influence of × on the overall probability of a recall (i.e., threshold effect), and $\boldsymbol{\delta}$ measures the additional influence of × on the probability of a recall given that the woman has cancer (i.e., accuracy effect). If $\boldsymbol{\delta} = 0$, then the covariate × influences the false-positive rate and sensitivity equally. This model allowed us to evaluate differences in interpretive performance that reflect a threshold effect (i.e., a shift along an ROC curve; in Fig. 1, movement from point A to point B) versus an accuracy effect (i.e., differences that reflect performance on a different ROC curve; in Fig. 1, movement from point A to point C). We report multivariable results for specificity, sensitivity, and overall accuracy. Odds ratios (ORs) for sensitivity and specificity reflect how well physicians performed with respect to a given covariate along an ROC curve (if the accuracy effect is not statistically significant), whereas odds ratios for accuracy reflect a shift associated with a given covariate to a new ROC curve. For example, given an overall ROC curve for physicians, a statistically significant positive accuracy effect means a given covariate is associated with a shift to a different ROC curve that reflects better performance. An improvement in accuracy can reflect a statistically significant increase in the specificity without a corresponding statistically significant reduction in the sensitivity, a statistically significant increase in the sensitivity without a statistically significant decrease in the specificity, or an improvement in both sensitivity and specificity. If the accuracy effect is not statistically significantly different from 1, changes in specificity or sensitivity associated with a covariate reflect a shift along an ROC curve as opposed to a shift to a different ROC curve (Fig. 1). The models were fit by way of generalized estimating equations *(34)* with an independent working covariance matrix by

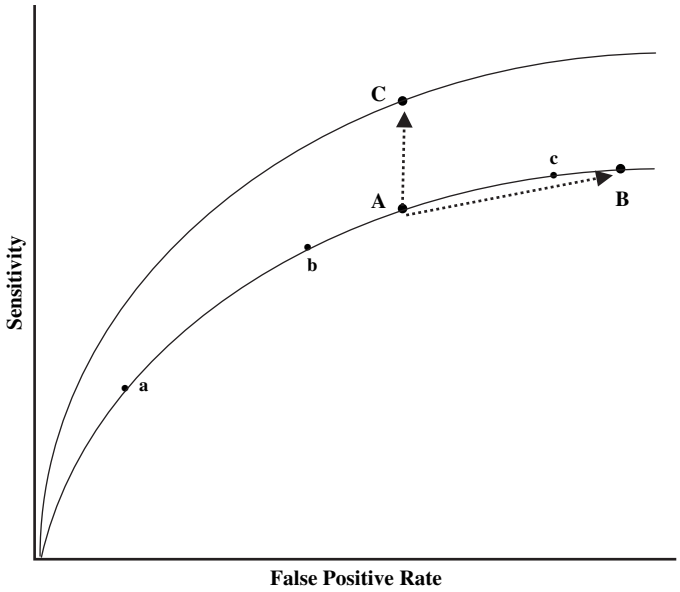

Fig. 1. Mammography accuracy and the interpretation threshold. Differences in physician performance that reflect an improvement in accuracy are shown by the shift from point A to C. Differences in physician performance that reflect a shift in the threshold used to interpret an examination as abnormal are shown by a shift from point A to B.

use of the GENMOD procedure in the SAS package (version 8.2; SAS Institute, Cary NC) of programs to account for the correlation among multiple mammograms interpreted by the same physician.

To demonstrate the real-world implications of differences in accuracy, we used the estimated sensitivity, specificity, and positive predictive values for all possible combinations of annual volume and screening focus to calculate expected numbers of cancers detected and false-positive diagnoses per 10 000 women screened annually, standardized to a single population of women with the covariate distribution and the same number of cancers (5.9 cancers per 1000 mammograms) as observed in this cohort. All statistical tests were two-sided.

Results

The study subjects were 209 physicians who interpreted 1 220 046 screening mammograms between January 1, 1995, and December 31, 2000, including 7143 (5.9 per 1000 mammograms) diagnosed as breast cancer within 12 months of the screening mammogram. Each physician interpreted a mean of 6011 screening mammograms (95% CI = 4998 to 6677) of which a mean of 34 (95% CI = 28 to 40) were from women diagnosed with breast cancer within 12 months of the index mammogram.

Sensitivity and Specificity of Mammography by Patient Characteristics

The sensitivity and specificity of mammographic interpretation varied substantially and statistically significantly by patient characteristics (Table 2). For example, for subsequent screening mammograms, as patient age increased from younger than 40 years to older than 70 years, the false-positive rate decreased from 10.5% (95% CI = 10.1 to 10.9) to 6.5% (95% CI = 6.4 to 6.6) and the sensitivity increased from 52.7% (95% CI = 39.5 to 65.9) to 79.7% (95% CI = 77.6 to 81.9). The false-positive rate was lower, and sensitivity was higher when breast density was predominantly fat or contained scattered fibroglandular densities. Lower false-positive rates were observed for subsequent examinations than for first examinations, whereas higher sensitivities were observed for first screening examinations.

Physician Variability in Mammography Sensitivity and False-Positive Rates

Physicians exhibited wide variations in mammography sensitivity and specificity. The mean sensitivity was 77% (range = 29%–97%, 95% CI = 76% to 79%), and the mean false-positive rate was 10% (range = 1%–29%, 95% CI = 9% to 10%). The

Table 2. Accuracy of screening mammography by patient characteristics

	First screening mammogram					Subsequent screening mammograms				
				Likelihood ratio					Likelihood ratio	
	No.	Sensitivity, % (95% CI)*	False-positive rate, % (95% CI)	Positive	Negative	No.	Sensitivity, % (95% CI)	False-positive rate, % (95% CI)	Positive	Negative
Patient age†										
<40 y	51 494	84.6 (79.2 to 90.1)	15.0 (14.7 to 15.3)	5.64	0.18	18 454	52.7 (39.5 to 65.9)	10.5 (10.1 to 10.9)	5.0	0.53
40–49 y	131 272	81.6 (78.4 to 84.8)	13.5 (13.3 to 13.7)	6.05	0.21	246 198	68.6 (65.4 to 71.8)	9.2 (9.1 to 9.4)	7.4	0.35
50–59 y	69 150	82.5 (79.2 to 85.8)	12.5 (12.3 to 12.8)	6.60	0.20	278 559	75.2 (72.9 to 77.4)	8.4 (8.3 to 8.5)	8.9	0.27
60–69 y	43 162	85.6 (82.1 to 89.1)	11.3 (11.0 to 11.6)	7.58	0.16	178 278	77.1 (74.7 to 79.4)	7.6 (7.5 to 7.7)	10.1	0.25
≥70 y	41 038	87.9 (85.4 to 90.5)	9.7 (9.4 to 10.0)	9.07	0.13	162 441	79.7 (77.6 to 81.9)	6.5 (6.4 to 6.6)	12.3	0.22
Density‡										
Almost entirely fat	16 615	94.0 (89.0 to 99.1)	6.5 (6.1 to 6.9)	14.50	0.06	47 516	88.8 (83.3 to 94.3)	3.6 (3.4 to 3.7)	25.0	0.12
Scattered fibroglandular densities	73 156	90.1 (87.4 to 92.8)	12.1 (11.8 to 12.3)	7.46	0.11	243 996	82.3 (80.1 to 84.5)	7.5 (7.4 to 7.6)	11.0	0.19
Heterogeneously dense	56 936	82.0 (77.8 to 86.1)	12.5 (12.2 to 12.8)	6.55	0.22	180 201	72.4 (69.8 to 75.1)	8.8 (8.7 to 9.0)	8.2	0.30
Dense	19 708	77.8 (70.8 to 84.8)	14.7 (14.2 to 15.2)	5.29	0.26	45 643	65.9 (60.4 to 71.5)	10.5 (10.2 to 10.8)	6.3	0.38
Unknown	169 701	83.1 (81.0 to 85.2)	13.6 (8.5 to 13.7)	6.12	0.20	366 574	74 (72.2 to 75.9)	8.6 (8.5 to 8.7)	8.6	0.28

*CI = confidence interval.

†The point estimates changed little when calculated on the basis of a standardized distribution of breast density; therefore, the crude results are provided.

‡The point estimates changed little when calculated on the basis of a standardized distribution of patient age; therefore, the crude results are provided.

mean sensitivity for 95% of the physicians was between 48% and 95%, and the mean false-positive rate for 95% of the physicians was between 2% and 22%. Physicians with the highest false-positive rates tended to have the highest sensitivity, whereas physicians with the lowest false-positive rates tended to have the lowest sensitivity (Fig. 2). Thus, some of the difference among physician false-positive rates reflects their threshold for calling examinations abnormal (reflected as a tradeoff between sensitivity and specificity). However, some of the variation in sensitivity and specificity (and thus overall accuracy) was not the result of differences in threshold because at each false-positive rate, there was substantial variation in sensitivity between physicians. For example, at a false-positive rate of approximately 10%, the sensitivity ranged from 33% to 96%.

Sensitivity and Specificity of Mammography by Physician Characteristics

To identify physician characteristics that could explain the variation in physician accuracy, we first calculated physician sensitivity and specificity without adjusting for patient mixture. We found variations in the false-positive rates that paralleled physician experience (Fig. 3). In general, the false-positive rate declined (i.e., specificity improved) with increasing physician age, with increasing time since receipt of medical degree, and with increasing annual volume. For example, among subsequent screening mammograms (Fig. 3, B), the false-positive rate was 10.3% among physicians younger than age 40 years but only 6.8% among physicians aged 60–69 years. Additionally, physicians who had a higher focus on screening mammography than on diagnostic mammography had a lower false-positive rate (among subsequent examinations, 6.7% vs. 10.2%). Differences in sensitivity by physician experience were smaller and the confidence intervals largely overlapped, suggesting that the differences were not statistically significant.

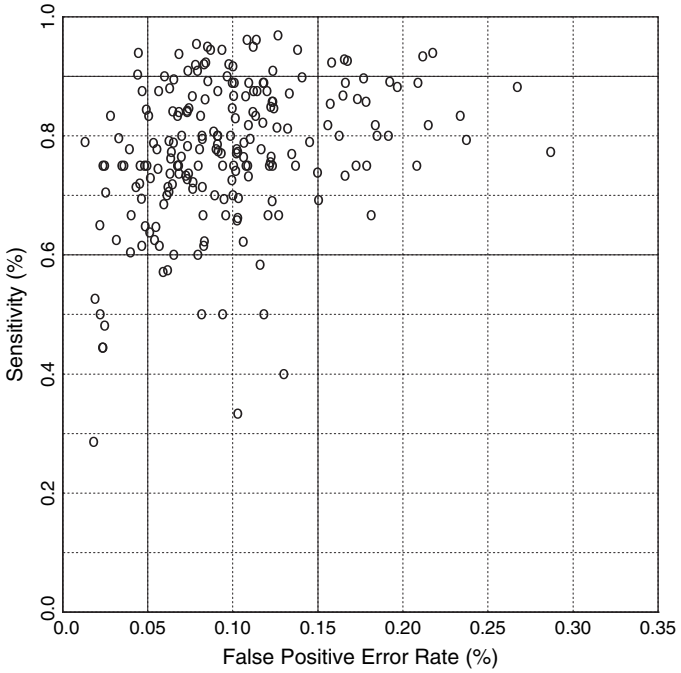

Fig. 2. Sensitivity versus the false-positive rate of screening mammography interpretation. Each physician is represented by a single point.

Combined Effects of Patient and Physician Characteristics on the Sensitivity and Specificity of Mammography

From the multivariable logistic regression analysis, several patient characteristics were associated with specificity (Table 3). A statistically significant increase in specificity was associated with an increase in patient age, with subsequent examinations, and with a breast density that was almost entirely fat. The following physician characteristics were also associated with a statistically significant increase in specificity: at least 25 years (versus less than 10 years) since receipt of medical degree (for physicians 25–29 years, OR = 1.54, 95% CI = 1.14 to 2.08), interpretation of 2500–4000 (versus 481–750) mammograms annually (OR = 1.30, 95% CI = 1.06 to 1.59), and a higher focus on screening mammography than on diagnostic mammography (OR = 1.59, 95% CI = 1.37 to 1.82). Interpretation of 1500–2500 mammograms was associated with a non–statistically significant improvement in specificity (OR = 1.16, 95% CI = 0.97 to 1.39).

Several patient characteristics were strongly associated with sensitivity. Increased sensitivity was associated with increased patient age, with first mammographic examinations, and with a breast density that was almost entirely fat or contained scattered fibroglandular densities. Physician characteristics were less consistently associated with sensitivity. A higher focus on screening mammography than on diagnostic mammography was associated with a lower sensitivity (OR = 0.82, 95% CI = 0.69 to 0.98), but sensitivity was not statistically significantly associated with a physician's annual volume or time since receipt of medical degree.

Overall accuracy is presented in Table 3. A statistically significant increase in overall accuracy was associated with a patient age older than 50 years and with breast density other than extremely dense. A statistically significant increase in overall accuracy was associated with 25–35 years since receipt of medical degree (e.g., for 25–29 years since receipt of their medical degree, OR for accuracy = 1.54, 95% CI = 1.05 to 2.26; P = .025). This result primarily reflects improved specificity (OR = 1.54, 95% CI = 1.14 to 2.08; P = .006) without a statistically significant change in sensitivity (OR = 1.0, 95% CI = 0.72 to 1.40; Table 3). A statistically significant increase in accuracy was also associated with a higher focus on screening mammography than on diagnostic mammography (OR = 1.29, 95% CI = 1.08 to 1.55), reflecting a statistically significant increase in specificity (OR = 1.59, 95% CI = 1.37 to 1.82) with a smaller reduction in sensitivity (OR = 0.82, 95% CI = 0.69 to 0.98). There was no statistically significant difference in accuracy as a function of physicians' annual volume (none of the groups was different than the lowest volume category), suggesting that the differences in specificity by annual volume largely reflect differences among physicians in their threshold for calling a mammogram abnormal. Interpretation of 751–1000 mammograms annually was associated with improved accuracy (OR = 1.33, 95% CI = 0.97 to 1.83), as characterized by small increases in both sensitivity (OR = 1.17, 95% CI = 0.87 to 1.56) and specificity (OR = 1.14, 95% CI = 0.93 to 1.41). However, this level of mammogram interpretation was not statistically significant (P = .08).

Association of Physician Experience with False-Positive Rates and Cancer Detection Rates

Physicians who had a higher focus on screening mammography than on diagnostic mammography or annual volume of

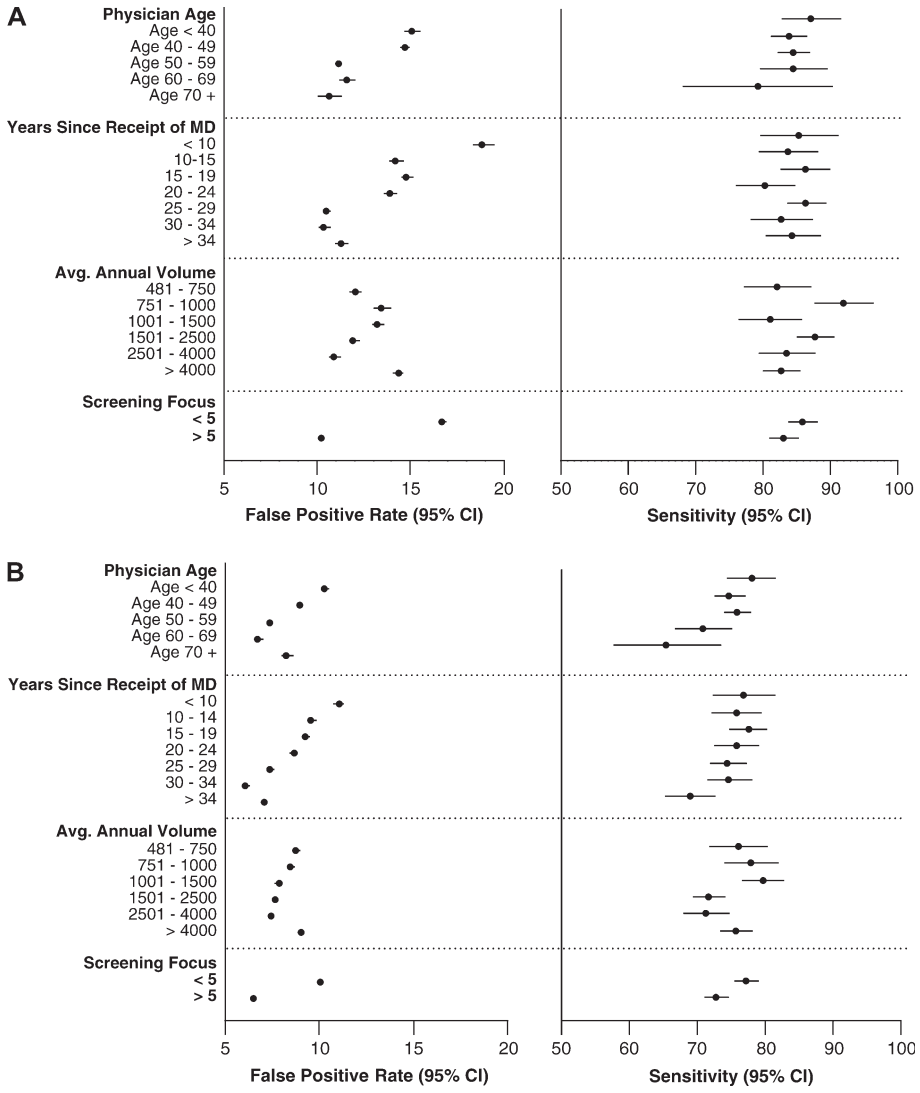

Fig. 3. False-positive rate and sensitivity (and 95% confidence intervals [CIs]) of screening mammography by physician characteristics for first (**A**) and subsequent (**B**) screening examinations. Error bars = 95% CIs. Some error bars are not visible because they do not extend beyond the symbol.

2500–4000 mammograms compared with 480–750 mammograms had lower false-positive rates. For physicians with a higher screening focus, this result reflects improved accuracy (defined as improved performance along a more accurate ROC curve). For physicians with a higher volume, this result reflects a shift along a ROC curve to operate in an area that emphasizes improved specificity. The difference in how these physicians perform will substantially affect the patients whose mammograms they interpret. Compared with physicians who interpret the minimum number of mammograms annually (i.e., 481–750 mammograms) and had a low screening focus (ratio less than 5), physicians who interpret 2500–4000 mammograms annually and had a high screening focus (ratio greater than or equal to 5) had approximately 50% fewer false-positive examinations (674 versus 1279 false-positive examinations per 10 000 screening examinations) and detected only a few less cancers (44 versus 47 per 10 000 screening examinations) (Table 4). Thus, a physician who interprets 3000 mammograms annually and has a high focus on screening mammography would have approximately 182 fewer false-positive examinations and would miss approximately one cancer per year, compared with a low-volume physician who does not focus to the same degree on screening mammography. A physician who interprets 1500–2500 mammograms annually and has a high focus on screening mammography would have approximately 40% fewer false-positive examinations and miss approximately one cancer per 5000 screening examinations, compared with the low-volume physician who does not focus to the same degree on screening mammography. These differences in sensitivity and specificity are reflected by the positive predictive value of mammography, which is nearly twice as high as in the high-volume, high-screening-focus category as in the low-volume, low-screening-focus category (6.1% vs. 3.6%).

Discussion

We found substantial physician variation in mammographic sensitivity and specificity that was not explained by the characteristics of patients whose mammograms they interpreted. The most dramatic difference was in the false-positive rate, which varied from 1% to 29%. In general, the most experienced physicians had the lowest false-positive rates. Physicians who had been practicing the longest, who interpreted 2500–4000 mammograms annually, and who emphasized screening, as opposed to diagnostic, mammography had lower false-positive rates than their less-experienced counterparts. For physicians who had practiced the longest and who had a high focus on screening mammography, overall accuracy was improved as well, meaning that they had higher specificity without an equal loss in sensitivity. For physicians

Table 3. Influence of patient and physician characteristics on the odds of a correct mammogram interpretation in women with and without breast cancer simultaneously adjusting for threshold*

	Specificity		Sensitivity		Accuracy†	
	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*
Patient age						
<40 y	0.90 (0.83 to 0.98)	<.012	0.94 (0.61 to 1.4)	.770	0.85 (0.56 to 1.28)	.428
40–49 y	1.0 (referent)		1.0 (referent)		1.0 (referent)	
50–59 y	**1.08** (1.04 to 1.10)	<.001	**1.25** (1.06 to 1.47)	.007	**1.34** (1.14 to 1.58)	<.001
60–69 y	**1.16** (1.11 to 1.20)	<.001	**1.36** (1.12 to 1.64)	.002	**1.57** (1.29 to 1.91)	<.001
≥70 y	**1.32** (1.25 to 1.41)	<.001	**1.51** (1.26 to 1.81)	<.001	**2.00** (1.66 to 2.40)	<.001
Screening						
First	1.0 (referent)		1.0 (referent)		1.0 (referent)	
Subsequent	**1.59** (1.49 to 1.67)	<.001	**0.52** (0.45 to 0.61)	<.001	**0.82** (0.70 to 0.98)	.024
Density						
Almost entirely fat	**2.38** (1.67 to 3.33)	<.001	**3.98** (2.20 to 7.17)	<.001	**9.37** (5.07 to 17.32)	<.001
Scattered fibroglandular densities	1.19 (0.91 to 1.54)	.200	**2.19** (1.64 to 2.93)	<.001	**2.60** (1.98 to 3.45)	<.001
Heterogeneously dense	1.05 (0.81 to 1.37)	.704	1.34 (1.00 to 1.79)	0.046	**1.41** (1.05 to 1.90)	0.024
Extremely dense	1.0 (referent)		1.0 (referent)		1.0 (referent)	
Unknown	0.91 (0.68 to 1.20)	.500	**1.48** (1.07 to 2.04)	.017	1.34 (0.99 to 1.82)	0.058
Time since receipt of medical degree						
<10 y	1.0 (referent)		1.0 (referent)		1.0 (referent)	
10–14 y	1.16 (0.88 to 1.54)	.282	0.98 (0.68 to 1.43)	.921	1.14 (0.73 to 1.78)	.564
15–19 y	1.22 (0.92 to 1.64)	.172	1.07 (0.79 to 1.46)	.654	1.31 (0.92 to 1.87)	.135
20–24 y	1.18 (0.88 to 1.59)	.276	0.96 (0.70 to 1.33)	.817	1.14 (0.78 to 1.64)	.501
25–29 y	**1.54** (1.14 to 2.08)	.006	1.00 (0.72 to 1.40)	.999	**1.54** (1.05 to 2.26)	.025
30–34 y	**1.67** (1.25 to 2.22)	<.001	0.86 (0.63 to 1.19)	.367	**1.44** (0.99 to 2.12)	.060
>34 y	**1.59** (1.12 to 2.22)	.008	0.76 (0.55 to 1.04)	.084	1.20 (0.82 to 1.73)	.347
Average annual volume mammogram interpretation						
481–750 mammograms	1.0 (referent)		1.0 (referent)		1.0 (referent)	
751–1000 mammograms	1.14 (0.93 to 1.41)	.216	1.17 (0.87 to 1.56)	.292	1.33 (0.97 to 1.83)	.080
1001–1500 mammograms	1.05 (0.85 to 1.30)	.657	1.07 (0.80 to 1.44)	.643	1.13 (0.87 to 1.46)	.373
1501–2500 mammograms	1.16 (0.97 to 1.39)	.092	0.91 (0.72 to 1.15)	.449	1.06 (0.86 to 1.32)	.571
2501–4000 mammograms	**1.30** (1.06 to 1.59)	.011	0.83 (0.63 to 1.10)	.197	1.08 (0.82 to 1.42)	.586
>4000 mammograms	1.03 (0.85 to 1.25)	.789	0.96 (0.74 to 1.23)	.719	0.98 (0.77 to 1.25)	.878
Ratio of screening to diagnostic mammographic interpretation						
<5	1.0 (referent)		1.0 (referent)		1.0 (referent)	
>5	**1.59** (1.37 to 1.82)	<.001	**0.82** (0.69 to 0.98)	.026	**1.29** (1.08 to 1.55)	.005

**P* values correspond with the odds ratio (OR) to the left. CI = confidence interval. Statistically significant ORs ($P<.05$) are shown in boldface type.

†Improved sensitivity at given specificity or improved specificity at given sensitivity.

who interpreted at a high volume (2500–4000 mammograms annually), the difference in performance reflected a shift in the threshold used by these physicians to interpret an examination as abnormal (thus, a shift along an ROC curve). The differences in sensitivity were of much smaller magnitude than the differences in the false-positive rate; consequently, the higher-volume physicians did not miss many cancers even with the higher threshold they used to interpret an examination as abnormal (approximately one cancer per year).

Our results have important implications for the practice of screening mammography. We estimated that, compared with physicians who interpreted the minimum number allowed by

Table 4. Estimated differences in patient outcomes stratified by physician differences in screening mammography*

Annual volume: No. mammograms interpreted	Focus on screening	Sensitivity, %	Specificity, %	False-positive rate, %	No. cancers detected	No. false-positive diagnoses	PPV
480–750	Low	80.8	87.1	12.9	47	1279	3.6
	High	77.7	91.4	8.6	45	855	5.0
750–1000	Low	83.0	88.5	11.5	49	1143	4.1
	High	80.1	92.4	7.6	47	759	5.8
1000–1500	Low	81.8	87.7	12.3	48	1226	3.8
	High	78.8	91.8	8.2	46	818	5.3
1500–2500	Low	79.4	88.7	11.3	46	1121	4.0
	High	76.1	92.5	7.5	45	744	5.6
2500–4000	Low	77.9	89.7	10.3	46	1020	4.3
	High	74.4	93.2	6.8	44	674	6.1
>4000	Low	80.1	87.4	12.6	47	1250	3.6
	High	76.9	91.6	8.4	45	834	5.1

*Estimates assume that 10 000 women underwent screening mammography, that the multivariable distribution of patient characteristics, and that the total number of cancers (5.9 per 1000 mammograms) was the same as it is in this cohort. PPV = positive predictive value.

Mammography Quality Standards Act (i.e., 480–750 mammograms per year) and who have a lower screening focus, physicians who interpret 2500–4000 mammograms annually and have a higher screening focus have 50% fewer false-positive diagnoses (168 vs. 320 per 2500 examinations) and miss approximately one cancer per 2500 mammograms interpreted. We found that physicians with a higher screening focus have substantially improved specificity, slightly lower sensitivity, and overall improved accuracy. Our results indicate that physicians who focus on screening are better at screening than those who do not. One possible explanation is that physicians who have a larger proportion of diagnostic examinations (i.e., a low screening focus) may expect higher underlying rates of cancer, which might lead them to recall a larger percentage of patients.

There is considerable debate over how to analyze data describing the accuracy of diagnostic testing. Although ROC analyses have been a mainstay of diagnostic imaging research, there are several limitations of this method for evaluating the accuracy of mammography. ROC curve analysis cannot be used to understand the actual sensitivity and specificity in clinical practice *(35)*, and some ROC analyses, such as those that rely on the area under the curve, assume that every location along an ROC curve is equivalent. For example, if physician a has a sensitivity of 20% and a false-positive rate of 1%, physician b has a sensitivity of 85% and a false-positive rate of 5%, and physician c has a sensitivity of 90% and a false-positive rate of 30% (Fig. 1), all physicians can be said to perform along a single ROC curve, with each physician using a different threshold to interpret mammograms as abnormal. Although the performance of all three physicians can be plotted on the same ROC curve, it is not the case that each point along the curve reflects equally desirable performance. Yet area under the ROC curve analysis would not detect differences between these physicians. Specificity will tend to impact many more individuals than sensitivity. Thus, for physician c, the slightly higher sensitivity needs to be weighed against the substantially higher false-positive rate, and the performances of physicians b and c should not be considered comparable. Lastly, in some instances, a clinically relevant improvement in test accuracy (such as an improvement in sensitivity with only a small change in specificity) may not be regarded as an improvement via a ROC curve analysis, if the curve appears relatively steep in that region so that both points fall along the same curve *(35)*. Thus, we used the calculated sensitivity and specificity of each physician as the important outcome, because they are clinically relevant and easily understood. We used ROC curve analysis to determine whether the differences we detected were caused by threshold differences between physicians. We have identified physician characteristics that are associated with accuracy (time since receipt of medical degree and a high focus on screening mammography), as well as physician characteristics that are associated with a shift along an ROC curve (high annual volume).

Our results are consistent with those of previous studies *(12,13)* that used practice sets and found that more experienced physicians have lower false-positive rates. Our findings are in contrast with those of Beam et al. *(15)* who used a practice set and found that the most recently trained physicians perform better and that annual volume is not an important predictor of accuracy. In that study, physicians' performance on the practice set differed dramatically from what we found in our study using actual clinical mammograms. The mean sensitivity of mammography was 90% in the Beam study (versus 77% with actual clinical mammograms in our study), and the mean false-positive rate was 38% (versus 10% with actual clinical mammograms in this study). Thus, mammogram interpretation in routine clinical practice appears to differ substantially from that the testing situation described in the Beam study *(15)* in which the high proportion of cancers probably lowers the threshold for interpreting examinations as abnormal *(1,16,17)*. Additionally, the Beam study's nonstandard analysis method (each mammogram, via its BI-RADS score, contributed several estimates to each physician's accuracy) could also account for the differing results. Lastly, given the ROC method used in the Beam study, the authors could not differentiate physicians who performed on the same ROC curve—i.e., who differed in characteristics that influenced the threshold but not the accuracy.

Our results support the three studies of mammographic accuracy and volume that used prospectively interpreted clinical data. Sickles et al. *(19)* demonstrated that three physicians with special training in mammography had lower false-positive rates and higher cancer detection rates than seven general physicians who each interpreted only sufficient numbers of mammograms to satisfy federal regulations. Kan et al. *(18)* demonstrated that the physicians in British Columbia, each of whom interpreted 2000–4000 mammograms annually, had lower false-positive rates than physicians who interpreted less than 2000 mammograms annually or more than 4000 annually. Théberge et al. (36) demonstrated that radiologists who read more than 1500 mammograms annually had higher breast cancer detection rates while maintaining lower false-positive rates. Our finding of improved specificity among more experience physicians agree with those of Barlow et al. *(37)*. Whereas we found that experienced physicians were also more accurate, they found that experienced physicians tended to increase the threshold they used to consider a mammogram abnormal without improved accuracy. Our results also differed with respect to annual volume. Paralleling the other measures of experience, we found that increased volume (up to 4000 mammograms per year) is associated with improved specificity, whereas Barlow et al. found that increased volume is associated with worse specificity but improved sensitivity. There are several differences in our research methods that may account for these differences. First, Barlow et al. used physician's self-reported annual volume, rather than actual volume, and physicians may have incorrectly estimated their annual volume. The physicians in the study of Barlow et al. reported reading many more mammograms than we found; 25% of physicians read fewer than 1000 mammograms annually in Barlow's study compared with 45% in our study. Similarly, whereas 37% of physicians in the Barlow study reported having read more than 2000 mammograms annually, we found only half as many physicians (21%) read at such high volumes. Although we may have underestimated annual volume for physicians who interpret mammograms at facilities that do not participate in the Breast Cancer Surveillance Consortium, we believe that this would have had only limited impact on overall estimates of annual volume. Three of the four registries that we included (Vermont, San Francisco, and New Mexico) have almost complete population-based capture of mammograms, and thus we almost certainly captured the majority of mammograms for those physicians in the study. Second, Barlow et al. used broad categories to characterize physician annual volume, combining all physicians with annual volumes of more than 2000 into a single category. We found, as have others *(18)*, that specificity improves as volume increases up to 4000

mammograms annually but that physicians with volumes of more than 4000 have worse specificity. Combining all physicians with volumes of more than 2000 mammograms annually could have masked trends. Additionally, volume was assessed in only a single year in the Barlow study, whereas we averaged physician volume over 4 years, to account for variability across the years. Lastly, Barlow et al. used ROC methodology similar to that used by Beam et al. *(15)*, in which the full range of BI-RADS assessments are analyzed by use of an ordinal regression model rather than by dichotomizing the interpretation as normal or abnormal as occurs in clinical practice. Surprisingly, by use of this ROC methodology, Barlow et al. found that patient age does not impact the accuracy of mammography, which contrasts with our work and the work of many others *(1)*. These unexpected results raise questions about the ROC results of that study.

Our study demonstrated that annual mammographic volume, time since receipt of medical degree, and a focus on screening mammography are important contributors to mammographic accuracy. However, these factors did not explain all of the variation in physician performance. Many other factors potentially contribute to mammographic accuracy, such as whether physicians regularly assess their outcomes (learn from their mistakes), which types of ongoing medical education they complete, and perhaps whether they have concerns about medical malpractice.

We recommend that there be explicit discussion of what the goals of mammography should be. Should physicians maximize sensitivity at the expense of having very high false-positive rates or should they maximize sensitivity while achieving a lower, but reasonable, false-positive rate? Some of the large variation that we found among physicians may reflect differences in their individual expectations about ideal mammography performance (with some physicians choosing to emphasize sensitivity at the expense of very high false-positive rates). If the goal is to maximize sensitivity while achieving a reasonable false-positive rate, one action could be to raise the minimum number of mammograms physicians must interpret annually. An argument against raising the minimum is that this approach would decrease the supply of physicians who can interpret mammograms. Our data, however, suggest that the impact would be small if the minimum level is raised moderately. For example, if the minimum level is raised to 750 mammograms annually, although 30% fewer physicians would interpret mammograms, only 10% more screening mammograms would have to be interpreted by the remaining higher-volume physicians. Although an annual volume of 2500 mammograms seems ideal from a performance perspective if minimizing the false-positive rate were a goal, this change would need to occur slowly to prevent a shortage of physicians who interpret mammograms. A compromise of 1500 mammograms might be a practical solution because it would probably lead to a substantial reduction in the false-positive rate (40% in our estimate) yet would not create as much of a burden on the remaining higher-volume physicians.

A strength of our study is that the data were collected from actual clinical practice in four geographic areas across the United States and that 95% of physicians in those areas who practice at facilities that participate in the Breast Cancer Surveillance Consortium were included in this analysis. A limitation of our study is that we do not know whether greater experience, higher annual volume, and a greater focus on screening mammography improve interpretations or whether the better physicians simply choose to interpret more examinations. That is, it is not possible to disentangle what is cause and what is effect. Nonetheless, physicians who are interpreting more screening mammograms are doing a better job. Another limitation is sample size; although our sample size was large, it was not large enough to look separately at ductal carcinoma in situ and invasive cancer.

Although some variation in physician performance is inevitable, the degree of variation that we found, particularly for the false-positive rates, is large. Consequently, finding and implementing interventions to minimize this variation should be a priority. The false-positive rate in the United States is higher than that in other countries *(38)*, and it is twice as high as the rate in the United Kingdom *(5)*, although cancer detection rates are similar in the two countries. One of the major factors producing these differences in rates between the United States and the United Kingdom could be the annual volume of mammograms interpreted by physicians. The median annual number of mammograms that physicians interpreted in our sample (1053 mammograms) contrasts starkly with the median annual number of mammograms that physicians interpret in the United Kingdom (7000 mammograms) *(21)*. In the United States, the minimum value required by the Mammography Quality Standards Act is very low, approximately two mammograms per clinical workday, and the mean is fewer than five mammograms per clinical workday. Most factors that influence the sensitivity of mammography are not easily modified, e.g., a woman's age, mammographic breast density, and a physician's years of experience. Physician volume and screening focus can be altered, particularly because the Mammography Quality Standards Act is actively involved in the monitoring of physician volume. Raising the annual volume requirements in the Mammography Quality Standards Act might improve the overall quality of screening mammography in the United States.

References

(1) Carney P, Miglioretti D, Yankaskas B, Kerlikowske K, Rosenberg R, Rutter C, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the performance of screening mammography. Ann Intern Med 2003;138:168–73.

(2) Kerlikowske K, Grady D, Barclay J, Sickles EA, Eaton A, Ernster V. Positive predictive value of screening mammography by age and family history of breast cancer. JAMA 1993;270:2444–50.

(3) May D, Lee N, Nadel M, Henson RM, Miller DS. The National Breast and Cervical Cancer Early Detection Program: report on the first 4 years of mammography provided to medically underserved women. AJR Am J Roentgenol 1998;170:97–104.

(4) Sickles E, Ominsky S, Sollitto R, Galvin HB, Monticciolo DL. Medical audit of a rapid-throughput mammography screening practice: methodology and results of 27,114 examinations. Radiology 1990;175:323–7.

(5) Smith-Bindman R, Chu PW, Miglioretti DL, Sickles EA, Blanks R, Ballard-Barbash R, et al. Comparison of screening mammography in the United States and the United Kingdom. JAMA 2003;290:2129–37.

(6) Elmore JG, Wells CK, Howard DH, Feinstein AR. The impact of clinical history on mammographic interpretations. JAMA 1997;277:49–52.

(7) Kerlikowske K, Grady D, Barclay J, Sickles ES, Ernster V. Likelihood ratios for modern screening mammography. Risk of breast cancer based on age and mammographic interpretation. JAMA 1996;276:39–43.

(8) Hannan EL, Racz M, Ryan TJ, McCallister BD, Johnson LW, Arani DT, et al. Coronary angioplasty volume-outcome relationships for hospitals and cardiologists. JAMA 1997;277:892–8.

(9) Halm EA, Lee C, Chassin MR. Is volume related to outcome in health care? A systematic review and methodologic critique of the literature. Ann Intern Med 2002;137:511–20.

(10) Linver MN, Paster SB, Rosenberg RD, Key CR, Stidley CA, King WV. Improvement in mammography interpretation skills in a community radiology practice after dedicated teaching courses: 2-year medical audit of 38,633 cases [erratum ain Radiology 1992;184:878]. Radiology 1992;184: 39–43.
(11) 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.
(12) Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? J Women's Health 1998;7: 443–9.
(13) Esserman L, Cowley H, Eberle C, Kirkpartrick A, Change S, Berbaum K, et al. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst 2002; 94.
(14) Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. Arch Intern Med 1996;156:209–13.
(15) Beam C, Conant E, Sickles E. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. J Natl Cancer Inst 2003;95:282–90.
(16) Egglin T, Feinstein A. Context bias. A problem in diagnostic radiology. JAMA 1996;276:1752–5.
(17) Elmore JG, Miglioretti DL, Carney PA. Does practice make perfect when interpreting mammography? Part II. J Natl Cancer Inst 2003;95:250–2.
(18) Kan L, Olivotto I, Burhenne LW, Sickles E, Coldman A. Standardized abnormal interpretation and cancer detection ratios to assess reading volume and reader performance in a breast screening program. Radiology 2000;215:563–7.
(19) Sickles E, Wolverton D, Dee K. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 2002;224:861–9.
(20) The Mammography Quality Standards Act of 1992, Pub. L. No. 102-539.
(21) Department of Health, U.K., Statistical Bulletin, Breast Screening Programme, England: 1999–2000. National Statistics; March 2000.
(22) National Mammography Quality Assurance Advisory Committee. Summary minutes, Washington, DC, January 23–25, 1995.
(23) 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.
(24) Barlow W, Lehman C, 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.
(25) Breast Cancer Surveillance Consortium Web Site. http://www.breastscreening.cancer.gov/elements.html#questionnaires. [Last accessed: February 2, 2005.]
(26) 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.
(27) Kerlikowske K, Carney PA, Geller B, Mandelson MT, Taplin SH, Malvin K, et al. Performance of screening mammography among women with and without a first-degree relative with breast cancer. Ann Intern Med 2000;133:855–63.
(28) Kerlikowske K, Miglioretti DL, Ballard-Barbash R, Weaver DL, Buist DS, Barlow WE, et al. Prognostic characteristics of breast cancer among postmenopausal hormone users in a screened population. J Clin Oncol 2003;21:4314–21.
(29) Miglioretti DL, Rutter CM, Geller BM, Cutter G, Barlow WE, Rosenberg R, et al. Effect of breast augmentation on the accuracy of mammography and cancer characteristics. JAMA 2004;291:442–50.
(30) National Cancer Institute Monograph: BCSC report: evaluating screening performance in practice. Available at: http://breastscreening.cancer.gov/espp.pdf. [Last accessed: February 2, 2005.]
(31) American Medical Association CoSA. http://www.ama-assn.org/ama/pub/category/12850.html. [Last accessed: February 2, 2005.]
(32) American College of Radiology. Breast imaging reporting and data system (BI-RADS). 3rd ed. Reston (VA): The College; 1998.
(33) Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time-varying covariates. Biostatistics 2004;5:381–98.
(34) Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
(35) Pepe MS, Urban N, Rutter C, Longton G. Design of a study to improve accuracy in reading mammograms. J Clin Epidemiol 1997;50:1327–38.
(36) Théberge I, Hébert-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;172:195–199.
(37) 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 2004:24;1840–50.
(38) Elmore JG, Yakano CY, Koepsell TD, et al. International variation in screening mammography Interpretations in community-based programs. J Natl Cancer Inst 2003;95:1384–93.

NOTES

[1]*Editor's note:* SEER is a set of geographically defined, population-based, central cancer registries in the United States, operated by local nonprofit organizations under contract to the National Cancer Institute (NCI). Registry data are submitted electronically without personal identifiers to the NCI on a biannual basis, and the NCI makes the data available to the public for scientific research.

This work was supported in part by the National Cancer Institute (CA86032 and Breast Cancer Surveillance Consortium cooperative agreements U01CA63740, U01CA86076, U01CA63736, U01CA70013 and U01CA69976) and The Department of Defense (DAMD179919112 and DAMD170010193).

Manuscript received August 6, 2004; revised November 4, 2004; accepted January 11, 2005.