Title

Radiologists' interpretive skills in screening vs. diagnostic mammography: are they related?

Journal

Clinical Imaging, 40(6)

ISSN

0899-7071

Authors

Elmore, Joann G

Cook, Andrea J

Bogart, Andy

et al.

Publication Date

2016-11-01

DOI

10.1016/j.clinimag.2016.06.014

Peer reviewed

HHS Public Access

Author manuscript

Clin Imaging. Author manuscript; available in PMC 2017 November 01.

Published in final edited form as:

Clin Imaging. 2016 ; 40(6): 1096–1103. doi:10.1016/j.clinimag.2016.06.014.

Radiologists' Interpretive Skills in Screening vs. Diagnostic Mammography: Are They Related?

Joann G. Elmore, MD, MPH[a], **Andrea J. Cook, PhD**[b], **Andy Bogart, MS**[c], **Patricia A. Carney, PhD**[d], **Berta Geller, EdD**[e], **Stephen Taplin, MD**[f], **Diana SM Buist, PhD**[b], **Tracy Onega, PhD**[g], **Christoph I. Lee, MD, MSHS**[h], and **Diana L. Miglioretti, PhD**[b,i]

[a]Division of General Internal Medicine, University of Washington, 325 Ninth Ave, Box 359780, Seattle, WA 98104, USA

[b]Group Health Research Institute, Group Health Cooperative, 1730 Minor Ave, Suite 1600, Seattle, WA 98101, USA

[c]RAND Corporation, 1776 Main Street, Santa Monica, CA, USA 90407

[d]Department of Family Medicine, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, Mail Code: FM, Portland, OR 97239, USA

[e]University of Vermont, Burlington, 1 South Prospect Street, UHC, Burlington, VT 05401, USA

[f]Healthcare Delivery Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute National Institutes of Health, 9609 Medical Center Drive, Rockville, MD 20850, USA

[g]Dartmouth Medical School, One Medical Center Drive, HB7937, Lebanon, NH 03756, USA

[h]Department of Radiology, University of Washington School of Medicine; Department of Health Services, University of Washington School of Public Health, 825 Eastlake Ave E, G3-200, Seattle, WA 98109, USA

[i]Division of Biostatistics, Department of Public Health Sciences, University of California Davis School of Medicine, One Shields Ave, Med Sci 1C, Rm 144, Davis, CA 95616, USA

Abstract

Purpose—To determine whether radiologists who perform well in screening also perform well in interpreting diagnostic mammography.

Materials & Methods—We evaluated the accuracy of 468 radiologists interpreting 2,234,947 screening and 196,164 diagnostic mammograms. Adjusting for site, radiologist, and patient characteristics, we identified radiologists with performance in the highest tertile and compared to those with lower performance.

Corresponding author, jelmore@u.washington.edu.

Results—A moderate correlation was noted for radiologists' accuracy when interpreting screening versus their accuracy on diagnostic exams: sensitivity ($r_{spearman}$=0.51, 95% CI: 0.22, 0.80; P=0.0006), specificity ($r_{spearman}$=0.40, 95% CI: 0.30, 0.49; P<0.0001).

Conclusion—Different educational approaches to screening and diagnostic imaging should be considered.

1. INTRODUCTION

Interpretation of diagnostic imaging requires a radiologist to evaluate images tailored to examine a specific abnormality associated with a patient's specific symptoms or abnormalities identified at screening. In contrast, interpretation of screening exams requires evaluation of standard images from a large population of individuals without specific clinical signs or symptoms. The screening interpretive process requires visual pattern recognition when scanning a high volume of images, while diagnostic interpretations require careful analysis of specific abnormalities often using spot compression and magnification views. Diagnostic interpretation also benefits from reports of physical findings made by the patient or physician or additional imaging by other modalities.

Screening and diagnostic examinations involve different patient populations, divergent disease probability, variable numbers and projections of images, and distinct interpretative approaches (e.g., batch reading of screening examinations versus individual reading of diagnostic examinations).[1, 2] Additionally, management recommendations for abnormal assessments usually differ between screening and diagnostic imaging, with suspicious screening examinations often leading to additional diagnostic imaging, and suspicious diagnostic examinations leading to biopsy. These factors suggest that radiologists use different interpretive processes, skills, and thresholds for noting abnormalities when assessing screening versus diagnostic examinations. However, little attention has been paid to this topic.

One previous study of radiologists' interpretations of a screening and diagnostic mammogram test set found little correlation between their accuracy in interpreting screening and diagnostic examinations.[3] This paper describes the imbalance in radiologists' skill development and proficiency between screening and diagnostic interpretation as "expertise disequilibrium."[3] To our knowledge, this topic has not been examined outside of test set conditions. As screening exams continue to be added to the field of radiology (e.g., lung cancer screening, MRI of the breast, screening in high-risk women, etc.) this topic is of increased importance.

In the present study, we examined the correlation between screening and diagnostic interpretive accuracy among individual radiologists using data from real-world settings. We analyzed detailed performance data from the Breast Cancer Surveillance Consortium (BCSC) mammography registries,[4] studying a large group of practicing U.S. radiologists. Data included screening and diagnostic mammogram interpretations accompanied by information on cancer outcomes merged with survey information on radiologist demographics, training, and other characteristics collected from the Factors Associated with Variability of Radiologists (FAVOR) study.[5] Our overarching goal was to evaluate whether

radiologists with the highest performance when interpreting screening mammograms also have the highest performance when interpreting diagnostic mammograms.

2. MATERIAL AND METHODS

2.1 Study Population

Our community-based, multicenter study included radiologists and breast imaging specialists throughout the United States who participate in the BCSC.[6] Seven mammography registries contributed data: San Francisco Bay Area, Colorado, North Carolina, New Mexico, New Hampshire, Vermont, and western Washington. These registries collect patient demographic and clinical information at mammography examinations conducted at a participating facility.[4] This information is linked to regional cancer registries and pathology databases to determine cancer outcomes. Each registry and the Statistical Coordinating Center received IRB approval for either active or passive consenting processes, or for a waiver of consent to enroll participants, link data, and perform analytic studies. All procedures were Health Insurance Portability and Accountability Act-compliant, and all of the registries and the Statistical Coordinating Center have received a Federal Certificate of Confidentiality and other protection for the identities of the patients, physicians, and facilities involved in this research.

Included in these analyses are interpretive performance data from all seven BCSC registry sites on 468 radiologists who interpreted at least one screening and one diagnostic mammogram between January 1, 2001 and December 31, 2006. These dates matched the FAVOR study's survey period to correlate radiologist characteristics with interpretive performance.

2.2 Definitions of Screening and Diagnostic Mammography

We defined screening and diagnostic mammography according to the standard BCSC definitions.[7] A screening mammogram was defined as a bilateral examination indicated by the radiologist or technician as having been conducted for screening purposes; in addition, it had to be performed at least nine months after any prior breast imaging on a woman with no history of breast cancer, reconstruction, or augmentation. We excluded screening mammograms performed on women who self-reported a breast lump or nipple discharge (<2% of screening examinations) because these mammograms may be interpreted differently than routine screening mammograms performed on asymptomatic women.

A diagnostic mammogram was defined as an examination performed to evaluate a breast concern (i.e., a clinical sign or symptom). We excluded short-interval follow-up mammograms and mammograms obtained for further evaluation of a recent screening mammographic examination. These exclusions were based on our overarching goal to assess diagnostic acumen outside of these screening situations because these follow-up examinations are typically obtained to assess findings noted during a screening examination.

Data from a self-administered survey provided information on the individual characteristics and clinical experience of a subset of radiologists.[5, 8] The survey included questions on demographics, clinical training, and previous breast imaging experience. The survey was

mailed to only 277 of the original cohort of radiologists because some radiologists had stopped practicing at a BCSC facility by the date of mailing. Responses are available for 195 (70%) of the subset of 277 radiologists.

2.2 Measurements (Sensitivity and Specificity)

We used the standard BCSC definitions based on the ACR BI-RADS® 4th edition guidelines (which was the standard during the study period) [9] to measure radiologists' interpretive performance.[7] Screening mammograms were classified as positive if they received an initial BI-RADS® assessment of 0 (needs additional imaging), 4 (suspicious abnormality), or 5 (highly suggestive of malignancy). An initial BI-RADS® assessment of 3 (probably benign) with a recommendation for immediate follow-up was also considered positive. Screening mammograms were classified as negative if they received a BI-RADS® assessment of 1 (negative), 2 (benign), or 3 (probably benign) without a recommendation for immediate follow-up.

Diagnostic mammograms were classified as positive if they received a final BI-RADS® assessment, after all diagnostic imaging was performed, of 0, 4, 5, or 3 with a recommendation for biopsy or surgical consult. Diagnostic mammograms were classified as negative if they received a BI-RADS assessment of 1, 2 or 3 without a recommendation for biopsy or surgical consult. The BCSC makes the distinction between BI-RADS 3 assessments with and without a recommendation for biopsy or surgical consult due to the differing clinical recommendations of more invasive biopsy versus non-invasive imaging follow-up.

Consistent with current standards in assessing mammography, breast cancer was defined as ductal carcinoma in situ or invasive breast cancer.[9-11] For screening mammograms, outcome status was defined by breast cancer diagnosis within one year after the mammogram and before the next screening mammogram. For diagnostic mammograms, outcome status was defined by whether a breast cancer diagnosis was recorded in the 30 days prior to or up to 1 year following the diagnostic examination. This was done because the diagnosis may have been dated to the first evidence of breast cancer (potentially prior to the mammogram) for women with signs and symptoms.

We considered a mammogram assessed as positive to be a true positive if a diagnosis of breast cancer was reported within the follow-up period. We considered a mammogram assessed as negative to be a true negative if breast cancer was not reported within the follow-up period.

2.3 Statistical Analyses

We calculated the frequency distributions of self-reported radiologist characteristics and created plots to compare the unadjusted screening and diagnostic interpretations for sensitivity and specificity. Data from radiologists with and without survey data are overlaid in these plots to facilitate comparison of joint distributions of performance measures by survey participation. We also calculated the Spearman correlation between screening and diagnostic performance for sensitivity and specificity of radiologists. Confidence intervals were obtained via bootstrap using 10,000 replicates.[12]

We constructed models providing adjusted performance estimates for each radiologist as described in detail in the Appendix. Briefly, we modeled sensitivity and specificity among radiologists using hierarchical logistic regression adjusted for patient-level characteristics and accommodated the correlation due to multiple mammography records for each radiologist.[13]

The estimated radiologist-specific effects from these models provided the basis for categorizing radiologists. For each performance measure, we calculated the tertiles (33rd and 67th percentiles) of the radiologist-specific effects distributions for screening and diagnostic mammograms separately from the logistic regression model, and then we used these values as thresholds for classification. Our primary interest was identifying the highest performers in sensitivity and specificity for each type of mammogram after adjustment for patient characteristics. We define high performance as at or above the 67th percentile of the performance distribution and low to average performance as anything below the 67th percentile. With these definitions, we categorized the radiologists into one of four groups: 1) high sensitivity and high specificity, 2) high sensitivity and low to average specificity, 3) low to average sensitivity and high specificity, or 4) low to average performance in both sensitivity and specificity. We compared radiologists' characteristics across performance groups for the 195 subjects with survey data, testing differences between the highest performing group and all other groups combined using Fisher's Exact Test.

To obtain an estimate of the adjusted correlation between screening and diagnostic performance, we refit the logistic regression models without assuming the random effects were independent. We used SAS 9.3 software[14] for all analyses, and R version 3.1.1 software for generating plots[15] and estimating Spearman correlations and their confidence bounds.

3. RESULTS

A total of 468 radiologists interpreted 2,234,947 screening mammograms of 1,029,363 women and 196,164 diagnostic mammograms of 154,167 women during the study period. The subgroup of 195 radiologists who completed the survey interpreted 1,231,794 screening mammograms of 692,102 women and 112,128 diagnostic mammograms of 92,661 women. Overall, 76,325 women had both screening and diagnostic mammogram records. Of these, 38,044 women had at least one of each type performed by a radiologist who completed the survey.

The demographic characteristics of the women who received screening and diagnostic examinations and the descriptions of the imaging studies are shown in **Table 1** for the exams interpreted by any of the 468 radiologists, as well as by the subset of 195 radiologists who responded to the FAVOR survey. Women who received diagnostic examinations were younger and less likely to have comparison films available.

Characteristics of the 195 study radiologists who returned the FAVOR survey are shown in **Table 2**. Approximately one in three were female, 9.9% reported a primary affiliation with an academic medical center, and 7.7% reported fellowship training in breast imaging. The

radiologists represent a range of ages, years of mammography interpretation, and percentage of time spent working in breast imaging.

Plots of unadjusted screening versus diagnostic sensitivity and specificity in individual radiologists are shown in **Figures 1** and **2** with black triangles signifying radiologists with survey data and grey circles signifying results from radiologists with no survey data. The unadjusted Spearman correlation between screening and diagnostic sensitivity (Figure 1) was 0.31 (95% CI: 0.16, 0.45) for those with survey data, and 0.15 (95% CI: 0.02, 0.28) for those without survey data. The unadjusted Spearman correlation between screening and diagnostic specificity was 0.35 (95% CI: 0.22, 0.46) for those with survey data and 0.34 (95% CI: 0.22, 0.44) for those without survey data.

Figures 3 and **4** depict the joint distributions of adjusted screening versus diagnostic radiologist-specific effect estimates which assumed these effects were independent. Vertical dashed lines depict the 33rd and 67th percentiles of the screening random effects distribution, and horizontal lines do so for the diagnostic random effects. A dashed diagonal line indicates where estimates would fall if adjusted screening and diagnostic performance were equal. For example, subjects for whom data points are plotted in the upper, rightmost portion of Figure 3 had high screening and high diagnostic performance, relative to their peers, after adjustment for patient-level characteristics. The adjusted correlation estimate obtained from logistic regression models for which radiologist-specific effects were not assumed to be independent was 0.51 (95% CI: 0.22, 0.80, P=0.0006) for sensitivity and 0.40 (95% CI: 0.30, 0.49, P<0.0001) for specificity

Table 3 presents radiologist characteristics across groups defined by adjusted performance in screening and diagnostic sensitivity. The low to average and high performance categories were assigned according to each radiologist's performance relative to a threshold set at the 67th percentile of the distributions of radiologist-specific random effects for screening sensitivity (0.065) and diagnostic sensitivity (0.103). P-values presented in the table are from comparisons of the top performers (high screening and diagnostic sensitivity performance, 4th column) to the combined group of radiologists with low to average performance in screening or diagnostic sensitivity (columns 1, 2, and 3). None of the radiologist characteristics were significantly associated with being among the top performers in sensitivity.

Table 4 presents similar comparisons of radiologist characteristics as Table 3 in groups defined by adjusted specificity performance. Groups are defined as described above, based on the 67th percentile of screening specificity random effect estimates and diagnostic specificity random effect estimates. The smallest proportion of females is in the top performer group for specificity (16.2%; p=0.047), though the difference is small and there was no adjustment for multiple comparisons. No other radiologist characteristics were associated with top performance of specificity. Radiologists in the highest performance level for both screening and diagnostic mammograms were not different from the other radiologists in regard to their age, academic affiliation, fellowship training, years of experience in the field of mammography, percent of time spent in breast imaging, or hours per week working in this clinical field.

4. DISCUSSION

This is the first study, to our knowledge, that correlated radiologists' interpretive performance on screening to their interpretive performance on diagnostic mammography using data derived from clinical practice. We found only a moderate correlation between radiologists' performance interpreting screening examinations and their performance interpreting diagnostic examinations. Additionally, radiologists' clinical experience and practice setting were not associated with being a top performer for both types of examinations.

Wide variability in mammographic interpretation by radiologists has been noted in both screening [10, 16-21] and diagnostic performance.[21-24] A study by Beam et al. (2006), questioned whether interpretive performance in screening is similar to interpretive performance in diagnostic interpretations.[3] Using a test set of mammography cases, Beam et al., found only a moderate correlation between screening and diagnostic interpretations and suggested that proficiency in one did not signify proficiency in the other. This is likely one reason BI-RADS recommends that follow-up and outcome monitoring of individual radiologist performance be done separately for screening and diagnostic examinations.[9] Our findings corroborate this earlier study and current practice recommendations and extend the previous study design by including radiologists' real-world interpretive performance and associated practice and individual characteristics.

While physician characteristics and experience level have been studied as possible predictors of mammographic accuracy, we found no association between physician characteristics and top performance for both types of examinations. Radiologists with fellowship training in breast imaging have previously been found to have both improved cancer detection in screening mammography and higher false-positive rates.[5] Raising the annual volume requirements in the Mammography Quality Standards Act has also been suggested to improve overall mammography quality in the U.S., as radiologists focusing on screening mammography have been associated with fewer false-positive screening examinations.[21, 25] Indeed, increasing the minimum interpretive volume requirements while adding a minimum requirement for diagnostic interpretation has been suggested to reduce the number of false-positive work-ups without hindering cancer detection.[21, 26, 27]

The performance data in this study were drawn from a large community-based cohort of practicing radiologists rather than a test situation. While some study questions can only be addressed using test set cases, we consider real practice to be the best indicator of radiologists' performance. Previous studies have not consistently found a correlation between radiologists' accuracy on test sets and their accuracy in actual clinical practice, while others have found a correlation.[28-30] Another strength of this study is the large size of the BCSC database, which includes data on more than 2.25 million mammography examinations, facilitating reliable sensitivity assessments among a large number of radiologists at the individual level.

Despite the obvious strengths and quantity of BCSC data, study limitations need to be considered. Data from screen-film and digital mammography examinations were restricted

to coincide with collection of radiologist characteristics captured in the FAVOR study survey; thus, our findings should be verified with newer imaging modalities such as digital breast tomosynthesis. We used BI-RADS assessments and clinical recommendations based on the 4th edition of the BI-RADS atlas, as these represented the interpretive standards during the study period. The 5th edition of the BI-RADS atlas categorizes BI-RADS 3 as positive screening and diagnostic assessments regardless of whether or not a biopsy or surgical consult is recommended.

Additionally, some radiologists interpreted relatively few mammograms for patients with breast cancer during the study period. However, our analytic methods account for variations in mammography volume. The average radiologist who interpreted both screening and diagnostic examinations had a predicted radiologist-specific effect close to zero, conditional on the observed patient population. If radiologists had unusually high performance values, even after accounting for patient characteristics, their predicted radiologist-specific effects for both screening and diagnostic examinations were positive and large. This approach allowed us to account for variations in patient populations across radiologists and BCSC sites and to distinguish among radiologists at all levels of performance for both types of examinations.

Interpretive volume and diagnostic performance have a complex multifaceted relationship and were not considered in the current analyses. Our study also did not evaluate physician performance on diagnostic examinations obtained as a result of abnormal screening, and it is possible that an association might have been noted between performance on screening mammography and this narrower type of diagnostic mammography examination.[27]

Our findings have several important applications to radiology resident training as well as continuing professional development of practicing radiologists. Core curriculum in residency should address screening and diagnostic mammography training separately. Interpretive skillsets may be distinct enough that radiologists-in-training need separate minimum training standards for diagnostic mammography distinct from the batch reading for screening mammograms that occurs in actual clinical practice. This is relevant given the *2015 Diagnostic Radiology Milestones*, designed for programs to use in semi-annual reviews of resident performance and reporting to the ACGME.[31] Residents' progress will be examined according to the Milestone Levels 1 to 5, with the Level 5 resident [31] advancing beyond performance targets set for residency and demonstrating "aspirational" goals, which might describe the performance of someone who has been in practice for several years. However, it is also important to create a stronger link between the diagnostic workup of abnormalities found on screening examinations to close the loop on patient processes of care. In the era of "big data" and performance analytics, [32] we need to harness reporting data from imaging examinations to guide development of training programs and feedback systems.

Our findings also have important ramifications with regard to clinical practice improvement efforts. Our study corroborates recommendations that practices separate performance benchmarks for radiologists' screening and diagnostic interpretive performance.[11] Additionally, given the weak-to-moderate correlation between screening and diagnostic

mammography, our study suggests that accuracy would not be sacrificed with the complete separation of screening and diagnostic breast imaging services among interpreting radiologists.

5. CONCLUSIONS

In summary, we found a statistically significant, but weak-to-moderate correlation between radiologists' interpretive performance in screening mammography and their interpretive performance in diagnostic mammography. Our findings have several important ramifications with regard to education and practice improvement efforts. Further research regarding the associations of interpretive accuracy, volume of interpretations, and types of imaging studies might elucidate best approaches to clinical practices, audit feedback systems, and educational training programs.

ACKNOWLEDGMENTS

We thank the Breast Cancer Surveillance Consortium (BCSC) investigators, participating mammography facilities, radiologists and women undergoing mammography 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/.

Supported by public health service grant R01 HS-010591 from the Agency for Healthcare Research and Quality, National Cancer Institute grants R01 CA-107623 and K05 CA-104699. Data collection and statistical analysis for this work was supported by the NCI-funded BCSC (HHSN261201100031C).

The collection of cancer incidence 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://breastscreening.cancer.gov/work/acknowledgement.html.

Electronic Appendix

Statistical Methods

We constructed models providing adjusted performance estimates for each radiologist. We modeled sensitivity and specificity among all 468 eligible radiologists using hierarchical logistic regression adjusted for patient-level characteristics and accommodated the correlation due to multiple mammography records for each radiologist.[13] For sensitivity, we modeled the probability of a positive assessment among mammograms obtained for women who were diagnosed with breast cancer. For specificity, we modeled the probability of a negative assessment among mammograms for women who were not diagnosed with breast cancer within one year of follow-up.

Each model included fixed effects to indicate a screening versus diagnostic mammogram as well as patient-level characteristics of age, family history of breast cancer, time since previous mammogram, and BCSC registry site. Each model also included two radiologist-specific random effects: one for screening and one for diagnostic examinations. Radiologist-specific effects were assumed to be independent and normally distributed around zero, with positive estimates indicating above-average performance and negative estimates indicating below-average performance, after adjustment for model covariates. We assumed the two random effects were independent so that the model would not artificially impose a

correlation between types of examination. On average, radiologists who perform well/poorly on screening also perform well/poorly on diagnostic mammography. Therefore, if we had imposed a correlation structure between screening and diagnostic random effects, this would have masked radiologists who performed differently by type of mammogram since their performance is different than the average radiologist who has similar performance by type of mammogram (e.g. random effects would have been shrunk toward the average performance between screening and diagnostic since the model would have assumed they should be positively correlated). The purpose of the analysis was to take into account patient characteristics and number of mammograms when defining performance while still allowing radiologists to have different performance by type of exam. We then obtained random effect estimates using the empirical Bayes approach.[33] We plotted these adjusted estimates to compare screening versus diagnostic performance, controlling for patient characteristics.

The estimated radiologist-specific effects from these models provided the basis for categorizing radiologists. For each performance measure, we calculated the tertiles (33^{rd} and 67^{th} percentiles) of the radiologist-specific effects distributions for screening and diagnostic mammograms separately from the logistic regression model, and then we used these values as thresholds for classification.

Our primary interest was identifying the highest performers in sensitivity and specificity for each type of mammogram after adjustment for patient characteristics. We define high performance as at or above the 67^{th} percentile of the performance distribution and low to average performance as anything below the 67^{th} percentile. With these definitions, we categorized the radiologists into one of four groups: 1) high sensitivity and high specificity; 2) high sensitivity and low to average specificity; 3) low to average sensitivity and high specificity; or 4) low to average performance in both sensitivity and specificity. We compared radiologist characteristics across performance groups for the 195 subjects with survey data, testing differences between the highest performing group to all other groups combined using Fisher's Exact Test.

To obtain an estimate of the adjusted correlation between screening and diagnostic performance, we refit the logistic regression models without assuming the random effects were independent. We used SAS 9.3 software [14] for all analyses, and R version 3.1.1 software for generating plots [15] and estimating Spearman correlations and their confidence bounds.

REFERENCES

1. Henderson LM, Miglioretti DL, Kerlikowske K, Wernli KJ, Sprague BL, Lehman CD. Breast cancer characteristics associated with digital versus screen-film mammography for screen-detected and interval cancers. AJR American journal of roentgenology. 2015; 205(3):676–84. [PubMed: 26295657]
2. Harris; Jay, R.; Lippman, ME.; Osborne, K.; Morrow, M. Diseases of the Breast. Lippincott Williams & Wilkins; 2012.
3. Beam CA, Conant EF, Sickles EA. Correlation of radiologist rank as a measure of skill in screening and diagnostic interpretation of mammograms. Radiology. 2006; 238(2):446–53. [PubMed: 16436811]

4. National Cancer Institute. Breast Cancer Surveillance Consortium [Web]. National Cancer Institute: National Institutes of Health; 2015. [updated 7/6/2015]. Available from: http://breastscreening.cancer.gov/ [2015 Dec 10]
5. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, et al. Variability in Interpretive Performance of Screening Mammography and Radiologist Characteristics Associated with Accuracy. Radiology. 2009; 253:641–51. [PubMed: 19864507]
6. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. AJR American journal of roentgenology. 1997; 169(4):1001–8. [PubMed: 9308451]
7. Breast Cancer Surveillance Consortium. BCSC Standard Definitions. National Cancer Institute; 2009. [updated September 16, 2009]. Available from: http://breastscreening.cancer.gov/data/bcsc_data_definitions.pdf [Dec 10, 2015]
8. Breast Cancer Surveillance Consortium. National Survey of Mammography Practices. National Cancer Institute; 2006. [updated Sep 9, 2009]. Available from: http://breastscreening.cancer.gov/collaborations/favor_ii_mammography_practice_survey.pdf [Dec 9, 2015]
9. American College of Radiology. Breast Imaging Atlas. 4th ed.. American College of Radiology; Renton, VA: 2013. ACR Breast Imaging Reporting and Data System (BI-RADS)..
10. Rosenberg RD, Yankaskas BC, Abraham LA, Sickles EA, Lehman CD, Geller BM, et al. Performance benchmarks for screening mammography. Radiology. 2006; 241(1):55–66. [PubMed: 16990671]
11. Sickles EA, Miglioretti DL, Ballard-Barbash R, Geller BM, Leung JW, Rosenberg RD, et al. Performance benchmarks for diagnostic mammography. Radiology. 2005; 235(3):775–90. [PubMed: 15914475]
12. Efron B. Better Bootstrap Confidence Intervals. Journal of the American Statistical Association. 1987; 82(397):171–85.
13. Pinheiro JC, Bates DM. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. Journal of Computational and Graphical Statistics. 1995; 4(1):12–35.
14. SAS Institute Inc.. SAS Software 9.3 for Windows. Cary; North Carolina: 2008.
15. R Development Core Team. R: A language and environment for statistical computing. Vienna: 2014. Available from: www.r-project.org [2014 Dec 09]
16. Elmore JG, Miglioretti DL, Reisch LM, Barton MB, Kreuter W, Christiansen CL, et al. Screening mammograms by community radiologists: variability in false-positive rates. Journal of the National Cancer Institute. 2002; 94(18):1373–80. [PubMed: 12237283]
17. Elmore JG, Nakano CY, Koepsell TD, Desnick LM, D'Orsi CJ, Ransohoff DF. International variation in screening mammography interpretations in community-based programs. Journal of the National Cancer Institute. 2003; 95(18):1384–93. [PubMed: 13130114]
18. Beam CA, Conant EF, Sickles EA. Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. Journal of the National Cancer Institute. 2003; 95(4):282–90. [PubMed: 12591984]
19. Barlow WE, Chi C, Carney PA, Taplin SH, D'Orsi C, Cutter G, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. Journal of the National Cancer Institute. 2004; 96(24):1840–50. [PubMed: 15601640]
20. Tan A, Freeman DH Jr. Goodwin JS, Freeman JL. Variation in false-positive rates of mammography reading among 1067 radiologists: a population-based assessment. Breast cancer research and treatment. 2006; 100(3):309–18. [PubMed: 16819566]
21. Assessing and Improving the Interpretation of Breast Images: Workshop Summary. National Academies Press; Washington DC: 2015.
22. Miglioretti DL, Smith-Bindman R, Abraham L, Brenner RJ, Carney PA, Bowles EJ, et al. Radiologist characteristics associated with interpretive performance of diagnostic mammography. Journal of the National Cancer Institute. 2007; 99(24):1854–63. [PubMed: 18073379]
23. Jackson SL, Taplin SH, Sickles EA, Abraham L, Barlow WE, Carney PA, et al. Variability of interpretive accuracy among diagnostic mammography facilities. Journal of the National Cancer Institute. 2009; 101(11):814–27. [PubMed: 19470953]

24. Aiello Bowles EJ, Miglioretti DL, Sickles EA, Abraham L, Carney PA, Yankaskas BC, et al. Accuracy of short-interval follow-up mammograms by patient and radiologist characteristics. AJR American journal of roentgenology. 2008; 190(5):1200–8. [PubMed: 18430832]
25. Smith-Bindman R, Chu P, Miglioretti DL, Quale C, Rosenberg RD, Cutter G, et al. Physician predictors of mammographic accuracy. Journal of the National Cancer Institute. 2005; 97(5):358–67. [PubMed: 15741572]
26. Buist DS, Anderson ML, Haneuse SJ, Sickles EA, Smith RA, Carney PA, et al. Influence of annual interpretive volume on screening mammography performance in the United States. Radiology. 2011; 259(1):72–84. [PubMed: 21343539]
27. Buist DS, Anderson ML, Smith RA, Carney PA, Miglioretti DL, Monsees BS, et al. Effect of radiologists' diagnostic work-up volume on interpretive performance. Radiology. 2014; 273(2): 351–64. [PubMed: 24960110]
28. Soh BP, Lee W, McEntee MF, Kench PL, Reed WM, Heard R, et al. Screening Mammography: Test Set Data Can Reasonably Describe Actual Clinical Reporting. Radiology. 2013; 268(1):46–53. [PubMed: 23481165]
29. Egglin TK, Feinstein AR. Context bias. A problem in diagnostic radiology. JAMA. 1996; 276(21): 1752–5. [PubMed: 8940325]
30. Rutter CM, Taplin S. Assessing mammographers' accuracy. A comparison of clinical and test performance. Journal of clinical epidemiology. 2000; 53(5):443–50. [PubMed: 10812315]
31. Vydareny K, Amis E, Becker G, Borgstede J, Bulas D, Collins J, et al. The Diagnostic Radiology Milestone Project. Accreditation Council for Graduate Medical Education and American Board of Radiology. Jul.2015 Report No.
32. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Affairs. 2014; 33(7):1163–70. [PubMed: 25006142]
33. Harville D. Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. The Annals of Statistics. 1976; 4(2):384–95.

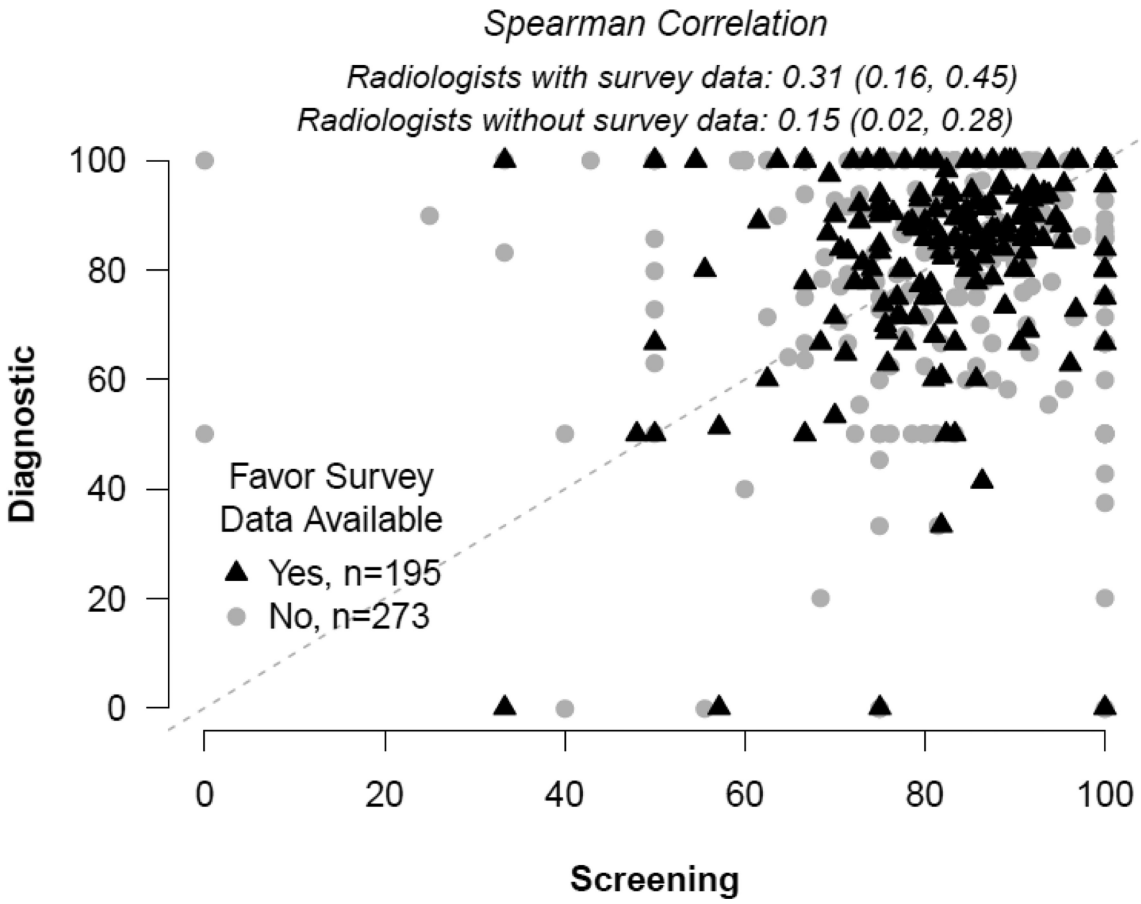

Figure 1.
Sensitivity.

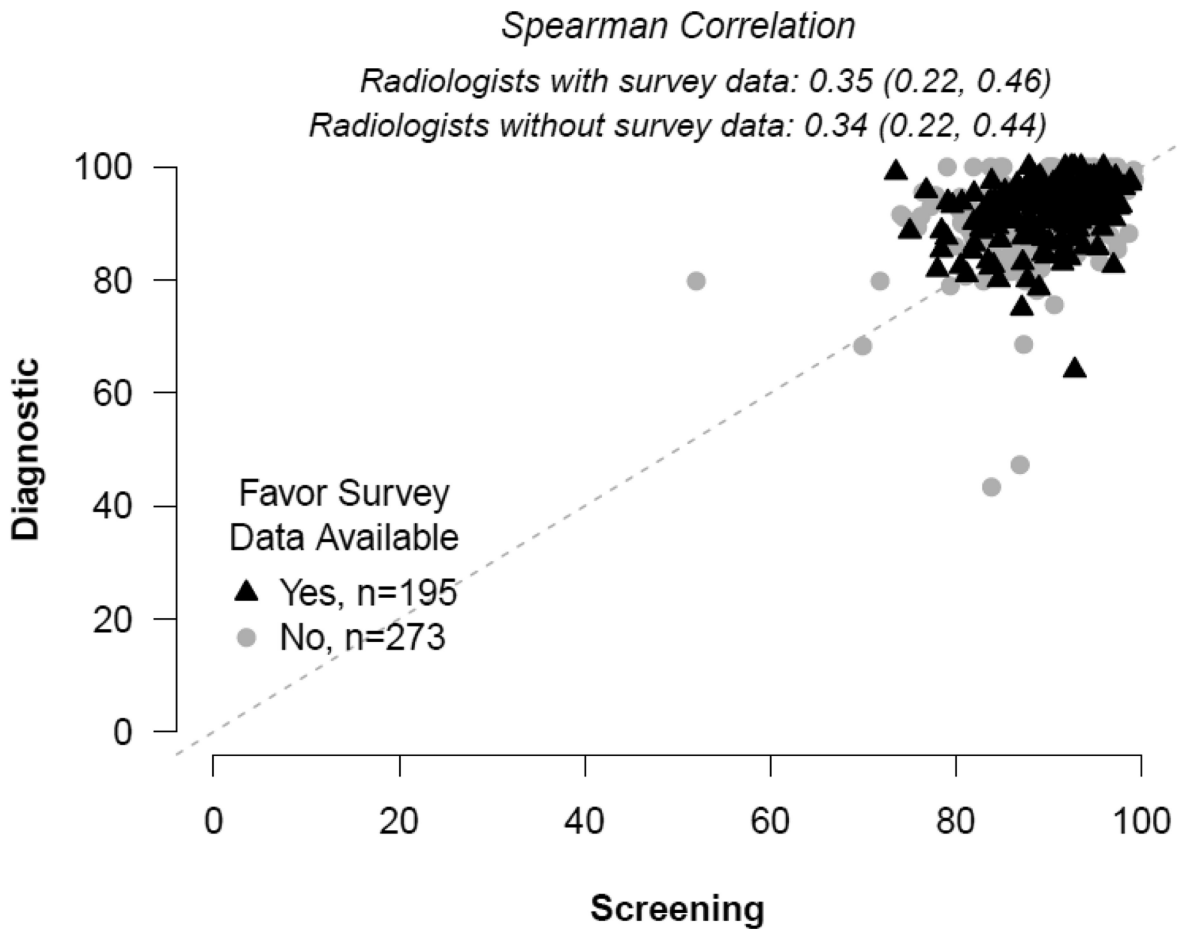

Figure 2.
Specificity.

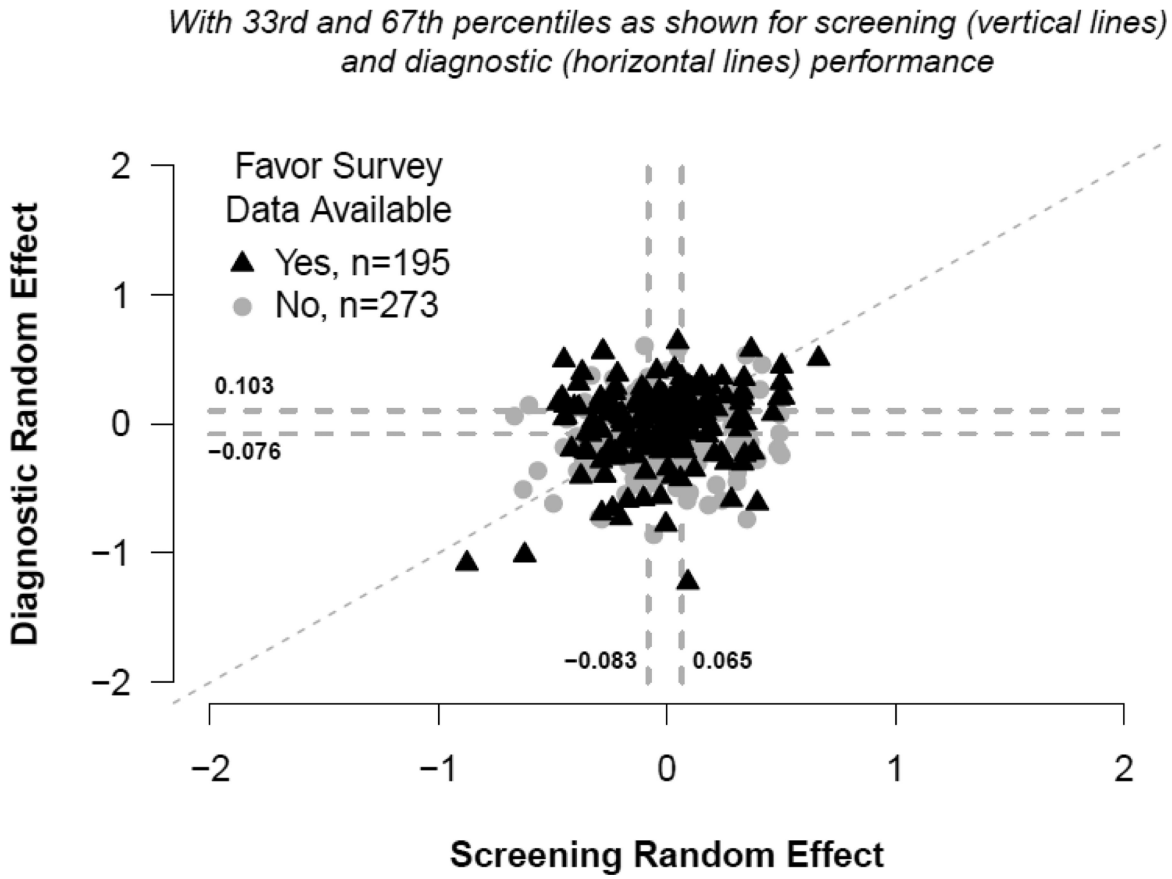

Figure 3.
Sensitivity random effects.

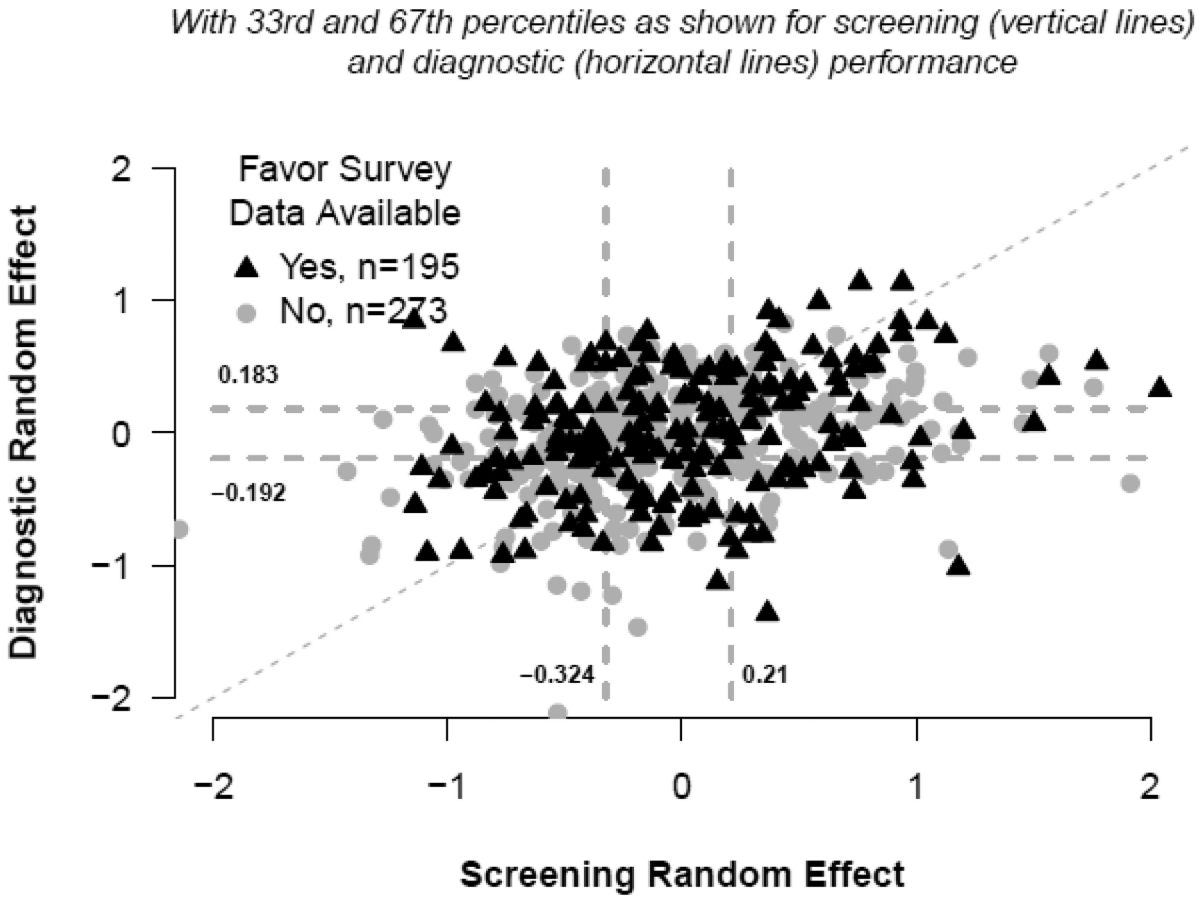

Figure 4.
Specificity random effects.

Table 1

Characteristics of mammograms interpreted by BCSC radiologists.[1]

	Interpreted by any of 468 BCSC Radiologists		Interpreted by one of 195 BCSC Radiologists who completed the FAVOR survey	
	Screening exams	Diagnostic exams	Screening exams	Diagnostic exams
Patient characteristics at exam time	**n = 2,234,947**	**n = 196,164**	**n = 1,231,794**	**n = 112,128**
Age at exam, %				
< 40	3.6	18.0	3.3	17.3
40-49	28.8	29.3	28.1	29.1
50-59	30.9	24.4	31.0	24.8
60-69	19.5	14.6	19.9	14.9
70+	17.3	13.6	17.7	13.9
BI-RADS® breast density, %				
Almost entirely fat (<25% fibroglandular)	6.7	4.8	7.7	5.6
Scattered fibroglandular tissue (25%-50%)	36.3	26.8	38.1	27.2
Heterogeneously dense (50%-75%)	32.2	29.3	33.4	30.2
Extremely dense (>75%)	6.6	9.3	7.0	10.0
Unknown	18.1	29.7	13.8	26.9
Prior mammography, %				
< 3 years	79.0	69.7	78.3	69.9
3 years or more	7.8	7.2	7.9	7.1
No prior mammogram	4.4	10.3	4.2	9.4
Unknown	8.8	12.8	9.6	13.6
Comparison film available at time of exam, %				
No	9.7	28.4	10.3	28.1
Yes	77.8	54.8	77.5	56.0
Unknown	12.5	16.8	12.1	15.8
Family history of breast cancer, %				
No	81.8	78.6	81.2	78.4
Yes	14.2	15.5	14.8	16.3
Unknown	4.0	5.9	4.0	5.3
Current hormone therapy use, %				
No	72.5	76.8	71.7	75.9
Yes	18.1	12.3	17.3	11.6
Unknown	9.4	10.9	11.1	12.6
Self-reported breast symptoms, %				
None	91.3	36.6	91.1	38.9
Pain	1.6	8.5	1.8	8.4
Other not including pain	1.9	8.8	1.6	8.1
Nipple discharge	0.0	2.5	0.0	2.4

	Interpreted by any of 468 BCSC Radiologists		Interpreted by one of 195 BCSC Radiologists who completed the FAVOR survey	
	Screening exams	Diagnostic exams	Screening exams	Diagnostic exams
Patient characteristics at exam time	**n = 2,234,947**	**n = 196,164**	**n = 1,231,794**	**n = 112,128**
Lump	0.0	35.4	0.0	33.3
Unknown	5.3	8.2	5.4	9.0

[1]Percentages shown here are calculated among all screening mammograms and all diagnostic mammograms read by the 468 Breast Cancer Surveillance Consortium (BCSC) radiologists (first two columns), or the subset of 195 BCSC radiologists who responded to the Factors Associated with Variability of Radiologists (FAVOR) survey (second two columns), between January 2001 and December 2006. The cohort of 468 BCSC radiologists reviewed screening mammograms from 1,029,363 individual women and diagnostic mammograms from 154,167 women. The subset of 195 radiologists who responded to the FAVOR survey reviewed screening mammograms from 692,102 individual women, and diagnostic mammograms from 92,661 women. Women who had both screening and diagnostic mammograms contribute information to both the screening and diagnostic cohort summaries above. 76,325 women had at least one screening and one diagnostic mammogram interpreted by a BCSC radiologist. Of these, 38,044 women had at least one screening and at least one diagnostic mammogram interpreted by FAVOR study radiologists.

Table 2

Radiologist Characteristics among those who completed the FAVOR survey[1]

Radiologist Characteristic	N=195 (100%)
Gender, n (%)	
Female	59 (30.3%)
Male	136 (69.7%)
Age group, n (%)	
32 – 44	47 (24.1%)
45 – 54	71 (36.4%)
55 or older	77 (39.5%)
Academic Medical Center Affiliation, n (%)	
None	161 (83.9%)
Adjunct	12 (6.25%)
Primary	19 (9.9%)
Fellowship training, n (%)	
Not-fellowship trained	180 (92.3%)
Fellowship trained	15 (7.7%)
Years of mammography experience, n (%)	
0 – 9	49 (25.5%)
10 – 19	63 (32.8%)
20 or more	80 (41.7%)
Percent of time spent in breast imaging, n (%)	
0 – 19%	46 (24.5%)
20 – 39%	53 (28.2%)
40 – 79%	29 (15.4%)
80 – 100%	60 (31.9%)
Hours per week spent in breast imaging, n (%)	
0 – 8	43 (23.2%)
9 – 16	63 (34.1%)
17 – 32	32 (17.3%)
33 or more	47 (25.4%)

[1]Radiologists who interpreted a minimum of 1 screening and 1 diagnostic exam and also completed the Factors Associated with Variability of Radiologists (FAVOR) survey. Percentages shown correspond to non-missing observations. Missing data on survey questions: Academic medical center affiliation n=3; Years of mammography experience n=3; Percentage of time spent in breast imaging n=7; Hours per week spent in breast imaging n=10.

Table 3

Radiologist characteristics by categories of screening and diagnostic sensitivity.[1]

	Low to average performance in screening or diagnostic performance			Top performers	
Adjusted Screening Performance:	**Low to Average**	**High**	**Low to Average**	**High**	**P value comparing highest performers to all others combined**
Adjusted Diagnostic Performance:	**Low to Average**	**Low to Average**	**High**	**High**	
	N=87	**N=35**	**N=44**	**N=29**	
Gender, n (%)					0.189
Female	20 (23.0%)	11 (31.4%)	16 (36.4%)	12 (41.4%)	
Male	67 (77.0%)	24 (68.6%)	28 (63.6%)	17 (58.6%)	
Age group, n (%)					0.136
32 – 44	17 (19.5%)	9 (25.7%)	11 (25.0%)	10 (34.5%)	
45 – 54	30 (34.5%)	11 (31.4%)	18 (40.9%)	12 (41.4%)	
55 or older	40 (46.0%)	15 (42.9%)	15 (34.1%)	7 (24.1%)	
Academic Medical Center Affiliation, n (%)					0.261
None	73 (84.9%)	31 (91.2%)	35 (81.4%)	22 (75.9%)	
Adjunct	6 (6.98%)	0 (0%)	4 (9.30%)	2 (6.90%)	
Primary	7 (8.14%)	3 (8.82%)	4 (9.30%)	5 (17.2%)	
Fellowship training, n (%)					0.247
Not-fellowship trained	82 (94.3%)	34 (97.1%)	39 (88.6%)	25 (86.2%)	
Fellowship trained	5 (5.75%)	1 (2.86%)	5 (11.4%)	4 (13.8%)	
Years of mammography experience, n (%)					0.314
0 – 9	21 (24.4%)	8 (23.5%)	13 (30.2%)	7 (24.1%)	
10 – 19	24 (27.9%)	10 (29.4%)	16 (37.2%)	13 (44.8%)	
20 or more	41 (47.7%)	16 (47.1%)	14 (32.6%)	9 (31%)	
Percent of time spent in breast imaging, n (%)					0.558
0 – 19%	23 (27.4%)	10 (30.3%)	8 (19.0%)	5 (17.2%)	
20 – 39%	26 (31.0%)	5 (15.2%)	11 (26.2%)	11 (37.9%)	
40 – 79%	10 (11.9%)	7 (21.2%)	7 (16.7%)	5 (17.2%)	
80 – 100%	25 (29.8%)	11 (33.3%)	16 (38.1%)	8 (27.6%)	
Hours per week spent in breast imaging, n (%)					0.204
0 – 8	20 (24.1%)	13 (39.4%)	5 (12.5%)	5 (17.2%)	
9 – 16	25 (30.1%)	9 (27.3%)	15 (37.5%)	14 (48.3%)	
17 – 32	15 (18.1%)	5 (15.2%)	6 (15.0%)	6 (20.7%)	
33 or more	23 (27.7%)	6 (18.2%)	14 (35.0%)	4 (13.8%)	

[1] Percentages shown correspond to non-missing observations. P-values compare the 29 top performers (4^{th} column) to the 166 others combined (columns 1-3) using Fisher's Exact Test. Low, Average, and High performance categories were assigned relative to the 33^{rd} and 67^{th} tertiles of the distribution of radiologist-specific random effects for screening sensitivity (−0.083, 0.065) and diagnostic sensitivity (−0.076, 0.103). Missing data on survey questions: Academic medical center affiliation n=3; Years of mammography experience n=3; Percentage of time spent in breast imaging n=7; Hours per week spent in breast imaging n=10.

Table 4

Radiologist characteristics by categories of screening and diagnostic specificity.[1]

	Low to average performance in screening or diagnostic performance			Top performers	
Adjusted Screening Performance:	**Low to Average**	**High**	**Low to Average**	**High**	**P value comparing highest performers to all others combined**
Adjusted Diagnostic Performance:	**Low to Average**	**Low to Average**	**High**	**High**	
	N=90	**N=31**	**N=37**	**N=37**	
Gender, n (%)					0.047
Female	34 (37.8%)	7 (22.6%)	12 (32.4%)	6 (16.2%)	
Male	56 (62.2%)	24 (77.4%)	25 (67.6%)	31 (83.8%)	
Age group, n (%)					0.144
32 – 44	26 (28.9%)	4 (12.9%)	11 (29.7%)	6 (16.2%)	
45 – 54	36 (40%)	10 (32.3%)	14 (37.8%)	11 (29.7%)	
55 or older	28 (31.1%)	17 (54.8%)	12 (32.4%)	20 (54.1%)	
Academic Medical Center Affiliation, n (%)					0.325
None	73 (82%)	28 (90.3%)	30 (83.3%)	30 (83.3%)	
Adjunct	5 (5.62%)	0 (0%)	3 (8.33%)	4 (11.1%)	
Primary	11 (12.4%)	3 (9.68%)	3 (8.33%)	2 (5.56%)	
Fellowship training, n (%)					0.311
Not-fellowship trained	81 (90%)	29 (93.5%)	34 (91.9%)	36 (97.3%)	
Fellowship trained	9 (10%)	2 (6.45%)	3 (8.11%)	1 (2.7%)	
Years of mammography experience, n (%)					0.106
0 – 9	28 (31.1%)	5 (16.1%)	11 (31.4%)	5 (13.9%)	
10 – 19	32 (35.6%)	8 (25.8%)	12 (34.3%)	11 (30.6%)	
20 or more	30 (33.3%)	18 (58.1%)	12 (34.3%)	20 (55.6%)	
Percent of time spent in breast imaging, n (%)					0.124
0 – 19%	19 (21.8%)	6 (19.4%)	12 (32.4%)	9 (27.3%)	
20 – 39%	23 (26.4%)	8 (25.8%)	8 (21.6%)	14 (42.4%)	
40 – 79%	15 (17.2%)	4 (12.9%)	8 (21.6%)	2 (6.06%)	
80 – 100%	30 (34.5%)	13 (41.9%)	9 (24.3%)	8 (24.2%)	
Hours per week spent in breast imaging, n (%)					0.432
0 – 8	16 (18.8%)	4 (13.3%)	13 (35.1%)	10 (30.3%)	
9 – 16	27 (31.8%)	11 (36.7%)	13 (35.1%)	12 (36.4%)	
17 – 32	17 (20%)	5 (16.7%)	4 (10.8%)	6 (18.2%)	
33 or more	25 (29.4%)	10 (33.3%)	7 (18.9%)	5 (15.2%)	

[1] Percentages shown correspond to non-missing observations. P-values compare the 37 top performers (4^{th} column) to the 158 others combined (columns 1-3) using Fisher's Exact Test. Low, Average, and High performance categories were assigned relative to the 33^{rd} and 67^{th} tertiles of the distribution of radiologist-specific random effects for screening specificity (−0.324, 0.210) and diagnostic specificity (−0.192, 0.183). Missing data on survey questions: Academic medical center affiliation n=3; Years of mammography experience n=3; Percentage of time spent in breast imaging n=7; Hours per week spent in breast imaging n=10.