Clinical Imaging 40 (2016) 1096–1103

Contents lists available at ScienceDirect

Clinical Imaging

journal homepage: http://www.clinicalimaging.org

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

Joann G. Elmore [a,*], Andrea J. Cook [b], Andy Bogart [c], Patricia A. Carney [d], Berta M. Geller [e], Stephen H. Taplin [f], Diana S.M. Buist [b], Tracy Onega [g], Christoph I. Lee [h,i], Diana L. Miglioretti [b,j]

[a] Division of General Internal Medicine, University of Washington, 325 Ninth Avenue, Box 359780, Seattle, WA, 98104, USA
[b] Group Health Research Institute, Group Health Cooperative, 1730 Minor Avenue, Suite 1600, Seattle, WA, 98101, USA
[c] RAND Corporation, 1776 Main Street, Santa Monica, CA, 90407, USA
[d] Department of Family Medicine, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Mail Code: FM, Portland, OR, 97239, USA
[e] University of Vermont, 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, 825 Eastlake Avenue E, G3-200, Seattle, WA, 98109, USA
[i] Department of Health Services, University of Washington School of Public Health, 1959 NE Pacific St., Box 357660, Seattle, WA, 98195, USA
[j] Division of Biostatistics, Department of Public Health Sciences, University of California Davis School of Medicine, One Shields Avenue, Med Sci 1C, Room 144, Davis, CA, 95616, USA

ARTICLE INFO

*Article history:*
Received 4 March 2016
Received in revised form 3 June 2016
Accepted 29 June 2016
Available online xxxx

*Keywords:*
Screening mammography
Diagnostic mammography
Interpretive performance
Sensitivity
Specificity

ABSTRACT

**Purpose:** This study aims to determine whether radiologists who perform well in screening also perform well in interpreting diagnostic mammography.
**Materials and 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.
**Results:** A moderate correlation was noted for radiologists' accuracy when interpreting screening versus their accuracy on diagnostic examinations: sensitivity ($r_{spearman}$=0.51, 95% CI: 0.22, 0.80; $P$=.0006) and specificity ($r_{spearman}$=0.40, 95% CI: 0.30, 0.49; $P$<.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 examinations 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 examinations 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.

* Corresponding author. Division of General Internal Medicine, University of Washington, 325 Ninth Avenue, Box 359780, Seattle, WA, 98104, USA.
*E-mail address:* jelmore@u.washington.edu (J.G. Elmore).

http://dx.doi.org/10.1016/j.clinimag.2016.06.014
0899-7071/

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. Materials 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 an institutional review board 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 9 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.

### 2.3. 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 noninvasive 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 1 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.

### 2.4. 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