UCSF
UC San Francisco Previously Published Works

Title

Mammography Facility Characteristics Associated With Interpretive Accuracy of Screening Mammography

Permalink

https://escholarship.org/uc/item/2kj5p1cg

Journal

Journal of the National Cancer Institute, 100(12)

ISSN

0027-8874

Authors

Taplin, Stephen
Abraham, Linn
Barlow, William E
et al.

Publication Date

2008-06-18

DOI

10.1093/jnci/djn172

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at https://creativecommons.org/licenses/by-nc/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Mammography Facility Characteristics Associated With Interpretive Accuracy of Screening Mammography

Stephen Taplin, Linn Abraham, William E. Barlow, Joshua J. Fenton, Eric A. Berns, Patricia A. Carney, Gary R. Cutter, Edward A. Sickles, Carl D'Orsi, Joann G. Elmore

Background Although interpretive performance varies substantially among radiologists, such variation has not been examined among mammography facilities. Understanding sources of facility variation could become a foundation for improving interpretive performance.

Methods In this cross-sectional study conducted between 1996 and 2002, we surveyed 53 facilities to evaluate associations between facility structure, interpretive process characteristics, and interpretive performance of screening mammography (ie, sensitivity, specificity, positive predictive value [PPV1], and the likelihood of cancer among women who were referred for biopsy [PPV2]). Measures of interpretive performance were ascertained prospectively from mammography interpretations and cancer data collected by the Breast Cancer Surveillance Consortium. Logistic regression and receiver operating characteristic (ROC) curve analyses estimated the association between facility characteristics and mammography interpretive performance or accuracy (area under the ROC curve [AUC]). All *P* values were two-sided.

Results Of the 53 eligible facilities, data on 44 could be analyzed. These 44 facilities accounted for 484 463 screening mammograms performed on 237 669 women, of whom 2686 were diagnosed with breast cancer during follow-up. Among the 44 facilities, mean sensitivity was 79.6% (95% confidence interval [CI] = 74.3% to 84.9%), mean specificity was 90.2% (95% CI = 88.3% to 92.0%), mean PPV1 was 4.1% (95% CI = 3.5% to 4.7%), and mean PPV2 was 38.8% (95% CI = 32.6% to 45.0%). The facilities varied statistically significantly in specificity ($P < .001$), PPV1 ($P < .001$), and PPV2 ($P = .002$) but not in sensitivity ($P = .99$). AUC was higher among facilities that offered screening mammograms alone vs those that offered screening and diagnostic mammograms (0.943 vs 0.911, $P = .006$), had a breast imaging specialist interpreting mammograms vs not (0.932 vs 0.905, $P = .004$), did not perform double reading vs independent double reading vs consensus double reading (0.925 vs 0.915 vs 0.887, $P = .034$), or conducted audit reviews two or more times per year vs annually vs at an unknown frequency (0.929 vs 0.904 vs 0.900, $P = .018$).

Conclusion Mammography interpretive performance varies statistically significantly by facility.

J Natl Cancer Inst 2008;100:876–887

Given that more than 27 million women have a screening mammogram each year, there is growing interest in evaluating and improving radiologists' interpretive performance of mammography (1). Two major groups of factors are known to influence the interpretation of mammograms: the characteristics of the women who are being screened (eg, breast density, age, and time-since-last mammogram) (2–4) and the characteristics of the interpreting radiologists (eg, number of years of experience in mammography interpretation and reading volume) (5–7). However, although patient and radiologist characteristics influence measures of interpretive performance, the characteristics that have been examined to date account for only 10% of the measured variation in perfor-

Affiliations of authors: Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD (ST); Group Health, Center for Health Studies, Seattle, WA (LA); Cancer Research and Biostatistics, Seattle, WA (WEB); Department of Family and Community Medicine, University of California Davis Health System, Sacramento, CA (JJF); Department of Radiology, Lynn Sage Comprehensive Breast Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, IL (EAB); Oregon Health Sciences University, Portland, OR (PAC); Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL (GRC); Department of Radiology, University of California School of Medicine, San Francisco, CA (EAS); Department of Radiology, Emory University, Atlanta, GA (CDO); Department of Internal Medicine, University of Washington School of Medicine, Seattle, WA (JGE).

Correspondence to: Stephen Taplin, MD, MPH, 6130 Executive Blvd, MSC 7344, EPN 4005, Bethesda, MD 20892-7344 (e-mail: taplins@mail.nih.gov).

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

DOI: 10.1093/jnci/djn172

© Oxford University Press 2008.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

mance (7). Furthermore, even if such characteristics accounted for a greater percentage of the variation, some cannot be modified (eg, patient age) and some cannot be easily ascertained by a woman who is having a mammogram (eg, radiologist reading volume).

Although interpretive performance varies substantially among radiologists (8), such variation has not, to our knowledge, been examined among mammography facilities. Knowledge of the sources of such variation among facilities is relevant to people who are concerned about the accuracy of mammography interpretation, including health plan leaders, health-care policy-makers, radiologists, women who are undergoing screening mammography, and the physicians who are ordering the tests. For example, women and their referring physicians could choose facilities that have specific characteristics or use practices that have been shown to be associated with better interpretive performance. In addition, such knowledge could inform the facilities themselves about practice changes they could make to improve interpretive performance. We therefore examined whether interpretive performance and accuracy vary across mammography facilities after accounting for known determinants of mammography interpretive performance.

CONTEXT AND CAVEATS

Prior knowledge

Mammography interpretive performance is known to be influenced by characteristics of the women who are being screened and of the interpreting radiologists. However, the extent to which screening mammography interpretive performance varies by facility-level characteristics is unclear.

Study design

A cross-sectional survey-based study that examined whether interpretive performance and accuracy vary across mammography facilities after accounting for known determinants of mammography interpretive performance and included 44 facilities, 484 463 screening mammograms performed on 237 669 women, and 2686 breast cancer diagnoses.

Contribution

Mammography interpretive accuracy was higher among facilities that offered screening mammograms alone vs those that offered screening and diagnostic mammograms; had a breast imaging specialist interpreting mammograms vs not; did not perform double reading vs independent vs consensus double reading; or conducted audit reviews two or more times per year vs annually or at an unknown frequency.

Implications

Understanding how facility characteristics influence interpretive accuracy could allow women and physicians to choose a mammography facility based on characteristics that are more likely to be associated with higher quality. Radiologists could also change the facilities' structures or processes to include practices that improve interpretive accuracy.

Limitations

Characterization of double reading of mammograms was limited. Unmeasured variation among women and radiologists may account for some of the variation associated with facilities. Some facilities were excluded from the analyses because of missing data. Associations were assessed at a single point in time. A number of selection biases may have affected the results.

Methods

Study Design, Setting, and Population

The data for this cross-sectional study were contributed by three geographically dispersed mammography registries of the Breast Cancer Surveillance Consortium (BCSC; http://breastscreening.cancer.gov/) (9): Group Health's Breast Cancer Surveillance System, a nonprofit health plan in the Pacific Northwest that includes more than 100 000 women aged 40 years or older (10); the New Hampshire Mammography Network (11), which provides mammography to more than 85% of the women in New Hampshire; and the Colorado Mammography Advocacy Program (9), which provides mammography to approximately half of the women who reside in the six-county metropolitan area of Denver, CO. As described in more detail below, these three registries collect data on women's characteristics, radiologists' mammography interpretations, and breast cancer diagnoses to evaluate mammography performance. Cancer ascertainment for the seven BCSC sites has been estimated to exceed 94.3% (12).

The study protocol was approved by the Human Subject Review Committees of the University of Washington School Of Medicine, Group Health, Dartmouth College, Northwestern University, and the Colorado Mammography Advocacy Program.

Mammography Registry Data

Each mammography registry provided data on every mammography examination that was performed at participating facilities from January 1, 1996, through December 31, 2001, including the date of the interpretation, the patient's age and breast density, and the date of her most recent previous mammogram. The mammography interpretation and breast density estimate were both collected according to the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) (13). The facility where the mammography was performed and the interpreting radiologist were also noted for each mammogram. A screening mammogram was one that the facility designated as having been performed for routine screening. To insure that our study focused on screening mammography, we excluded mammograms that were performed on women who had a history of breast cancer, had breast implants, or had undergone breast imaging within the previous 9 months. We ascertained breast cancer outcomes through December 31, 2002, by linkage with regional Surveillance, Epidemiology, and End Results registries; local tumor registries; and/or breast pathology databases that were maintained by the mammography registries.

Surveys

We used two mailed surveys in this study—a facility survey and a radiologist survey—that were developed by a multidisciplinary panel of experts in mammography, physics, economics, health services research, and epidemiology. The facility survey concerned the policies and practices of mammography facilities and was based on a conceptualization of factors that are known or suspected to influence mammography interpretation, including characteristics of the patient, of the radiologist, and of the facility (14–16). Drafts of this survey were pilot tested extensively for accuracy and reproducibility

of responses at three facilities that did not participate in the three study registries but were located in the same geographic regions as those that did. Cognitive interviewing of respondents to the pilot surveys led to iterative improvements in the survey and resulted in a final survey that required approximately 15 minutes to complete.

We mailed the facility survey to the designated contact person at each facility in December 2001; data collection was completed by September 2002. If we received no response after a second mailing, we sought a response by telephone or through site visits. The facility surveys were completed by one or more of the following employees at each facility: a technologist, a radiologist, the radiology department manager, and/or the facility business manager.

The radiologist survey was mailed to all radiologists (n = 168) who interpreted mammograms at the facilities that had responded to the first survey (n = 43); this survey was completed and returned by 128 radiologists (76%) who were still practicing in 2002 (*see* Figure 1). Detailed results of the radiologist survey are reported elsewhere (14). The radiologist and facility surveys are available at http://breastscreening.cancer.gov/collaborations/favor.html.

Specification of Study Variables

Outcome Variables. The primary outcomes were the sensitivity, specificity, positive predictive value (PPV; calculated as described below), and the area under the receiver operating characteristic (ROC) curve (AUC) of a screening mammogram. Mammography interpretations were recorded using the numerical value for each BI-RADS assessment according to the version in use at the time (13). To calculate the performance measures, we defined a negative interpretation as BI-RADS 1 (negative), BI-RADS 2 (benign finding), or BI-RADS 3 (probably benign) when the latter category of interpretation was associated with a recommendation for short-interval follow-up (eg, 6 months) but not when it was associated with a recommendation for immediate work-up. Rarely, a BI-RADS category 3 interpretation was associated with a normal follow-up interval (1 or 2 years), and we counted those as negative interpretations as well, although it was an incorrect use of the BI-RADS lexicon. A mammogram was classified as positive if it was given a BI-RADS category of 0 (additional imaging required), 4 (needs evaluation), or 5 (highly suggestive of malignancy). For the main analysis, we categorized mammograms that had a BI-RADS category of 3 and a recommendation for immediate follow-up as a BI-RADS 0 interpretation (13). For each screening mammogram, we determined whether breast cancer was diagnosed within 1 year of the examination or anytime before the next screen (follow-up period). Women who had a screening examination performed between 9 months and 1 year after their previous screen were considered to be cancer free, and their follow-up periods were truncated on the date of the subsequent screening mammogram. Sensitivity was defined as the proportion of all mammograms with a positive interpretation at the time of screening and a breast cancer during the follow-up period. Specificity was defined as the proportion of all mammograms with a negative interpretation and no breast cancer diagnosed during the follow-up period. For positive predictive value, we used two definitions that were consistent with those recommended by the American College of Radiology for medical audits (13). Positive predictive value 1 (PPV1) was defined as the proportion of screens that were associated with a breast cancer diagnosis within the follow-up period among those with a positive interpretation as noted above for the main analysis. Positive predictive value 2 (PPV2) was defined as the number of mammograms that were associated with a breast cancer diagnosis during follow-up among those with a recommendation for biopsy, surgical consultation, or fine-needle aspiration (BI-RADS 4,5).

We used ROC analysis to determine whether facility characteristics were associated with the overall accuracy of screening interpretations, as measured by the AUC. We chose the AUC because it compares sensitivity against 1−specificity over a range of values for these two measures and therefore incorporates both measures of interpretive performance in a single analysis. AUC values can range from 0 to 1, but only values greater than 0.5 reflect interpretive accuracy that is greater than chance alone. A higher AUC reflects an increased ability to discriminate between mammograms that do and do not show the presence of a cancer (ie, better accuracy). Nonetheless, because the AUC is computed from a curve that varies over the entire range of possible values for sensitivity and specificity, we also considered whether these values were in an acceptable range for screening mammography by dichotomizing the ordinal BI-RADS scale and reporting sensitivity and specificity as described above.

Independent Variables. Because interpretive performance is influenced by characteristics of both the patient and the radiologist, we built logistic regression models of facility performance measures that included covariates for patients and radiologists and therefore accounted for these factors. For each patient, we included characteristics that were associated with variation in interpretive performance of mammography in other studies, that is, age at the time of the mammogram (in 5-year intervals) (4,17), BI-RADS breast density category (3), and the number of months since the previous mammogram (4). For each radiologist, we included the self-reported number of years spent interpreting mammograms and the facility-reported number of screening mammograms recorded in the registry in the year before the survey (2001) because these variables have been found to be associated with variation in performance measures (5,8).

To be included in this analysis, each facility had to have conducted at least 300 screening mammography examinations with no missing data on patient characteristics during the study period. We evaluated two groups of facility characteristics that could affect all radiologists: 1) structural and organizational characteristics, which included whether the facility had a financial incentive (profit vs not-for-profit) or an affiliation with an academic medical center (yes/no), the type of breast imaging offered (screening only vs screening and diagnostic), and the annual facility volume of mammograms (<1500, 1501–2500, 2501–6000, or >6000, computed from registry data) and 2) characteristics of the interpretive process, which included the presence of a breast imaging specialist (ie, a radiologist who spent at least 50% of his or her time doing breast imaging), the percentage of mammograms read off-site at other facilities (0%, 1%–79%, or 80%–100%), the percentage of screens that were "batch read" in groups of 10 or more (0%–50% vs 51%–100%), whether mammograms were read by two or more radiologists (ie, double reading; yes/no), whether double reading was done with or without knowledge of the other reader's results (consensus

vs independent), the number of times per year that feedback (audit data) was provided to radiologists about their interpretive performance (once vs two or more times per year), and the method of reviewing audit results (together with other radiologists, with a manager only, by the radiologist alone, or unknown). Reviewing together with other radiologists could occur in a variety of ways that were not specified. The usual approach to these group reviews was that definitions of the measures were given; the individual performance measures and group means were also reported. A review of the images of specific troublesome cases might also have occurred. The categories for volume, batch reading, and proportion of films read off-site were chosen based on the distributions of the data and in an attempt to have the facilities distributed as evenly as possible across the values. The data for the proportion of mammograms read at other facilities were severely skewed to either end of the distribution (ie, 0% and 80%–100%). Twelve facilities were missing a response to two survey questions related to audit data (two were missing data on the frequency of audits, five were missing data on the method of audit review, and five were missing both kinds of data). To avoid having to exclude these facilities from the analysis, we created a separate category (unknown) for these survey questions and included that variable value in the analysis.

We hypothesized that increased facility volume, presence of consensus double reading of mammograms, and reviewing audit data together in a meeting would be associated with increased accuracy but suspected that trade-offs in sensitivity and specificity would occur with other facility characteristics and that the other facility characteristics might not lead to a net change in accuracy. We expected that a high facility volume would be associated with improved accuracy through increased experience and an emphasis on screening mammography. We expected consensus double reading and reviewing audit results as a group to be associated with improved accuracy because social learning theory suggests that people learn from each other's behavior and from the behavior of role models (18).

Statistical Analysis

Analysis of Sensitivity, Specificity, and Positive Predictive Value. For each performance measure, we initially fit a multivariable model that included patient characteristics, radiologist characteristics, and an individual effect for facility (statistically called a random effect), but we did not include other facility factors as covariates. The purpose of these models was to determine whether statistically significant variation in each performance measure was associated with an individual facility after adjusting for patient and radiologist characteristics. The variance in each performance measure due to the facility was tested to determine whether this variation across facilities was greater than that expected by chance alone. For each facility, we computed the mean values and 95% confidence intervals (CIs) for sensitivity, specificity, PPV1, and PPV2. We then built multivariable models for each performance measure using facility characteristics that were relevant to our hypotheses (ie, facility volume, method of double reading, method of audit review) or that were associated with any outcome in univariate analyses at a P value of .1 or lower. We removed variables that were highly related to other variables in the model (presence of double reading, offering interventional procedures, and percentage of screening mammograms) or that were no longer statistically significant at a P level of .1 or lower (eg, affiliation with academic medical center). The final model for each performance measure therefore included patient characteristics, radiologist characteristics, and the facility characteristics that were part of our principal hypotheses or that remained associated with at least one performance measure at a P level of .1 or lower. Screens that were missing any of these factors were excluded; the final models included 360 149 screening mammograms (*see* Figure 1). All screening examinations that met the above criteria were included in the models; thus, multiple mammograms per woman could be included until a cancer was diagnosed.

The models for sensitivity examined the probability of a positive screen among screens with a cancer diagnosis during follow-up. The models for specificity estimated the probability of having a negative screen among those without a cancer diagnosis during follow-up. The models for PPV estimated the probability of having a cancer diagnosis given a positive screen, with the definition of positive screen differing for PPV1 and PPV2 as noted above. These mixed-effect logistic regression models were fit using the SAS procedure NLMIXED (19).

Receiver Operating Characteristic Analysis. We also used the SAS procedure NLMIXED (19) to perform ROC analysis, which was used to summarize the association between facility characteristics and the overall interpretive accuracy of screening mammography. We used the BI-RADS assessment codes as an ordinal response: 1, 2, 3 with no immediate work-up, 3 with immediate work-up, 0, 4, and 5. We then fit ordinal regression models with two random effects and covariates for patient and radiologist characteristics. The two random effects adjusted for the radiologist's likelihood of calling the mammogram positive and his or her ability to discriminate between mammograms with and without cancer present. The multivariable models enabled us to estimate ROC curves and the AUC associated with specific facility characteristics while adjusting for patient and radiologist characteristics. The actual AUC value is computed from the estimates associated with covariates. Confidence intervals are available for those estimates but are not routinely computed for AUC values, which are only summary statistics and primarily descriptive. Additional detail regarding the method of fitting the ROC curves is available in an earlier publication (5). Likelihood ratio statistics were used to determine whether each facility factor was statistically significantly associated with accuracy ($P \leq .05$). All P values are two-sided.

Results

Of the 53 facilities that were eligible for inclusion in this study, 45 (85%) completed a facility survey. One facility had conducted fewer than 300 screens and was excluded (Figure 1). The remaining 44 facilities provided data for the main analysis. Among the 44 facilities, 24 (54.6%) were structured for profit, 11 (25%) did not offer diagnostic mammograms, 17 (38.6%) had a radiologist on staff who specialized in breast care, and 19 (43.2%) provided radiologists with audit data two or more times per year (Table 1). These 44 facilities accounted for 484 463 screening mammograms performed on 237 669 women, of whom 2686 were diagnosed with breast cancer during follow-up (Figure 1). The mean number of cancers diagnosed in patients who were seen at the 44 facilities was 61 (median = 28, range

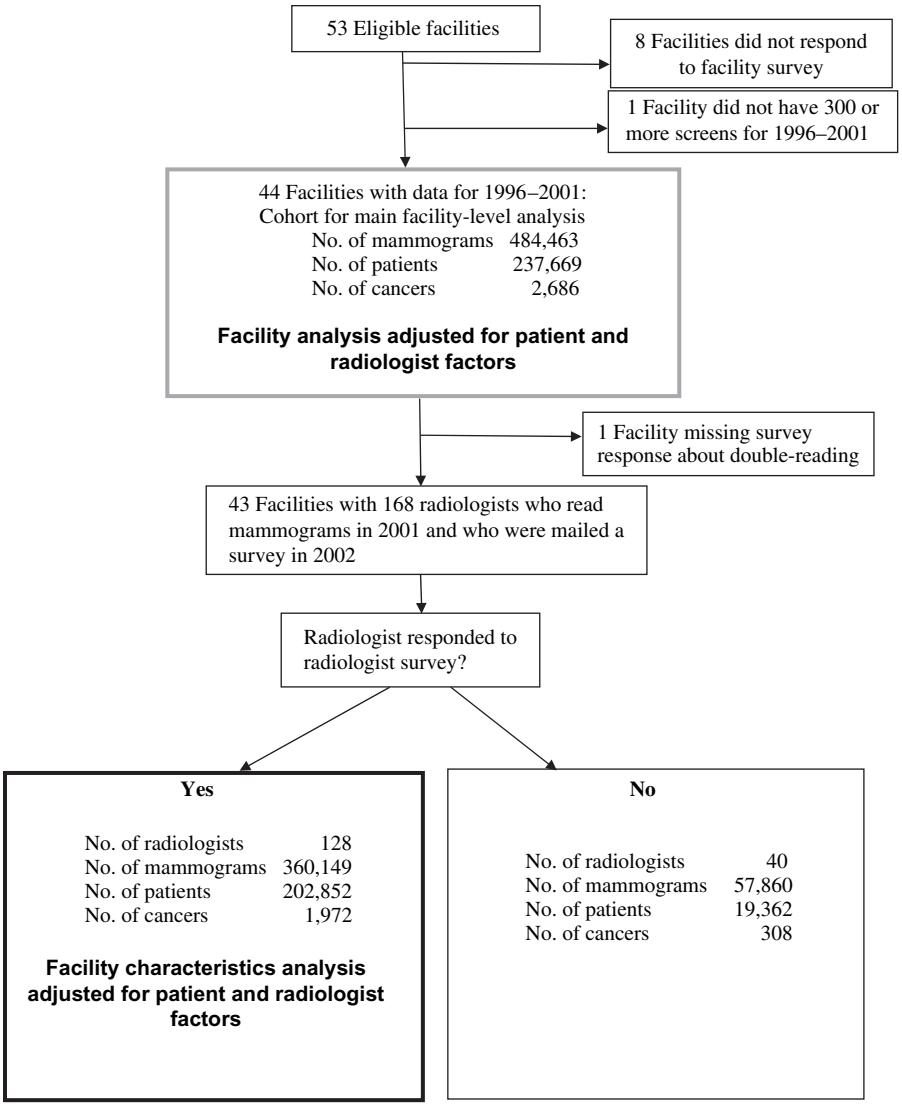

Figure 1. Data collection for the analysis of mammography performance.

= 0–342). The mean number of radiologists who interpreted mammograms at the 44 facilities was 20 (median = 18, range = 3–54).

Figure 2 shows the adjusted performance measures from the first multivariable models, which included a facility random effect and patient and radiologist factors, for each of the 44 facilities during this time period. Among the 44 facilities, mean sensitivity was 79.6% (95% CI = 74.3% to 84.9%; median = 81.9%, range = 0%–100%), mean specificity was 90.2% (95% CI = 88.3% to 92.0%; median = 90.8%, range = 55.6%–96.6%), mean PPV1 was 4.1% (95% CI = 3.5% to 4.7%; median = 4.0%, range = 0%–10.7%), and mean PPV2 was 38.8% (95% CI = 32.6% to 45.0%; median = 36.5%, range = 7.2%–100%). The 44 facilities varied statistically significantly in specificity ($P < .001$), PPV1 ($P < .001$), and PPV2 ($P = .002$) but not in sensitivity ($P = .99$).

The multivariable analysis was intended to more closely examine the factors associated with variation in facility performance, but not all the facilities provided the necessary data, and such facilities were omitted from this analysis. One facility was missing data on the double reading of mammograms and was excluded from the subanalysis that adjusted for radiologist characteristics, leaving 43 facilities. Of the 168 radiologists who worked at these 43 facilities, 128 (76%) responded to the radiologist survey. The models that accounted for radiologist characteristics were therefore based on data from 43 facilities, 128 radiologists, and 360 149 screening mammograms in 202 852 women, of whom 1972 were diagnosed with breast cancer during the follow-up period (Figure 1). Only one of the three performance measures—specificity—differed statistically significantly between radiologists who worked in facilities that responded to the facility survey and radiologists who worked in facilities that did not respond to the facility survey (mean sensitivity: 81.2% vs 84.9%, difference = 3.7%, 95% CI = −0.92% to 8.33%, $P = .14$; mean specificity: 89.4% vs 90.2%, difference = 0.8%, 95% CI = 0.52% to 1.03%, $P < .001$; mean PPV1: 3.97% vs 3.73%, difference = 0.24%, 95% CI = −0.28% to 0.75%, $P = .38$).

We examined the variation in performance measures among the facilities by looking more closely at the characteristics of the practice procedures based on data from the facility surveys (Table 1). Outcomes (performance measures) associated with a facility characteristic at a statistical significance level of less than .1 are shown in boldface in the table. Having a breast imaging specialist

Table 1. Facility-reported practice procedures and performance measures (N = 44 facilities with 484 463 screening mammography examinations)*

Facility characteristic	No. (%)	Specificity†, % (95% CI)	Sensitivity‡, % (95% CI)	PPV1§, % (95% CI)	PPV2‖, % (95% CI)
Facility structure and organization					
Is your mammography facility for-profit or not-for-profit?					
Not-for-profit	20 (45.5)	91.0 (89.6 to 92.4)	82.2 (77.8 to 86.7)	**4.9 (4.0 to 5.8)**	42.5 (34.5 to 50.6)
For-profit	24 (54.6)	89.5 (86.1 to 92.8)	77.3 (67.9 to 86.7)	**3.5 (2.7 to 4.3)**	35.5 (25.7 to 45.2)
Is this facility associated with an academic medical center?					
No	34 (77.3)	90.2 (87.9 to 92.6)	78.8 (72.3 to 85.2)	4.3 (3.6 to 5.0)	**41.4 (33.9 to 48.8)**
Yes	10 (22.7)	89.9 (86.8 to 93.0)	82.7 (74.9 to 90.6)	3.6 (2.1 to 5.2)	**28.6 (22.2 to 35.0)**
Facility volume (average no. of mammograms per year)¶					
≤1500	9 (20.5)	93.4 (91.3 to 95.6)	**64.6 (38.2 to 91.0)**	4.1 (1.5 to 6.7)	**32.7 (19.6 to 45.7)**
1501–2500	12 (27.3)	87.1 (80.4 to 93.8)	**87.5 (79.9 to 95.0)**	3.6 (2.3 to 4.9)	**24.9 (16.3 to 33.6)**
2501–6000	12 (27.3)	89.7 (88.2 to 91.3)	**83.1 (78.8 to 87.4)**	4.1 (3.4 to 4.9)	**47.6 (36.6 to 58.6)**
>6000	11 (25.0)	91.2 (89.4 to 93.1)	**78.1 (73.1 to 83.0)**	4.7 (4.0 to 5.5)	**46.4 (31.3 to 61.6)**
What percentage of your mammograms are screening mammograms?					
1–74	15 (34.1)	89.9 (88.1 to 91.7)	82.2 (77.4 to 87.0)	3.9 (3.0 to 4.7)	**47.9 (34.6 to 61.2)**
75–79	13 (29.5)	88.0 (81.9 to 94.1)	84.2 (77.7 to 90.8)	4.3 (3.3 to 5.4)	**30.9 (21.6 to 40.1)**
80–100	16 (36.4)	92.2 (90.5 to 93.8)	73.6 (60.5 to 86.6)	4.2 (2.8 to 5.6)	**36.9 (27.3 to 46.5)**
What percentage of the screening mammograms done at your facility are interpreted at another facility?#					
0	27 (61.4)	90.5 (89.2 to 91.8)	82.4 (78.7 to 86.1)	**4.6 (3.8 to 5.3)**	40.5 (34.1 to 46.9)
80–100	17 (38.6)	89.6 (84.9 to 94.4)	74.9 (61.5 to 88.3)	**3.4 (2.4 to 4.5)**	35.6 (21.1 to 50.1)
What percentage of screening mammograms are interpreted in groups of 10 or more?					
0–50	7 (17.5)	88.1 (74.6 to 101.7)	77.6 (57.8 to 97.4)	4.6 (1.5 to 7.7)	35.8 (19.4 to 52.3)
51–100	33 (82.5)	90.4 (89.3 to 91.4)	80.0 (73.9 to 86.1)	3.9 (3.3 to 4.5)	40.7 (33.1 to 48.4)
Missing	4				
Does this facility offer diagnostic mammograms?					
No	11 (25.0)	92.0 (89.8 to 94.1)	**68.6 (47.9 to 89.4)**	**3.2 (1.8 to 4.5)**	47.6 (25.2 to 70.1)
Yes	33 (75.0)	89.5 (87.1 to 92.0)	**82.9 (79.4 to 86.4)**	**4.5 (3.8 to 5.1)**	36.6 (30.5 to 42.7)
Does this facility offer interventional services (FNA, core or vacuum-assisted biopsy, cyst aspirations, needle localization or other procedures)?					
No	17 (39.5)	89.2 (84.5 to 93.9)	76.8 (65.0 to 88.7)	**3.5 (2.5 to 4.6)**	39.7 (25.4 to 54.0)
Yes	26 (60.5)	90.6 (89.3 to 91.9)	83.1 (79.2 to 86.9)	**4.6 (3.9 to 5.4)**	38.3 (32.1 to 44.5)
Missing	1				
Interpretive and audit processes					
Are any screening mammograms performed at your facility interpreted by radiologist(s) who specialize in breast care?					
No	27 (61.4)	**88.9 (86.0 to 91.9)**	**84.2 (80.5 to 87.8)**	**4.6 (3.8 to 5.3)**	**34.5 (27.4 to 41.6)**
Yes	17 (38.6)	**92.1 (90.9 to 93.3)**	**71.9 (59.1 to 84.7)**	**3.5 (2.5 to 4.5)**	**47.7 (35.6 to 59.8)**
Are any screening mammograms from your facility interpreted by more than one radiologist?					
No	21 (48.8)	89.6 (85.8 to 93.3)	**84.2 (77.5 to 90.8)**	4.3 (3.2 to 5.3)	**33.4 (25.4 to 41.5)**
Yes	22 (51.2)	90.8 (89.3 to 92.4)	**74.9 (66.4 to 83.5)**	4.0 (3.2 to 4.8)	**45.3 (36.0 to 54.6)**
Missing	1				
How are decisions made for mammograms interpreted by more than one radiologist?					
Double readings not performed	21 (48.8)	89.6 (85.8 to 93.3)	84.2 (77.5 to 90.8)	4.3 (3.2 to 5.3)	33.4 (25.4 to 41.5)
Independently	19 (44.2)	90.9 (89.2 to 92.6)	74.0 (64.0 to 84.0)	3.8 (2.9 to 4.7)	46.3 (35.4 to 57.2)
By consensus	3 (7.0)	90.4 (80 to 100.8)	80.6 (57.1 to 104.2)	5.2 (2.9 to 7.5)	39.8 (14.9 to 64.6)
Missing	1				

(Table continues)

Table 1 (continued).

Facility characteristic	No. (%)	Specificity†, % (95% CI)	Sensitivity‡, % (95% CI)	PPV1§, % (95% CI)	PPV2‖, % (95% CI)
How often is individual radiologist-level performance data shared with radiologists?					
Once a year	18 (40.9)	**89.9 (88.3 to 91.5)**	80.5 (69.3 to 91.7)	3.7 (2.9 to 4.5)	33.4 (25.3 to 41.5)
Two or more times per year	19 (43.2)	**91.9 (90.5 to 93.4)**	78.6 (75.0 to 82.2)	4.7 (3.6 to 5.8)	39.2 (29.1 to 49.3)
Unknown	7 (15.9)	**85.9 (73.3 to 98.6)**	79.9 (60.0 to 99.9)	3.8 (1.9 to 5.6)	52.1 (27.4 to 76.9)
How is this information reviewed?					
Reviewed together (with other radiologists) in a meeting	21 (47.7)	89.9 (88.5 to 91.2)	81.8 (77.3 to 86.3)	**4.4 (3.8 to 5.0)**	36.3 (26.3 to 46.4)
Reviewed by facility or department manager or by lead radiologist alone	8 (18.2)	92.3 (90.9 to 93.8)	77.2 (71.0 to 83.4)	**3.8 (2.2 to 5.4)**	41.9 (33.9 to 50.0)
Reviewed by each radiologist alone	5 (11.4)	93.6 (89.9 to 97.3)	84.6 (71.4 to 97.8)	**5.9 (1.5 to 10.4)**	40.4 (26.2 to 54.6)
Unknown	10 (22.7)	87.3 (79.0 to 95.7)	74.1 (51.4 to 96.7)	**3.0 (1.7 to 4.3)**	41.8 (21.9 to 61.6)

* Outcomes in bold have an overall *P* level of .1 or lower for a univariate association. CI = confidence interval; PPV1 = positive predictive value 1; PPV2 = positive predictive value 2; FNA = fine-needle aspiration; BI-RADS = American College of Radiology Breast Imaging Reporting and Data System.

† Specificity was defined as the percentage of screening examinations that were given a negative BI-RADS assessment (BI-RADS 1, 2, or 3 with a recommendation of normal or short-interval follow-up) among those that did not have a breast cancer diagnosis during the follow-up period.

‡ Sensitivity was defined as the percentage of screening mammograms that were given a positive BI-RADS assessment (BI-RADS 0, 4, 5, or 3 with a recommendation for immediate work-up) among those that had a diagnosis of invasive breast cancer or ductal carcinoma in situ during the follow-up period.

§ PPV1 was defined as the percentage of screens that were associated with a breast cancer diagnosis within 1 year of follow-up among those with a positive BI-RADS assessment.

‖ PPV2 was defined as the percentage of screens that were associated with breast cancer diagnosis during follow-up among those with a BI-RADS assessment of 4 or 5 and a recommendation for biopsy, surgical consultation, or FNA.

¶ Based on registry data.

No facilities reported that 1%–79% of the screening mammograms done at the facility were interpreted at another facility.

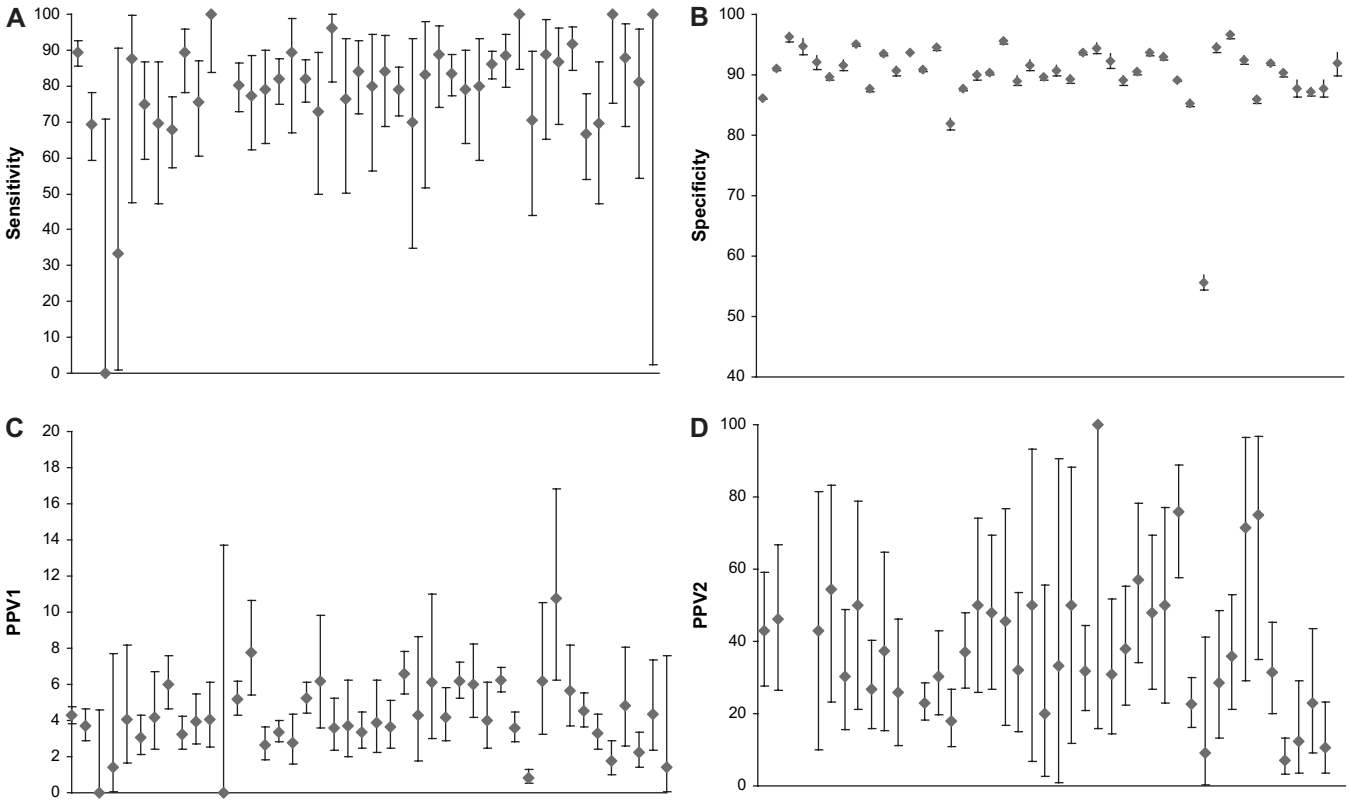

Figure 2. Screening mammography performance measures for the 44 facilities. **A**) Sensitivity. **B**) Specificity. **C**) Positive predictive value of any additional evaluation (PPV1). **D**) PPV of referral for biopsy (PPV2). **Diamonds** indicate mean values; **error bars** correspond to 95% confidence intervals. The 44 facilities varied statistically significantly in specificity ($P < .001$), PPV1 ($P < .001$), and PPV2 ($P = .002$) but not in sensitivity ($P = .99$).

who interpreted screening mammograms was the only facility characteristic that was associated with all four performance measures in the univariate analyses at a *P* level less than .1. Facilities that provided performance data to their radiologists two or more times per year had a higher specificity than those that provided performance data only once a year (mean specificity: 91.9% vs 89.9%, difference = 2%, 95% CI = 0.01% to 4.05%, *P* = .056). Facility affiliation with an academic medical institution was not associated with sensitivity, specificity, or PPV1 at a *P* level of .1 or lower; thus, it was not included in the multivariable models.

Several facility-level characteristics were associated with differences in measures of interpretive performance in multivariable models that controlled for patient and radiologist characteristics (Table 2). Specificity decreased with increasing facility volume

Table 2. Mixed-effects modeling of sensitivity, specificity, and positive predictive value 1 by facility characteristics with adjustment for patient and radiologist characteristics and mammography registry (N = 43 facilities with 360149 screening mammography examinations)*

	Odds of having a negative mammogram given no cancer diagnosis (specificity)		Odds of having a positive mammogram given a cancer diagnosis (sensitivity)		Odds of having a cancer diagnosis given a positive mammogram (PPV1)	
Facility characteristic	**OR (95% CI)**	**Overall *P***	**OR (95% CI)**	**Overall *P***	**OR (95% CI)**	**Overall *P***
Facility structure and organization						
Facility volume (average no. of mammograms per year)†						
≤1500	1.00 (referent)	.002	1.00 (referent)	.097	1.00 (referent)	.202
1501–2500	0.65 (0.48 to 0.88)		2.77 (1.15 to 6.65)		1.04 (0.70 to 1.55)	
2501–6000	0.66 (0.49 to 0.89)		2.29 (1.05 to 5.03)		0.84 (0.58 to 1.21)	
>6000	0.53 (0.38 to 0.74)		2.53 (1.17 to 5.44)		0.99 (0.68 to 1.43)	
Is your mammography facility for-profit or not-for-profit?						
Not-for-profit	1.00 (referent)	.315	1.00 (referent)	.324	1.00 (referent)	.057
For-profit	0.88 (0.67 to 1.14)		1.26 (0.79 to 2.01)		0.82 (0.66 to 1.01)	
Does this facility offer diagnostic mammograms?						
No	1.00 (referent)	.003	1.00 (referent)	.883	1.00 (referent)	<.001
Yes	0.66 (0.50 to 0.86)		0.95 (0.50 to 1.83)		0.63 (0.49 to 0.82)	
Interpretive and audit process						
Are any screening mammograms performed at your facility interpreted by a radiologist who specializes in breast care?						
No	1.00 (referent)	.083	1.00 (referent)	.652	1.00 (referent)	.039
Yes	1.26 (0.97 to 1.63)		0.90 (0.57 to 1.42)		1.23 (1.01 to 1.50)	
What percentage of the screening mammograms done at your facility are interpreted at another facility?‡						
0	1.00 (referent)	.002	1.00 (referent)	.093	1.00 (referent)	.072
80–100	0.70 (0.56 to 0.87)		1.59 (0.92 to 2.72)		0.82 (0.66 to 1.02)	
How are decisions made for mammograms interpreted by more than one radiologist?						
Double reads not performed	1.00 (referent)	.177	1.00 (referent)	.779	1.00 (referent)	.005
Independent double reads	0.84 (0.68 to 1.03)		0.90 (0.63 to 1.30)		0.90 (0.78 to 1.03)	
Double reads by consensus	0.76 (0.49 to 1.16)		1.03 (0.53 to 2.01)		0.62 (0.48 to 0.82)	
How often is individual radiologist-level audit data given back to radiologists on their performance?						
Once a year	1.00 (referent)	<.001	1.00 (referent)	.291	1.00 (referent)	.050
Two or more times per year	1.54 (1.23 to 1.94)		0.74 (0.50 to 1.09)		1.23 (1.03 to 1.47)	
Unknown	0.81 (0.61 to 1.06)		1.08 (0.60 to 1.96)		1.13 (0.91 to 1.40)	
How is this audit information reviewed?						
Reviewed together (with other radiologists) in meeting	1.00 (referent)	.001	1.00 (referent)	.609	1.00 (referent)	<.001
Reviewed by facility or department manager or by lead radiologist alone	0.60 (0.44 to 0.84)		1.22 (0.61 to 2.44)		0.94 (0.69 to 1.27)	
Reviewed by each radiologist alone	1.32 (0.94 to 1.83)		1.20 (0.49 to 2.93)		1.07 (0.73 to 1.58)	
Unknown	0.70 (0.53 to 0.92)		1.38 (0.83 to 2.30)		0.65 (0.54 to 0.79)	

* OR = odds ratio; CI = confidence interval; PPV1 = positive predictive value 1.

† Based on registry data.

‡ No facilities reported that 1%–79% of the screening mammograms done at the facility were interpreted at another facility.

Table 3. Facility characteristics and area under the receiver operating characteristic curve

Characteristic	AUC*	$P_{accuracy}$†
Facility structure and organization		
Facility volume (average no. of mammograms per year)‡		
≤1500	0.916	.117
1501–2500	0.937	
2501–6000	0.912	
>6000	0.911	
Is your mammography facility for-profit or not-for-profit?		
Not-for-profit	0.913	.534
For-profit	0.919	
Does this facility offer diagnostic mammograms?		
No	0.943	.006
Yes	0.911	
Interpretive and audit processes		
Are any screening mammograms performed at your facility interpreted by radiologists who specialize in breast care?		
No	0.905	.004
Yes	0.932	
What percentage of the screening mammograms done at your facility are interpreted off-site at another facility?§		
0	0.917	.139
80–100	0.900	
How are decisions made for mammograms interpreted by more than one radiologist?		
Double reads not performed	0.925	.034
Independent double reads	0.915	
Double reads by consensus	0.887	
How often is individual radiologist-level audit data given back to radiologists on their performance?		
Once a year	0.904	.018
Twice or more per year	0.929	
Unknown	0.900	
How is this audit information reviewed?		
Reviewed together with other radiologists in meeting	0.918	.158
Reviewed by facility or department manager or by lead radiologist alone	0.915	
Reviewed by each radiologist alone	0.937	
Unknown	0.899	

* AUC = area under the receiver operating characteristic curve.

† Based on a likelihood ratio statistic.

‡ Based on registry data.

§ No facilities reported that 1%–79% of the screening mammograms done at the facility were interpreted at another facility.

(*P* = .002), but sensitivity did not vary with facility volume (*P* = .097). Facilities that offered diagnostic mammograms had lower specificity (odds ratio [OR] = 0.66, 95% CI = 0.5 to 0.86; *P* = .003) and lower PPV1 (OR = 0.63, 95% CI = 0.49 to 0.82; *P* < .001) than facilities that did not. Having a breast imaging specialist who interpreted screening mammograms was associated with slightly higher specificity (OR = 1.26, 95% CI = 0.97 to 1.63; *P* = .83) and higher PPV1 (OR = 1.23, 95% CI = 1.01 to 1.50; *P* = .39), although neither association was statistically significant. Facilities that sent mammograms off-site for interpretation had statistically significantly lower specificity than those that did not (OR = 0.70, 95% CI = 0.56 to 0.87; *P* = .002). The profit status of the facility was not associated with any measure of interpretive performance.

Compared with no double reading of mammograms, the method of double reading was not associated with differences in specificity (*P* = .177) or sensitivity (*P* = .779) but was associated with a lower PPV1 (*P* = .005) (Table 2). Among the 23 facilities that reported double reading of mammograms, only four did so for all readings; 13 facilities double read 2%–30% of mammograms, and six did not report the percentage that were double read. Providing audit data to radiologists two or more times per year was associated with higher specificity and PPV1 than providing data only once per year. Reviewing audit information alone vs together with other radiologists was associated with lower specificity. None of the facility structure or interpretive process variables was associated with differences in PPV2 (data not shown).

The ROC analysis and comparison of AUCs showed that four facility characteristics were associated with higher overall accuracy of mammography after controlling for patient and radiologist characteristics (Table 3): AUC was higher among facilities offering screening mammograms alone than among those that offered both screening and diagnostic mammograms (0.943 vs 0.911, *P* = .006), among facilities having a breast imaging specialist interpreting mammograms vs not having one (0.932 vs 0.905, *P* = .004), among facilities not performing double reading vs independent vs consensus double reading (0.925 vs 0.915 vs 0.887, *P* = .034), and among facilities conducting audit reviews two or more times per year compared with annual or unknown frequency (0.929 [2/year] vs 0.904 [1/year] vs 0.900 [unknown], *P* = .018). Contrary to our hypotheses, neither higher mammogram volume in the facility (*P* = .117) nor reviewing audits together (*P* = .158) was associated with improved accuracy. Moreover, contrary to our expectation that consensus double reading would be associated with improved accuracy, we found that any method of double reading was associated with decreased accuracy (*P* = .034).

Discussion

To our knowledge, this is the first study to demonstrate that after controlling for radiologists' and patients' characteristics, screening mammography interpretive performance (specificity, PPV, AUC) varies by facility and is associated with facility-level characteristics. Variation in interpretive performance across radiologists is well recognized (8,17), but our findings suggest that interpretive performance also varies according to characteristics of the facilities where the radiologists work. The facility characteristics can be summarized in a way that distinguishes the facility structures and interpretive processes from each other and identifies facilities that are more likely to have a higher interpretive performance. These findings are important because a referring physician or the patient herself is much more likely to have the opportunity to choose the facility where the mammogram is performed and interpreted than they are to choose the radiologist who will interpret the mammogram.

Among the four interpretive performance measures, specificity and PPV1 were the most consistently associated with facility-level differences. We did not find facility-level differences in sensitivity. Whether higher sensitivity or higher specificity best identifies

higher quality is somewhat controversial because choosing one or the other as an indicator depends on a value judgment about whether finding more cancers or avoiding false-positive examinations is more important. To avoid this controversy, we used a single measure of accuracy that accounts for both sensitivity and specificity, the AUC. Using this summary measure, we showed that offering screening examinations only, not doing double readings, including a breast imaging specialist on staff, and conducting audits two or more times per year were associated with increased accuracy. If these findings are replicated in other studies, they could provide the basis for developing policy that examines and reports mammography performance at a facility level.

Although we used the AUC as our summary measure, its calculation depends on a distribution of interpretations across an ordinal scale with at least three values. Because radiologists' clinical recommendations for subsequent actions could be construed as dichotomous (eg, normal interval follow-up or further evalution), there could be concern that radiologists do not use the full ordinal scale needed to calculate the AUC. To confirm that the ordinal scale of BI-RADS interpretations met this assumption, in a previous study (5), we examined cancer rates across the full ordinal scale. In that study, we showed that our ordering of the BI-RADS assessment codes was associated with cancer rates that increased with each increase in the code, even though some of these BI-RADS codes (eg, BI-RADS 1 and 2) would result in the same management recommendation (ie, normal interval follow-up). The cancer rates per 1000 mammograms were as follows: 0.83 for BI-RADS 1, 1.43 for BI-RADS 2, 7.48 for BI-RADS 3 with no immediate work-up, 14.54 for BI-RADS 3 with immediate work-up, 32.25 for BI-RADS 0, 165.68 for BI-RADS 4, and 839.66 for BI-RADS 5 (5). Using the entire ordinal scale for ROC analysis is therefore justified and prevents loss of information compared with other methods of classifying BI-RADS interpretations. We therefore include ROC analysis in our evaluation and suggest that it offers an important estimate of performance.

Contrary to our hypotheses, neither annual facility volume nor the method of audit review was associated with greater interpretive accuracy. This lack of association does not, however, mean that these factors are unimportant. In this study, we conducted a nested analysis that controlled for some characteristics of women and radiologists. In this nested analysis, we examined facility volume after accounting for the radiologists' volume. The lack of association between facility volume and interpretive performance in our analysis means that a low-volume reader performs the same regardless of the volume of the facility. There was no additional effect of facility volume beyond that accounted for at the radiologist level. However, we still need to understand the effect of volume on radiologists' interpretations. A report from the Institute of Medicine (20) noted that persistent questions remain about the influence of volume on mammography interpretive performance. Our analysis does not change that.

Another surprise in our findings is that double reading was associated with lower accuracy. This finding is inconsistent with published results of randomized trials of independent and consensus double reading (21,22). However, the literature on double reading demonstrates a variety of approaches within each of the two broad categories of consensus and independent double reading, and we did not capture those differences in our survey. The negative association we observed between double reading and accuracy may mean that double reading is not implemented effectively in practice or that our characterization of double reading obscured the benefit of the double reading methods that work. An examination of double reading using a better characterization of the method is needed to identify whether the benefits of double reading demonstrated in trials are being realized in practice.

Our limited characterization of double reading was only one of the limitations of this observational study. It had other limitations as well. Although we accounted for important variation due to characteristics of women (breast density, age, and time-since-last mammogram) and radiologists (years of experience, reading volume), there is some risk that unmeasured variation in women and radiologists accounts for some of the variation we associated with facilities. We also had missing data for some questions and therefore had to exclude some facilities from analyses. For example, the data in Table 1 are based on screens from facilities that filled out the facility survey only, whereas the data in Table 2 are based on screens from facilities that filled out both a facility survey and a radiologist survey. The latter screens were a subset of the former, and the difference in which patients and radiologists were included in the analysis for Table 2 could have affected the results. It is also possible that facilities that employ a breast imaging specialist or perform diagnostic mammography may have placed an emphasis on treating breast disease, which resulted in a population of women whose characteristics were not accounted for by our model. These are limitations that cannot be completely avoided in observational research. Thus, we have emphasized that variation exists at the facility level rather than focusing on which factors might account for it. More work needs to be done to explain the variation, but we have shown that variation exists, and we suggest that it may be possible to explain some of the variation in terms of identifiable facility characteristics.

Understanding how facility characteristics influence interpretive accuracy is important because it could allow women and physicians to choose a mammography facility based on characteristics that are more likely to be associated with higher quality. Radiologists could also change the facilities' structures or processes to include practices that improve interpretive accuracy. For example, a facility might decide to include a breast imaging specialist in the practice if it was clear that such individuals would improve the facility's interpretive performance. It will also be important to determine exactly how breast imaging specialists make a difference in a facility's interpretive performance because of the limited number of breast imaging specialists that are available. In this study, a breast imaging specialist was defined based on the amount of time devoted to breast imaging. More information about qualifications, training, and the breast imaging specialists' interactions with other radiologists would provide greater insight into his/her effect on the interpretive performance of a facility. There is some evidence that radiologists who specialize in breast imaging have better interpretive skills than those who do not (23); it would be interesting to determine if the better interpretive performance we noted in facilities with a breast imaging specialist was a direct consequence of the interpretive skills of the breast imaging specialist or an indirect consequence of the specialist's professional interactions

with other radiologists at the facility during group interpretations of mammographic images.

Another facility practice that might be considered to identify high-quality facilities or to improve mammography performance is the use of double reading. Although this technique is discussed in the radiology literature and clinical practice, there are conflicting reports about its contribution to mammography interpretive performance (21,22). Our findings are consistent with a small negative effect of double reading on overall accuracy in that the AUC was lower for either double reading technique than for no double reading at all, but this negative effect on accuracy is largely the result of the effect of double reading on specificity (Table 2). Although we classified double reading as "independent" or "consensus" double reading, many subclassifications are possible—whether the radiologists are aware of the first reader's interpretation, required to commit to an interpretation before seeing the first reader's interpretation, required to always involve a third party when two radiologists disagree regarding an interpretation, or required to make a recommendation based on the most abnormal interpretation—all of which will affect performance measures and are differences that we did not ascertain. Furthermore, the facilities used double reading for varying proportions of their screening mammograms. Given the substantial potential variation in double reading practice and the weak negative association between the reading strategy and interpretive accuracy, it does not appear that facilities should be differentiated from each other based on whether or how they perform double reading until the different strategies and their effects are more clearly demonstrated in practice.

A final facility practice that might be considered to identify high-quality facilities or to improve mammography performance is whether the radiologists review their interpretive performance two or more times per year. This is the first report, to our knowledge, to examine the association between audit frequency and performance, but it is still too early to draw definitive conclusions. To understand how audits work and which activities improve accuracy, future studies should explore how audit information is provided to radiologists, whether radiologists review cases based on their audit reports every time they receive an audit, and what advantages are afforded by more frequent review.

This study has one additional limitation. Our results may also be affected by a number of selection biases; not all radiologists returned the survey and not all facilities reported their characteristics. Thus, the question remains whether our results are representative of all radiologists and facilities that could have participated. To address the effect of facility survey response bias, we compared the interpretive performance of radiologists practicing in facilities that responded to the facility survey and radiologists who practiced in facilities that did not respond to the facility survey. Only specificity differed statistically significantly between these two groups of radiologists. We suspect that including radiologists with this slightly higher specificity would have introduced more variation among the facilities but would not have changed our conclusions. Because this was a large study that included facilities from geographically diverse settings, it is unlikely that variation in interpretive performance was due to response bias alone. Therefore, we are confident that our findings show that facilities vary measurably in their performance and accuracy.

In summary, we found that interpretive performance differed statistically significantly across mammography facilities. We also found that higher interpretive accuracy of screening mammography was seen at facilities that offered screening examinations alone, that included a breast imaging specialist on staff, that did single reading (eg, did not do double reading), and that reviewed interpretive audits two or more times each year. These cross-sectional associations require prospective evaluation as we strive to improve the accuracy of mammography. It is possible that women and physicians could choose mammography facilities based on facility characteristics and that facilities could change their structure and processes to maximize quality. Identifying facility structures and processes that influence interpretive performance could be a foundation for improving the quality of mammography interpretive performance and choices among mammography facilities.

References

1. Joy JE, Penhoet EE, Petitti DB, eds. *Saving Women's Lives: Strategies for Improving Breast Cancer Detection and Diagnosis.* Washington, DC: The National Academies Press; 2005.
2. Kerlikowske K. Effect of age, breast density, and family history on the sensitivity of first screening mammography. *JAMA.* 1996;276(1):33–38.
3. Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Ann Intern Med.* 2003;138(3):168–175.
4. Yankaskas BC, Taplin SH, Ichikawa L, et al. Association between mammography timing and measures of screening performance in the United States. *Radiology.* 2005;234(2):363–373.
5. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. *J Natl Cancer Inst.* 2004;96(24):1840–1850.
6. Smith-Bindman R, Chu PW, Miglioretti DL, et al. Comparison of screening mammography in the United States and the United Kingdom. *JAMA.* 2003;290(16):2129–37.
7. Tan A, Freeman DHJ, Goodwin JS, Freeman JL. Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment. *Breast Cancer Res.* 2006;100(3):309–318.
8. Elmore JG, Miglioretti DL, Reisch LM, et al. Screening mammograms by community radiologists: variability in false-positive rates. *J Natl Cancer Inst.* 2002;94(18):1373–1380.
9. Ballard-Barbash R, Taplin SH, Yankaskas BC, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. *AJR Am J Roentgenol.* 1997;169(4):1001–1008.
10. Taplin SH. Evaluating organized breast cancer screening implementation: the prevention of late-stage disease? *Cancer Epidemiol Biomarkers Prev.* 2004;13(2):225–234.
11. Poplack SP, Tosteson AN, Grove MR, Wells WA, Carney PA. Mammography in 53,803 women from the New Hampshire mammography network. *Radiology.* 2000;217(3):832–840.
12. Ernster VL, Ballard-Barbash R, Barlow WE, et al. Detection of ductal carcinoma in situ in women undergoing screening mammography. *J Natl Cancer Inst.* 2002;94(20):1546–1554.
13. American College of Radiology. *Bi-Rads®—Mammography: Assessment Categories.* Reston, VA: American College of Radiology; 2003. February 2, 2004. http://www.acr.org/secondarymainmenucategories/quality_safety/biradsatlas.aspx.
14. Elmore JG, Taplin SH, Barlow WE, et al. Does litigation influence medical malpractice? The influence of community radiologists' medical malpractice perceptions and experience on screening mammography. *Radiology.* 2005;236(6):37–46.
15. Fishbein M. Factors influencing health behaviors: an analysis based on a theory of reasoned action. In: Landry F, ed. *Health Risk Estimation, Risk Reduction and Health Promotion.* Ottawa, Canada: CPHA; 1983:203–214.
16. Green LW, Kreuter MW. Health promotion today and a framework for planning. In: Green LW, Kreuter MW, eds. *Health Promotion Planning:*

An Educational and Environmental Approach. Mountain View, CA: Mayfield Publishing Company; 1991, p. 24.
17. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology breast imaging reporting and data system. *J Natl Cancer Inst.* 1999;90(23): 1801–1809.
18. Bandura A. Models of human nature and causality. In: Bandura A, ed. *Social Foundations of Thought and Action: a Social Cognitive Theory* Englewood Cliffs, NJ: Prentice-Hall, Inc; 1986:23.
19. *SAS* Cary, NC: SAS Institute Inc.; 2003. [computer program] Version 9.1.
20. Nass S, Ball J, eds. *Committee on Improving Mammography Quality Standards.* Washington, DC: National Academies Press; 2005.
21. Ciatto S, Ambrogetti D, Bonardi R, et al. Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme. *J Med Screen.* 2005;12(2):103–106.
22. Duijm LEM, Groenewould JH, Hendriks JHCL, de Koning H. Independent double reading of screening mammograms in the Netherlands: effect of arbitration following reader disagreements. *Radiology.* 2004;231(2):564–570.
23. Sickles EA, Wolverton DE, Dee KE. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology.* 2002;224(3):861–869.

Funding

Agency for HealthCare Research and Quality (public health service grant R01 HS010591 to J.G.E.); National Cancer Institute (R01 CA107623 to J.G.E., K05 CA104699 to J.G.E., U01 CA63731 to S.T., U01 CA86082-01 to P.A.C., 5 U01 CA63736-09 to G.R.C., and 5 U01 CA86076 to W.E.B).

Notes

C. D'Orsi holds stock options in Hologic Corporation (a developer, manufacturer, and supplier of breast imaging systems) and is a member of the company's medical board. The authors thank Sara Jackson for her careful review of and comments on the manuscript.

This work was conducted primarily while S. Taplin was at Group Health, although final writing was done while he was at the National Cancer Institute. All opinions are those of the coauthors and do not imply agreement or endorsement by the Federal Government or the National Cancer Institute. The analysis, interpretation, and reporting of the data; the writing of the manuscript; and the decision to publish were the sole responsibility of the authors.

Manuscript received November 5, 2007; revised April 7, 2008; accepted April 22, 2008.