Title

Correlation Between Screening Mammography Interpretive Performance on a Test Set and Performance in Clinical Practice

Journal

Academic Radiology, 24(10)

ISSN

1076-6332

Authors

Miglioretti, Diana L

Ichikawa, Laura

Smith, Robert A

et al.

Publication Date

2017-10-01

DOI

10.1016/j.acra.2017.03.016

Peer reviewed

HHS Public Access

Author manuscript

Acad Radiol. Author manuscript; available in PMC 2018 October 01.

Published in final edited form as:

Acad Radiol. 2017 October ; 24(10): 1256–1264. doi:10.1016/j.acra.2017.03.016.

Correlation between screening mammography interpretive performance on a test set and performance in clinical practice

Diana L. Miglioretti, PhD[1,2], **Laura Ichikawa, MS**[2], **Robert Smith, PhD**[3], **Diana S.M. Buist, PhD**[2], **Patricia A. Carney, PhD**[4], **Berta Geller, EdD**[5], **Barbara Monsees, MD**[6], **Tracy Onega, PhD**[7], **Robert Rosenberg, MD**[8], **Edward A. Sickles, MD**[9], **Bonnie Yankaskas, PhD**[10], and **Karla Kerlikowske, MD**[11]

[1]Division of Biostatistics, Department of Public Health Sciences, University of California Davis School of Medicine, One Shields Ave., Med Sci 1C, Room 145, Davis, CA 95616, USA

[2]Kaiser Permanente Washington Health Research Institute, 1730 Minor Ave., Suite 1600, Seattle, WA 98101, USA

[3]Cancer Control Department, American Cancer Society, 250 Williams Street, Suite 600, Atlanta, GA 30303, USA

[4]Departments of Family Medicine and Public Health and Preventive Medicine, School of Medicine: Mail Code FM, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239, USA

[5]Office of Health Promotion Research, University of Vermont, Department of Family Medicine, 106 Carrigan Dr., Rowell 235, Burlington, VT 05405, USA

[6]Department of Radiology, Washington University, Mallinckrodt Institute of Radiology, 510 S. Kingshighway Blvd, St. Louis, MO 63110, USA

[7]Department of Community and Family Medicine, Dartmouth Medical School, HB 7927 – Community & Family Medicine, HB 7927 Rubin 8, One Medical Center Dr., Lebanon, NH 03756, USA

[8]University of New Mexico-HSC and Radiology Associates of Albuquerque, 4411 The 25 Way NE Suite 150, Albuquerque, NM 87109, USA

[9]Department of Radiology, University of California, San Francisco Medical Center, Box 1667, San Francisco, CA 94143, USA

[10]Department of Radiology, University of North Carolina, 101 Manning Dr # 2, Chapel Hill, NC 27514, USA

Correspondence: Diana L. Miglioretti, PhD, Department of Public Health Sciences, UC Davis School of Medicine, One Shields Ave., Med Sci 1C, Room 145, Davis, CA 95616, Phone: (530) 752-7168, Fax: (530) 752-3239, dmiglioretti@ucdavis.edu.

[11]Departments of Medicine and Epidemiology and Biostatistics, University of California, San Francisco, CA; General Internal Medicine Section, 4150 Clement Street, Department of Veterans Affairs, University of California, San Francisco, CA 94121, USA

Abstract

Rationale and Objectives—Evidence is inconsistent about whether radiologists' interpretive performance on a screening mammography test set reflects their performance in clinical practice. This study aimed to estimate the correlation between test-set and clinical performance and determine if the correlation is influenced by cancer prevalence or lesion difficulty in the test set.

Materials and Methods—This institutional review board–approved study randomized 83 radiologists from six Breast Cancer Surveillance Consortium registries to assess one of four test sets of 109 screening mammograms each; 48 radiologists completed a fifth test set of 110 mammograms 2 years later. Test sets differed in number of cancer cases and difficulty of lesion detection. Test-set sensitivity and specificity were estimated using woman-level and breast-level recall with cancer status and expert opinion as gold standards. Clinical performance was estimated using women-level recall with cancer status as the gold standard. Spearman rank correlations between test-set and clinical performance with 95% confidence intervals (CI) were estimated.

Results—For test sets with fewer cancers (N=15) that were more difficult to detect, correlations were weak-to-moderate for sensitivity (woman-level=0.46, 95%CI=0.16, 0.69; breast-level=0.35, 95%CI=0.03, 0.61) and weak for specificity (0.24, 95%CI=0.01, 0.45) relative to expert recall. Correlations for test sets with more cancers (N=30) were close to zero and not statistically significant.

Conclusions—Correlations between screening performance on a test set and performance in clinical practice are not strong. Test set performance more accurately reflects performance in clinical practice if cancer prevalence is low and lesions are challenging to detect.

Keywords

Screening mammography; interpretive performance; test sets

Introduction

The interpretive performance of screening mammography varies extensively among U.S. radiologists.[1,2] Given U.S. radiologist have relatively low interpretive volume, on average,[3,4] and often do not work-up their own recalled cases,[5] they have limited opportunities to know, directly or indirectly, whether women they recalled or did not recall on screening mammograms experienced benign or malignant outcomes. A test set of selected mammography images could be an efficient method to assess radiologists' skill level and to identify potential opportunities for improvement. Additionally, test sets could help radiologists meet Part 2 of the American Board of Radiology's Maintenance of Certification requirements (Lifelong-learning and Self-assessment).[6]

Findings from prior studies are inconsistent about whether interpretive performance on screening mammography test sets is correlated with performance in clinical practice,

possibly due to small samples (of radiologists and/or images) and variability in test set composition, performance measures evaluated, and statistical approaches used.[7-9] In a study of 27 U.S. radiologists who interpreted a test set of 113 film screening mammography examinations (30 with cancer), Rutter and Taplin[9] found moderate correlation between the specificity of screening mammography interpreted in clinical and test settings (0.41; 95% Bayesian credible interval (BCI) 0.16, 0.62), but no evidence of correlation between clinical and test-set sensitivity (-0.18, 95% BCI -0.27, 0.59). In contrast, Soh et al.[7] found significant, moderate correlations of 0.30-0.57 between several clinical audit measures and two test set measures (location sensitivity and Jackknifing free response operating characteristic figure-of-merit) of 60 cases (20 with cancer) read by 20 radiologists, but no correlation with test set specificity. Similarly, Scott et al.[8] found significant, moderate correlations of 0.29-0.41 between several performance measures on the PERFORMs test set and clinical performance among 39 readers in the UK. None of these prior studies evaluated the influence of breast cancer prevalence or lesion difficulty on the strength of the correlations.

In this study, we created five tests sets with different cancer prevalence and varying levels of difficulty detecting cancerous lesions. We sought to determine whether performance on the test set was correlated with performance in clinical practice, and whether these associations depend on cancer prevalence or difficulty.

Material and Methods

Study Population

Radiologists interpreting mammography at facilities participating in one of six Breast Cancer Surveillance Consortium (BCSC) registries between January 2005 and December 2006 were invited to participate as part of a larger randomized trial that also included non-BCSC radiologists.[10] Participating BCSC registries included the Carolina Mammography Registry, Group Health Surveillance Registry in Washington State, New Hampshire Mammography Network, San Francisco Mammography Registry, New Mexico Mammography Project, and Vermont Breast Cancer Surveillance System. Because this study required an estimate of clinical performance, we only included radiologists with at least 10 screening mammograms with cancer for estimating sensitivity and/or 100 screening mammograms without cancer for estimating specificity in the BCSC database. A total of 83 radiologists with a sufficient number of screening mammograms for estimating clinical performance completed at least one test set.

Each site received Institutional Review Board approval for study activities. Informed consent was obtained from radiologists participating in the study. Active or passive consent and/or waivers of consent were obtained from women receiving mammograms at a BCSC facility. All procedures complied with the Health Insurance Portability and Accountability Act. Identities of women, physicians, and facilities are protected by a Federal Certificate of Confidentiality and other protections. Radiologists received up to eight free Category I continuing medical education credits for interpreting a test set.

Test Set Development

We developed five test sets, each with 110 cases. For test sets 1-4, one case was incorrectly uploaded into the system, leaving 109 cases for analysis. Test sets 1-4 shared 91 cases. Test set 5 shared 58 normal exams without cancer with one of the first four test sets.

Test set development is described in detail elsewhere.[11] Briefly, we sampled 314 screening mammograms performed at a BCSC facility from 2000-2003 on women aged 40-69 years who also had a previous mammogram within the prior 11-30 months for use as comparison. We excluded exams performed on women with a history of breast cancer, mastectomy and/or breast augmentation. Each test-set case consisted of craniocaudal and mediolateral oblique views of each breast with comparison views from the prior 11-30 months.

American College of Radiology (ACR) staff digitized the film-screen mammography images. We created an expert panel of three senior breast-imaging specialists who taught at academic medical centers.[12] Each expert independently reviewed the digitized images using custom-designed software while blinded to the woman's cancer status, and indicated whether the woman should be recalled. Examinations of insufficient quality or with marks were flagged for exclusion. For recalled images, experts classified the most significant finding as a mass, calcification, asymmetrical density, or architectural distortion and assigned a level of difficulty of identifying the lesion as obvious, intermediate, or subtle. Consensus expert opinion was taken to be the agreement of at least two of three experts for each measure and the remaining examinations were resolved during a consensus meeting.[12]

The test sets differed by cancer prevalence and case difficulty (Table 1). Test sets 1, 2, and 5 had lower cancer prevalence (15 cancer cases) than test sets 3 and 4 (30 cancer cases). Cancer cases in test sets 2, 4, and 5 were more challenging, with 33% considered subtle and 20% considered obvious by the experts, compared to tests sets 1 and 3, which had 13% considered subtle and 33% considered obvious.

Radiologists' Interpretive Performance on the Test Set

The 83 radiologists were evaluated using one of four test sets from July 2008 to August 2009; 48 of these radiologists also completed the fifth test set from June 2010 to January 2011. Radiologists completed a brief questionnaire of their personal characteristics before starting the test set. For the first test set round, radiologists participating in this study were randomized along with radiologists from the larger study to one of the four test sets using a block randomization scheme with stratification by BCSC registry and number of breast cancer cases in the BCSC database (less than versus at least 30 cases). As part of the larger study, all radiologists who took one of the first four test sets were invited to participate in a randomized controlled trial of two educational interventions, either a live seminar or a DVD.[10] The intervention groups were invited to take test set 5 after the intervention was completed and the control group was sent the intervention DVD after completing test set 5.

Radiologists took the test sets, which were sent to them on a DVD, using a computer of their choice, or a laptop provided by the study, with display requirements that allowed viewing two images concurrently with sufficient resolution relative to the state-of-the art technology available at that time. The display requirement for a desktop PC was a screen at least 17-

inches with a resolution of at least 1280 ×1024; a laptop needed to have at least a 15.4-inch screen with a resolution of 1440×900 or higher. Radiologists could magnify the images to inspect areas of interest.

Radiologists were instructed to interpret the images as they would in clinical practice, except to record only the most significant finding, if any. Radiologists were informed that the test sets had been cancer-enriched relative to a screening population but the cancer prevalence was not revealed. For each case, radiologists indicated whether they would recall or not recall the case, based on the ACR Breast Imaging Reporting and Data System (BI-RADS) 4th edition lexicon definition: recall if assessment codes 0, 4 or 5; and no recall if assessment codes 1 or 2.[13] For recalled cases, radiologists indicated the location by clicking on the image(s). *Cancer cases* were defined as women with a diagnosis of ductal carcinoma *in situ* or invasive carcinoma within 12 months of the screening examination. Non-cancer cases were defined as women without a cancer diagnosis for at least 24 months following the screen. Recalled cases without cancer were defined to be *appropriate recalls* if the expert panel determined additional imaging was necessary for a finding on the mammogram, because the chance of cancer based on the screening results was sufficiently high to necessitate further evaluation.

For the test set, sensitivity and specificity were calculated relative to two gold standards: cancer status and expert judgment of appropriate recall. *Expert recall* included true positives and false-positive recalls that the experts deemed appropriate to rule out cancer.[11] For *woman-specific sensitivity*, a cancer case or screen recalled by the experts was considered a true positive if either breast was recalled. For *breast-specific sensitivity*, the radiologist had to recall the breast with the significant finding for it to be a true positive; otherwise, if the radiologist recalled the other breast, it was considered a false negative. Specificity was only measured at the woman-level because this is what is most clinically relevant – either the woman was recalled or not, regardless of which breast was recalled. For *specificity*, a non-cancer case or screen not recalled by the experts was a false-positive if either breast was recalled and a true negative otherwise.

Clinical Performance on the Test Set Exams

For each screening exam included in the test sets, we recorded the clinical assessment. Of the 182 unique test set screening exams, 146 (80%) were interpreted clinically by radiologists who did not participate in this study. Mammograms were considered to be positive or negative at the woman-level based on the initial BI-RADS assessment using the ACR BI-RADS 4th edition definition of recall.[13] We used the woman-level measure of recall because we often did not have separate clinical assessments for each breast. *Cancer cases* were defined using the same definition as for test set performance. Woman-level sensitivity and specificity were calculated relative to cancer status.

Radiologists' Interpretive Performance in Clinical Practice

For estimation of each radiologist's clinical performance, we included screening mammograms performed on women age 40 years or older. We excluded unilateral examinations and examinations performed within 9 months of a prior mammogram to avoid

classifying diagnostic exams as screening. We also excluded examinations performed on women with a history of breast cancer, mastectomy, or breast augmentation. For comparison with test sets 1-4, we included screening exams interpreted in the 3 years prior to the date the radiologist started the test set. For comparison with test set 5, we included screening exams interpreted within one-year post-intervention for the intervention groups and 1 year prior to the intervention for the control group.

For calculating clinical performance, mammograms were considered to be positive or negative at the woman-level based on the initial BI-RADS assessment using the ACR BI-RADS 4th edition definition of recall.[13] We used the woman-level measure of recall because we often did not have separate clinical assessments for each breast. Woman-level clinical sensitivity and specificity were calculated using cancer status as the gold standard. *Cancer cases* were defined as a diagnosis of ductal carcinoma *in situ* or invasive cancer within 12 months of the screening examination and prior to the next screen. Otherwise, the case was considered not to have cancer.

Analysis

We compared mean test set interpretive performance to the performance of the same exams interpreted in clinical practice using the woman-level recall relative to cancer status. We estimated and compared mean test set and clinical performance by fitting logistic regression models using a three-step generalized estimating equations approach to account for correlation among examinations interpreted by the same radiologists and among radiologists interpreting the same examinations.[14,15] For sensitivity, we modeled the probability of a positive assessment among cancer cases. For specificity, we modeled the probability of a negative assessment among non-cancer cases.

Separately for each test set, we estimated the association between each radiologist's performance on the test set to their interpretive performance in clinical practice using the Spearman rank correlation coefficient, which is a nonparametric measure of the association between the ranking of two variables. In other words, it measures whether radiologists who perform well on the test set relative to others also perform well relative to others in clinical practice. We considered a correlation of less than 0.2 to be very weak, 0.2-0.4 to be weak, 0.4-0.6 to be moderate, and 0.6 or higher to be strong. We also calculated Spearman rank correlation coefficients combining test sets with the same cancer prevalence (Lower prevalence = test sets 1, 2, and 5 and higher prevalence = test sets 3 and 4) and with the same difficulty (less difficult cancers = test sets 1 and 3 and more difficult cancers = test sets 2, 4, and 5).

Results

Table 2 summarizes the characteristics of the participating radiologists. Only 6% self-identified as breast imaging specialists and only 8% had fellowship training in breast or woman's imaging. Most were not affiliated with academic institutions (92%), and the majority (67%) interpreted mammograms for >10 years. More than half of participating radiologists (70%) reported working ≤3 days a week in breast imaging, and 60% reported interpreting an average of <100 mammograms per week.

Table 3 shows the overall, woman-level sensitivity and specificity on each test set compared to performance for the same screening exams in clinical practice (80% of which were interpreted clinically by a radiologist who did not participate in this study) using cancer status as the gold standard. Sensitivity on the test sets ranged from 72%-83% and these values were not significantly different from the clinical sensitivity values for the same screens, which ranged from 73%-83% ($p > 0.12$ for all test sets). Specificity on the test sets ranged from 62%-69% compared to 47%-62% for the same screens interpreted in clinical practice; these differences in specificity were statistically significant for test sets 3 and 4 ($p \leq 0.001$) and borderline significant for test set 1 ($p = 0.07$).

eTable 1 shows the woman-level characteristics of the screening exams included in each test set and the 373,058 screens used to calculate the participating radiologists' clinical performance. Test-set exams had a younger age distribution by design, which also resulted in test-set exams having higher breast density. For measuring clinical performance, 87.6% of screens had a mammogram in the prior 1-2 years and 91% had comparison views available, compared to 100% of the screens included in the test sets by design. For the calculation of clinical specificity for comparison with test sets 1-4, the 83 participating radiologists interpreted an average of 3,554 screening mammograms without cancer (range 199 to 15,576) during the three years prior to taking the test set; 58 of these radiologists with at least 10 cancer cases for calculation of sensitivity interpreted an average of 26 screening mammograms with breast cancer (range 10 to 88). For comparison with test set 5, 48 radiologists interpreted an average of 1,580 screening mammograms without cancer (range 139 to 4,378) within one-year post-intervention for the intervention groups and 1 year prior to the intervention for the control group of the larger study; 20 of these radiologists with sufficient cancer cases for calculation of sensitivity interpreted an average of 14 screening mammograms with breast cancer (range 10 to 30).

Figures 1 and 2 show each radiologist's sensitivity and specificity in clinical practice versus their own performance on the test set using expert recall as the gold standard. Clinical sensitivity and specificity tended to be higher and less variable across radiologists. Clinical sensitivity averaged 87% and varied from 65% to 100% across radiologists compared to an average sensitivity of 69% on the test set (range 42% to 90%); average specificity was 91% in clinical practice (range 72% to 98%) compared to an average of 74% on the test set (range 30% to 97%). Combing data from all test sets, the Spearman rank correlations are very weak and not statistically significant (woman-level sensitivity correlation=0.15, 95% confidence interval (CI) = -0.08, 0.36, specificity correlation=0.14, 95% CI = -0.03, 0.30).

When evaluating correlations separately by test set, most correlations are not significantly greater than zero; however, confidence intervals are wide given the small number of radiologists who completed each test set (eTable 2). One exception is for test set 5, for which woman-level test-set sensitivity among cancer cases is significantly correlated with sensitivity in clinical practice (0.44, 95% CI = 0.0, 0.74, p=0.0496), and specificity among non-cancer test-set 5 cases was significantly, although weakly, correlated with specificity in clinical practice (0.29, 95% CI = 0.00, 0.53, p=0.048).

Table 4 shows the Spearman rank correlations between test set and clinical performance grouping tests sets with the same number of cancer cases, the same lesion difficulty, or both. Correlations were highest for test sets with lower cancer prevalence, especially if the lesions were considered by the experts to be more difficult to detect. Correlations were weak-to-moderate and significantly greater than zero for woman- and breast-level sensitivity among test-set cases recalled by the experts for test sets 2 and 5, which had 15 cancer cases with 33% considered subtle and 47% considered intermediate difficulty to detect (Spearman rank correlation = 0.46, 95% CI = 0.16, 0.69 for woman-level and 0.35, 95% CI = 0.03, 0.61 for breast-level sensitivity). For these same test sets, correlations for specificity were also significantly greater than zero, though they were only weak (Spearman rank correlation = 0.28, 95% CI 0.05, 0.48 among all non-cancer test set cases and 0.24, 95% CI 0.01, 0.45 among cases not recalled by the experts).

Discussion

Our study is the first to explore the effect of test set composition, in terms of cancer prevalence and lesion difficulty, on the correlation between the interpretive performance of screening mammography on a test set versus performance in clinical practice. We found, at best, only weak-to-moderate correlations and these correlations only reached statistical significance for test sets with lower cancer prevalence, which is somewhat closer to the experience in clinical practice. Correlations were strongest when lower cancer prevalence was coupled with lesions that were challenging to detect.

The correlations we found were similar in magnitude to prior studies; however, studies differ in which performance measures were significantly correlated.[7-9] Similar to Soh et al.,[7] we found stronger correlations for sensitivity than for specificity. This is in contrast to Rutter and Taplin,[9] who found moderate correlations for specificity but no evidence of correlation between clinical and test-set sensitivity. Scott et al.[8] also found a significant, moderate correlation between clinical specificity and the “correct return to screen percentage” on the test set, but no significant correlation for their measure of sensitivity (correct recall percentage); however, they found significant correlations for other performance measures including positive predictive value, false-negative rate, and cancer detection rate. Taken together, these studies suggest that test sets may reflect clinical practice, but only moderately. No study has directly measured whether this level of correlation is strong enough to accurately identify radiologists who might benefit from additional training or specific areas for improvement.

We found stronger correlations with clinical performance with lower test set cancer prevalence (N=15 cancers or 14% versus N=30 cancers or 28%), and when cancerous lesions were difficult to detect. It seems reasonable that obvious cancers, i.e., cancers that most radiologists should easily detect, would not help discriminate between radiologists with different levels of interpretative performance. It is less clear why lower cancer prevalence would improve the correlations. In a study of 14 radiologists who interpreted tests sets of chest images with 5 different prevalence levels of lung abnormalities, Gur et al. found no effect of prevalence on performance measured by the area under the receiver operating characteristic curves;[16] however, confidence that the abnormality of interest was

present was higher for test sets with lower prevalence.[17] Evans, et al.[18] found six radiologists missed 30% of 100 cancer cases that were “seeded” into clinical practice at a rate slow enough to maintain a 1% cancer prevalence, compared to only 12% when these same exams were interpreted on a test set for which 50% of the screens were cancers (a very high prevalence). The authors attribute this difference to the low cancer rate in clinical practice; however, this is difficult to confirm given the exams were interpreted in different settings.

We evaluated rank correlations between a radiologist's test set and clinical performance, because it is possible that a test set may accurately measure relative performance across radiologists, i.e., if some radiologists perform better or worse than others, even if they do not reflect the actual level of performance in clinical practice. Additionally, absolute (versus relative) performance on a test set will strongly depend on the test set composition, mostly the difficulty of the test set exams. Even when interpreting the same exams in a clinical practice versus on a test set, absolute performance measures may differ due to the setting. In a study of nine radiologists who interpreted 276-300 film-screen mammograms, Gur et al.[19] found that screening mammography performance was better and less variable in clinical practice than on the same examinations in a test set, but they did not assess whether test set performance reflected relative performance in clinical practice. In contrast, Soh et al.[20] compared the performance of ten radiologists interpreting 200 screening mammograms (up to 20 with cancer) both in clinical practice and on a test set, and they found good agreement if comparison images were available. Comparing performance on the same exams interpreted by different radiologists in different settings, we found test set sensitivity was slightly lower than, but not significantly different from, clinical sensitivity, but that that test set specificity was higher than clinical specificity, reaching statistical significance for two test sets. These small differences may be due to the films having been digitized and interpreted on a computer, which may have made both malignant and benign lesions less conspicuous in the test setting.

Test sets could be important mechanisms for meeting Maintenance of Certification Requirements of the American Board of Radiology, because they allow radiologists to understand how their interpretive performance compares to other radiologists and with a panel of experts. Test sets could also be useful for identifying specific areas for improvement, for example, based on lesion characteristics. Studies have shown that physicians do not change their behavior if they do not perceive there is a need to do so.[21] Thus, assessing radiologist skill levels and identifying areas for improvement is now a part of the Maintenance of Certification. In a screening environment, sensitivity and cancer detection rate are difficult to assess using clinical audits, because cancer is rare in a single clinical setting, and sufficient numbers of cancers with specific malignancy features are even rarer. It may take many years for a radiologist to interpret enough cancer cases to precisely estimate their sensitivity or cancer detection rate, or to identify specific opportunities for improvement. Additionally, most facilities cannot link screening mammograms to state or regional cancer registries to capture false negative exams, which are needed to calculate sensitivity and specificity. Test sets overcome the limitations of clinical audits by providing immediate feedback on a relatively large number and variety of cancer cases in a shorter period of time than possible in clinical practice. In addition, assessment in the clinical

setting over a duration of time sufficient to accumulate a large enough number of cancers could reflect performance over a period of time when interpretive skill was changing, whereas performance on a test set should reflect current skill levels, and thus, provide less ambiguous feedback.

In our study, an expert panel determined whether each case should be recalled for further work-up, regardless of true cancer status, creating a useful reference standard of "appropriate recall." In screening, some mammographic signs need to be recalled to determine if they are cancer or not, and it may not be appropriate to penalize radiologists in their outcome measures for recalling these examinations for further evaluation if they turn out to be benign. It would be very challenging to determine, on a large scale, which recalled benign findings in a clinical setting were "appropriate." Test sets can provide performance feedback on which benign findings did not need to be recalled based on lesion or other image characteristics, allowing radiologists to assess specific areas where they may benefit from additional training.

Our study has several limitations. We designed our study before the widespread diffusion of digital mammography; therefore, we professionally digitized films for inclusion in our test sets, which reduces image quality. In addition, test sets were interpreted using personal computers, which could have lower resolution and smaller field of view than clinical workstations. This may have compromised interpretation due to poorer image quality compared to that in clinical practice. All readers (and the experts) experienced the same limitation in image quality; however, it is possible that the overall reduction in image quality may have changed the magnitude of differences between more vs. less accurate interpreting radiologists. Statistical power was limited within each test set, resulting in wide confidence intervals; however, power was higher for comparisons of test sets with similar attributes.

In conclusion, we found test set interpretive performance of screening mammography is, at best, only weakly-to moderately correlated with performance in clinical practice. Our results suggest test sets may be more reflective of relative performance across radiologists in clinical practice if test set cancer prevalence is low and lesions are difficult to detect. The absence of a strong correlation between test set performance and clinical performance does not discount the importance and potential of test sets to measure interpretive skill levels and identify opportunities for improvement. Given the limited opportunity for radiologists to assess their sensitivity and specificity against desired benchmarks, and across the range of mammographic findings that are associated with both malignancy and benign conditions, future studies should evaluate whether test sets can be used to accurately identify opportunities for improvement in radiologists' skills and can motivate change.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the American Cancer Society, made possible by a generous donation from the Longaberger Company's Horizon of Hope® Campaign (SIRSG-07-271, SIRSG-07-272, SIRSG-07-273,

SIRSG-07-274-01, SIRSG-07-275, SIRGS-06-281, SIRSG-09-270-01, SIRSG-09-271-01, SIRSG-06-290-04), the Breast Cancer Stamp Fund, and the National Cancer Institute Breast Cancer Surveillance Consortium (HHSN261201100031C). The collection of cancer data used in this study was supported in part by several state public health departments and cancer registries throughout the U.S. For a full description of these sources, please see: http://www.breastscreening.cancer.gov/work/acknowledgement.html. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, the National Institutes of Health, or the American Cancer Society. We thank the participating women, mammography facilities and radiologists for the data they have provided for this study. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: http://breastscreening.cancer.gov/.

Funding Source: This work was supported by the American Cancer Society using a donation from the Longaberger Company's Horizon of Hope Campaign (SIRSG-07-271, SIRSG-07-272, SIRSG-07-273, SIRSG-07-274-01, SIRSG-07-275, SIRSG-06-281, SIRSG-09-270-01, SIRSG-09-271-01, SIRSG-06-290-04) and by the Breast Cancer Stamp Fund. Collection of clinical mammography data was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (BCSC; HHSN261201100031C). The collection of cancer data was supported in part by several state public health departments and cancer registries throughout the U.S.; For a full description of these sources, please see: http://breastscreening.cancer.gov/work/acknowledgement.html.

Role of the Funder: The funding agencies had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Appendix

eTable 1

Characteristics of screening mammography examinations included in the test sets and in the clinical performance data

	Screens in test sets										Screens in clinical data	
	Test set 1		Test set 2		Test set 3		Test set 4		Test set 5			
Characteristic	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Total number of exams	109		109		109		109		110		373,058	
Number of cancers	15		15		30		30		15		2,025	
Woman's age, years												
40-49	44	(40.4)	45	(41.3)	40	(36.7)	41	(37.6)	37	(33.6)	99,800	(26.8)
50-59	43	(39.4)	40	(36.7)	43	(39.4)	43	(39.4)	45	(40.9)	119,583	(32.1)
60-69	22	(20.2)	24	(22.0)	26	(23.9)	25	(22.9)	28	(25.5)	86,916	(23.3)
≥70											66,759	(17.9)
Time since last mammogram												
No previous											10,594	(3.0)
1-2 years	109	(100)	109	(100)	109	(100)	109	(100)	110	(100)	314,605	(87.6)
3-4 years											21,757	(6.1)
5+ years											12,106	(3.4)
Missing											13,996	
Comparison film												
Yes	109	(100)	109	(100)	109	(100)	109	(100)	110	(100)	281,976	(90.6)
No											29,089	(9.4)
Missing											61,993	
Family history of breast cancer												
Yes	15	(15.5)	15	(15.6)	16	(16.8)	17	(17.9)	10	(9.6)	61,847	(17.2)
No	82	(84.5)	81	(84.4)	79	(83.2)	78	(82.1)	94	(90.4)	296,718	(82.8)

	Screens in test sets										Screens in clinical data	
	Test set 1		Test set 2		Test set 3		Test set 4		Test set 5			
Characteristic	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Missing	12		13		14		14		6		14,493	
BI-RADS breast density												
Almost entirely fat	6	(6.1)	6	(6.1)	4	(4.2)	4	(4.1)	7	(6.9)	45,020	(14.1)
Scattered	34	(34.3)	34	(34.3)	31	(32.3)	30	(30.9)	31	(30.4)	140,207	(43.8)
Heterogeneously dense	50	(50.5)	51	(51.5)	52	(54.2)	53	(54.6)	54	(52.9)	109,016	(34.1)
Extremely dense	9	(9.1)	8	(8.1)	9	(9.4)	10	(10.3)	10	(9.8)	25,634	(8.0)
Missing	10		10		13		12		8		53,181	

eTable 2

Spearman rank correlation (95% confidence interval) between radiologists' test set and clinical performance measures.

	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5
Sensitivity					
Number of radiologists with at least 10 cancer cases in clinical practice	16	16	11	15	20
Test set cancer cases					
Woman-level, correlation	-0. 11 (-0.57, 0.41)	0.1 9 (-0.34, 0.63)	0.3 8 (-0.29, 0.80)	-0. 51 (-0.81, 0.01)	**0.4 4 (0.00, 0.74)**
Breast-level, correlation	-0.01 (-0.50, 0.49)	0.30 (-0.23, 0.69)	0.30 (-0.36, 0.76)	-0.47 (-0.79, 0.05)	0.36 (-0.10, 0.69)
Test set cases recalled by experts					
Woman-level, correlation	0.0 7 (-0.44, 0.54)	0.3 5 (-0.17, 0.72)	-0. 10 (-0.66, 0.53)	-0. 51 (-0.81, 0.00)	0.4 4 (0.00, 0.74)
Breast-level, correlation	0.10 (-0.42, 0.57)	0.20 (-0.33, 0.63)	-0.02 (-0.61, 0.59)	-0.60 (-0.85, -0.13)	0.37 (-0.09, 0.70)
Specificity					
Number of radiologists with at least 100 non-cancer cases in clinical practice	22	24	17	20	48
Test set non-cancer cases, correlation	0.13 (-0.31, 0.52)	0.22 (-0.20, 0.57)	-0.28 (-0.67, 0.23)	0.25 (-0.22, 0.62)	**0.29 (0.00, 0.53)**
Test set cases not recalled by experts, correlatio	0.05 (-0.38, 0.46)	0.17 (-0.25, 0.53)	-0.30 (-0.68, 0.21)	0.30 (-0.16, 0.66)	0.25 (-0.04, 0.50)

Numbers in bold are significantly greater than zero.

References

1. Elmore JG, Jackson S, Abraham L, Miglioretti DL, Carney PA, Geller B, Yankaskas BC, Kerlikowske K, Onega T, Rosenberg RD, Sickles EA, Buist DSM. Variability in interpretive performance at screening mammography and radiologist characteristics associated with accuracy. Radiology. 2009; 253(3):641–51. [PubMed: 19864507]
2. Rosenberg RD, Yankaskas BC, Abraham LA, Sickles EA, Lehman CD, Geller BM, Carney PA, Kerlikowske K, Buist DS, Weaver DL, Barlow WE, Ballard-Barbash R. Performance Benchmarks for Screening Mammography. Radiology. 2006; 241(1):55–66. [PubMed: 16990671]

3. Smith-Bindman R, Miglioretti DL, Rosenberg R, Reid RJ, Taplin SH, Geller BM, Kerlikowske K. Physician workload in mammography. AJR Am J Roentgenol. 2008; 190(2):526–32. Epub 2008/01/24. [PubMed: 18212242]
4. Lewis RS, Sunshine JH, Bhargavan M. A portrait of breast imaging specialists and of the interpretation of mammography in the United States. AJR Am J Roentgenol. 2006; 187(5):W456–68. [PubMed: 17056875]
5. Buist DS, Anderson ML, Smith RA, Carney PA, Miglioretti DL, Monsees BS, Sickles EA, Taplin SH, Geller BM, Yankaskas BC, Onega TL. Effect of radiologists' diagnostic work-up volume on interpretive performance. Radiology. 2014; 273(2):351–64. Epub 2014/06/25. [PubMed: 24960110]
6. American Board of Radiology. American Board of Radiology Maintenance of Certification. 2016. #x0005B;10/05/2016]; version 2.2.2.: [Available from: https://www.theabr.org/moc-gen-landing
7. Soh BP, Lee WB, Mello-Thoms C, Tapia K, Ryan J, Hung WT, Thompson G, Heard R, Brennan P. Certain performance values arising from mammographic test set readings correlate well with clinical audit. J Med Imaging Radiat Oncol. 2015 Epub 2015/04/02.
8. Scott HJ, Evans A, Gale AG, Murphy A, Reed J. The relationship between real life breast screening and an annual self assessment scheme. Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment. 2009; 7263:72631E.
9. Rutter CM, Taplin S. Assessing mammographers' accuracy. A comparison of clinical and test performance. J Clin Epidemiol. 2000; 53(5):443–50. [PubMed: 10812315]
10. Geller BM, Bogart A, Carney PA, Sickles EA, Smith R, Monsees B, Bassett LW, Buist DM, Kerlikowske K, Onega T, Yankaskas BC, Haneuse S, Hill D, Wallis MG, Miglioretti D. Educational interventions to improve screening mammography interpretation: a randomized controlled trial. AJR Am J Roentgenol. 2014; 202(6):W586–96. Epub 2014/05/23. [PubMed: 24848854]
11. Carney PA, Bogart TA, Geller BM, Haneuse S, Kerlikowske K, Buist DS, Smith R, Rosenberg R, Yankaskas BC, Onega T, Miglioretti DL. Association between time spent interpreting, level of confidence, and accuracy of screening mammography. AJR Am J Roentgenol. 2012; 198(4):970–8. Epub 2012/03/28. [PubMed: 22451568]
12. Onega T, Anderson ML, Miglioretti DL, Buist DS, Geller B, Bogart A, Smith RA, Sickles EA, Monsees B, Bassett L, Carney PA, Kerlikowske K, Yankaskas BC. Establishing a gold standard for test sets: variation in interpretive agreement of expert mammographers. Acad Radiol. 2013; 20(6): 731–9. Epub 2013/05/15. [PubMed: 23664400]
13. American College of Radiology. American College of Radiology Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas). 4th. Reston, VA: Am Coll Radiol; 2003.
14. Miglioretti DL, Heagerty PJ. Marginal modeling of multilevel binary data with time varying covariates. Biostatistics. 2004; 5(3):381–98. [PubMed: 15208201]
15. Miglioretti DL, Heagerty PJ. Marginal modeling of nonnested multilevel data using standard software. Am J Epidemiol. 2007; 165(4):453–63. Epub 2006/11/24. [PubMed: 17121864]
16. Gur D, Rockette HE, Armfield DR, Blachar A, Bogan JK, Brancatelli G, Britton CA, Brown ML, Davis PL, Ferris JV, Fuhrman CR, Golla SK, Katyal S, Lacomis JM, McCook BM, Thaete FL, Warfel TE. Prevalence effect in a laboratory environment. Radiology. 2003; 228(1):10–4. [PubMed: 12832568]
17. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The prevalence effect in a laboratory environment: Changing the confidence ratings. Acad Radiol. 2007; 14(1):49–53. [PubMed: 17178365]
18. Evans KK, Birdwell RL, Wolfe JM. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. PLoS One. 2013; 8(5):e64366. [PubMed: 23737980]
19. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, Perrin RL, Poller WR, Shah R, Sumkin JH, Wallace LP, Rockette HE. The “Laboratory” Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations. Radiology. 2008; 249(1):47–53. Epub 2008/08/07. [PubMed: 18682584]

20. Soh BP, Lee W, McEntee MF, Kench PL, Reed WM, Heard R, Chakraborty DP, Brennan PC. Screening mammography: test set data can reasonably describe actual clinical reporting. Radiology. 2013; 268(1):46–53. Epub 2013/03/14. [PubMed: 23481165]

21. Duffy FD, Holmboe ES. Self-assessment in lifelong learning and improving performance in practice: physician know thyself. JAMA. 2006; 296(9):1137–9. Epub 2006/09/07. [PubMed: 16954495]

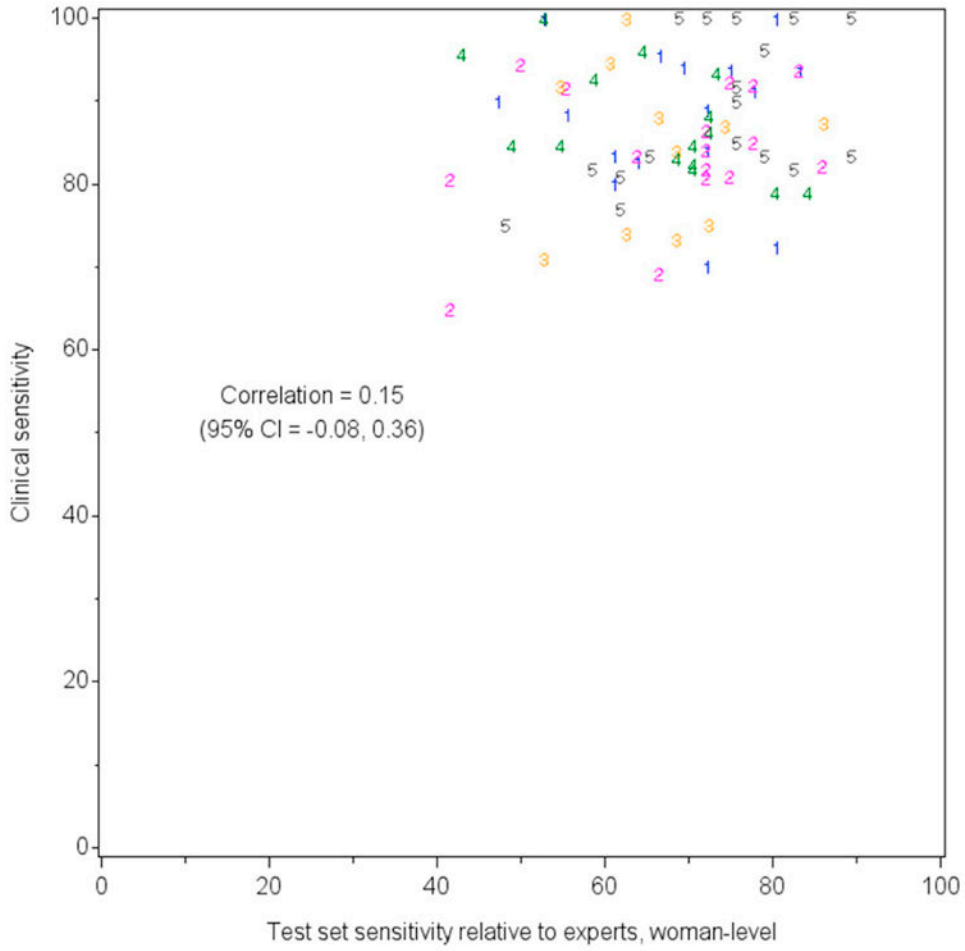

Figure 1.

Woman-level sensitivity in clinical practice versus woman-level test-set sensitivity among cases recalled by the expert panel. Symbols 1-5 indicate the test set number. Test set 1 had lower cancer prevalence and less difficult lesions, test sets 2 and 5 had lower cancer prevalence and more difficult lesions, test set 3 had higher cancer prevalence and less difficult lesions, and test set 4 had higher cancer prevalence and less difficult lesions. CI=confidence interval.

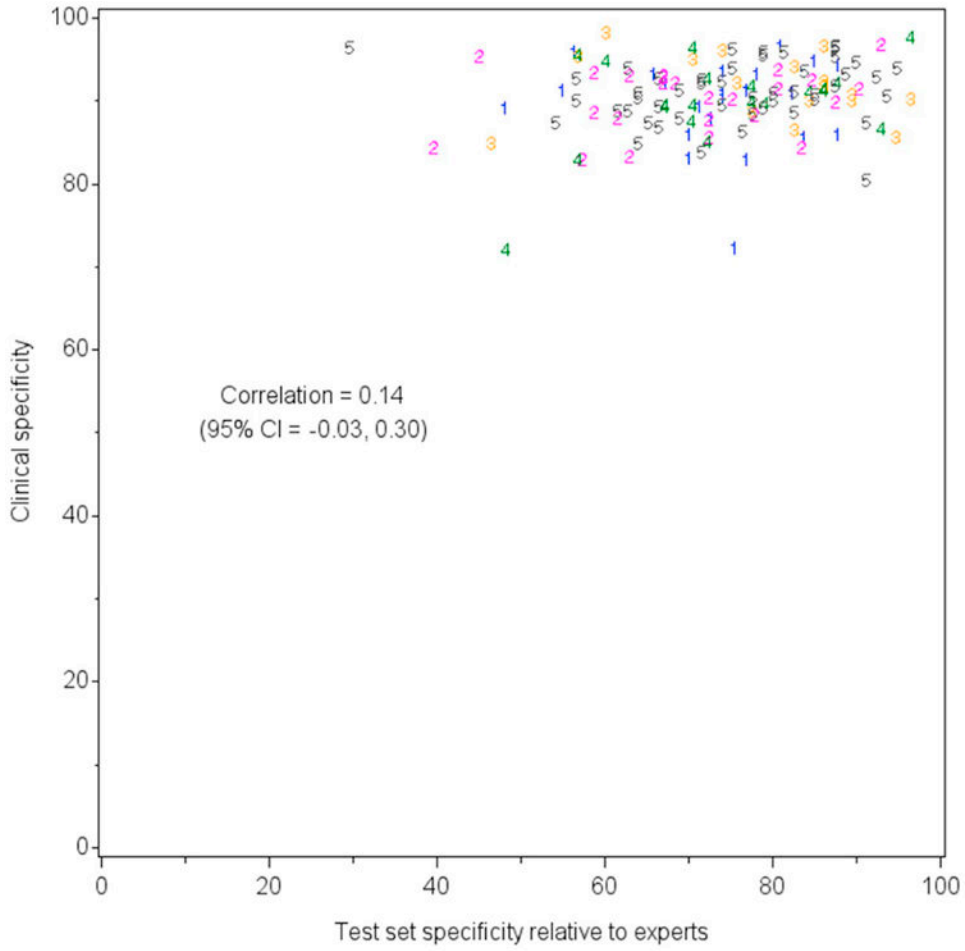

Figure 2.

Woman-level specificity in clinical practice versus woman-level test-set specificity among cases not recalled by the expert panel. Symbols 1-5 indicate the test set number. Test set 1 had lower cancer prevalence and less difficult lesions, test sets 2 and 5 had lower cancer prevalence and more difficult lesions, test set 3 had higher cancer prevalence and less difficult lesions, and test set 4 had higher cancer prevalence and less difficult lesions. CI=confidence interval.

Table 1

Test set composition.

	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5
Number of exams	109	109	109	109	110
Screens with cancer, *N*	15	15	30	30	15
Difficulty to detect, *N*(%)					
Obvious	5 (33 %)	3 (20 %)	10 (3 3%)	6 (20 %)	3 (20%)
Intermediate	8 (53%)	7 (47%)	16 (53%)	14 (47%)	7 (47%)
Subtle	2 (13%)	5 (33%)	4 (13%)	10 (33%)	5 (33%)
Finding type, *N*(%)					
Mass	4 (27 %)	3 (20 %)	9 (30 %)	6 (20 %)	3 (20%)
Calcification	7 (47%)	6 (40%)	11 (37%)	10 (33%)	6 (40%)
Asymmetric densities	2 (13%)	4 (27%)	5 (17%)	8 (27%)	4 (27%)
Architectural distortion	2 (13%)	2 (13%)	5 (17%)	6 (20%)	2 (13%)
Screens without cancer, *N*	94	94	79	79	95
Recalled by experts, *N*(%)	21 (22%)	21 (22%)	21 (27%)	21 (27%)	14 (15%)
Other non-cancers, *N*(%)	73 (78%)	73 (78%)	58 (73%)	58 (73%)	81 (85%)

*Screening examinations without cancer were defined to be *appropriate recalls* if the expert panel determined additional imaging was necessary for a finding on the mammogram, because the chance of cancer based on the screening results was sufficiently high to necessitate further evaluation.

Table 2
Characteristics of participating radiologists (*N*=83)

Characteristic	*N*	%
Breast imaging specialist		
Yes	5	6%
No	78	94%
Fellowship training in breast or women's imaging		
Yes	7	8%
No	76	92%
Main practice with academic radiology group		
Yes	7	8%
No	76	92%
Years interpreting mammography		
1-5	15	18%
6-10	12	15%
11-20	36	43%
>20	20	24%
Average days per week working in breast imaging		
≤1	21	25%
2	19	23%
3	18	22%
4	9	11%
5	16	19%
Mammographic examinations interpreted per week		
≤10	2	2%
11-49	17	21%
50-99	31	37%
100-199	20	24%
≥200	13	16%

Table 3

Mean (95% confidence interval) sensitivity and specificity on the test sets compared to performance for same screening examinations in clinical practice, using woman-level recall with cancer status as the gold standard.

	Test set 1: Lower cancer prevalence and less difficult lesions	Test set 2: Lower cancer prevalence and more difficult lesions	Test set 3: Higher cancer prevalence and less difficult lesions	Test set 4: Higher cancer prevalence and more difficult lesions	Test set 5: Lower cancer prevalence and more difficult lesions
Number of radiologists who took the test set	22	24	17	20	48
Sensitivity					
Test set	83% (68%-92%)	75% (61%-85%)	74% (63%-82%)	72% (60%-81%)	74% (62%-84%)
Clinical *	73% (48%-89%)	80% (53%-94%)	83% (68%-92%)	83% (69%-92%)	80%(53%-93%)
p-value	0.23	0.60	0.21	0.12	0.59
Specificity					
Test set	65% (58%-71%)	62% (55%-69%)	69% (61%-77%)	65% (57%-72%)	68% (62%-74%)
Clinical *	55% (44%-66%)	55% (44%-66%)	47% (36%-58%)	47% (36%-58%)	62% (51%-72%)
p-value	0.07	0.19	0.0001	0.001	0.22

* 80% of examinations were interpreted clinically by a radiologist who did not participate in this study.

Table 4

Spearman rank correlation (95% confidence interval) between radiologists' test set and clinical performance measures.

	Test sets 1, 2, 5: Lower cancer prevalence	Test sets 3, 4: Higher cancer prevalence	Test sets 1, 3: Less difficult lesions	Test sets 2, 4, 5: More difficult lesions	Test sets 2, 5: Lower cancer prevalence and more difficult lesions
Sensitivity					
Number of radiologists with at least 10 cancer cases in clinical practice	52	26	27	51	36
Test set cancer cases					
Woman-level, correlation	0.2 1 (-0.06, 0.46)	-0. 13 (-0.49, 0.27)	0.0 4 (-0.34, 0.42)	0.1 3 (-0.16, 0.39)	**0.3 7 (0.04, 0.62)**
Breast-level, correlation	0.21 (-0.07, 0.45)	-0.10 (-0.47, 0.30)	0.09 (-0.30, 0.46)	0.10 (-0.18, 0.37)	0.32 (-0.01, 0.59)
Test set cases recalled by experts					
Woman-level, correlation	**0.3 3 (0.06, 0.55)**	-0. 28 (-0.60, 0.12)	-0. 04 (-0.41, 0.35)	0.2 1 (-0.07, 0.46)	**0.4 6 (0.16, 0.69)**
Breast-level, correlation	0.25 (-0.03, 0.49)	-0.29 (-0.61, 0.11)	0.02 (-0.36, 0.39)	0.09 (-0.19, 0.36)	**0.35 (0.03, 0.61)**
Specificity					
Number of radiologists with at least 100 non-cancer cases in clinical practice	94	37	39	92	72
Test set non-cancer cases, correlation	**0.24 (0.04, 0.42)**	0.05 (-0.28, 0.37)	-0.05 (-0.36, 0.27)	**0.26 (0.06, 0.44)**	**0.28 (0.05, 0.48)**
Test set cases not recalled by experts, correlatio	0.19 (-0.01, 0.38)	0.08 (-0.25, 0.39)	-0.07 (-0.38, 0.25)	**0.23 (0.03, 0.42)**	**0.24 (0.01, 0.45)**