# Radiologist Agreement for Mammographic Recall by Case Difficulty and Finding Type

E-ONLY

*Tracy Onega, PhD[a], Megan Smith, MS[b], Diana L. Miglioretti, PhD[b,c], Patricia A. Carney, PhD[d], Berta A. Geller, EdD[e], Karla Kerlikowske, MD[f], Diana S. M. Buist, PhD[b], Robert D. Rosenberg, MD[g], Robert A. Smith, PhD[h], Edward A. Sickles, MD[f], Sebastien Haneuse, PhD[i], Melissa L. Anderson, MS[b], Bonnie Yankaskas, PhD[j]*

## Abstract

**Purpose:** The aim of this study was to assess agreement of mammographic interpretations by community radiologists with consensus interpretations of an expert radiology panel to inform approaches that improve mammographic performance.

**Methods:** From 6 mammographic registries, 119 community-based radiologists were recruited to assess 1 of 4 randomly assigned test sets of 109 screening mammograms with comparison studies for no recall or recall, giving the most significant finding type (mass, calcifications, asymmetric density, or architectural distortion) and location. The mean proportion of agreement with an expert radiology panel was calculated by cancer status, finding type, and difficulty level of identifying the finding at the patient, breast, and lesion level. Concordance in finding type between study radiologists and the expert panel was also examined. For each finding type, the proportion of unnecessary recalls, defined as study radiologist recalls that were not expert panel recalls, was determined.

**Results:** Recall agreement was 100% for masses and for examinations with obvious findings in both cancer and noncancer cases. Among cancer cases, recall agreement was lower for lesions that were subtle (50%) or asymmetric (60%). Subtle noncancer findings and benign calcifications showed 33% agreement for recall. Agreement for finding responsible for recall was low, especially for architectural distortions (43%) and asymmetric densities (40%). Most unnecessary recalls (51%) were asymmetric densities.

**Conclusions:** Agreement in mammographic interpretation was low for asymmetric densities and architectural distortions. Training focused on these interpretations could improve the accuracy of mammography and reduce unnecessary recalls.

**Key Words:** Mammography, breast cancer, screening, agreement

*J Am Coll Radiol 2012;9:788-794.* 

[a]Department of Community and Family Medicine, Dartmouth Medical School, Hanover, New Hampshire.

[b]Group Health Research Institute, Seattle, Washington.

[c]Department of Biostatistics, University of Washington, Seattle, Washington.

[d]Departments of Family Medicine and Public Health and Preventive Medicine, Oregon Health & Science University, Portland, Oregon.

[e]University of Vermont, Burlington, Vermont.

[f]Departments of Medicine and Epidemiology/Biostatistics and Radiology, University of California, San Francisco, San Francisco, California.

[g]Department of Radiology, University of New Mexico, Albuquerque, New Mexico.

[h]American Cancer Society, Atlanta, Georgia.

[i]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts.

[j]Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Corresponding author and reprints: Tracy Onega, PhD, Dartmouth Medical School, HB 7927, Rubin 8 - DHMC, One Medical Center Drive, Lebanon, NH 03756; e-mail: tracy.l.onega@dartmouth.edu.

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, and SIRSG-06-290-04), the Breast Cancer Stamp Fund, and the National Cancer Institute Breast Cancer Surveillance Consortium (U01CA63740, U01CA86076, U01CA86082, U01CA70013, U01 CA69976, U01CA63731, U01CA70040, and HHSN261201100031C). The collection of cancer and vital status data used in this study was supported in part by several state public health departments and cancer registries throughout the United States. For a full description of these sources, 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 or the National Institutes of Health.


1546-1440/16/$36.00 ■ http://dx.doi.org/10.1016/j.jacr.2016.09.035
Reprinted from: J Am Coll Radiol 2012; 9: 788-794.

## INTRODUCTION

The effectiveness of screening mammography depends on the accuracy of image interpretation. However, the interpretive performance of US radiologists in community practice varies widely for screening mammography [1]. Notable interobserver variability is seen, even when interpreting the same mammographic images [2-7]. Most studies comparing radiologists' interpretations of screening mammogram test sets have focused on agreement of the BI-RADS® assessment [4,6-9]. Berg et al [5] found considerable variability in the assessment of 103 screening mammograms interpreted by 5 experienced radiologists. Ciatto et al [6] and Kerlikowske et al [4] reported moderate interobserver agreement in studies of 100 examinations interpreted by 12 radiologists and 2,616 mammograms interpreted by 2 radiologists, respectively. Elmore et al [3], Lazarus et al [9], Venta et al [10], and Skaane et al [11] reported substantial interobserver variation in studies of 50 to 1,147 mammograms, with 5 to 10 interpreting radiologists.

Agreement of mammographic interpretations by finding type (calcification, mass, asymmetry, or architectural distortion) and case assessment difficulty has not been well characterized. Berg et al [5] examined the agreement of 5 radiologists experienced in mammography for lesion classification and defining features and found high interobserver agreement ($\kappa = 0.75$) for lesion classification but not for features and management. However, this prior study was limited by a small sample size and the lack of an expert consensus standard. Determining which mammographic findings have poor agreement with a well-defined standard could identify areas of focus for interventions aimed at improving interpretive accuracy.

The purpose of this study was to assess agreement between mammographic interpretations of community radiologists and a gold standard, defined as consensus interpretations of an expert panel. We hypothesized that certain findings or features would be mainly responsible for lack of agreement. These would be potential areas to target for improving mammographic performance. Our study included the largest number of participating radiologists to date for a mammographic interpretation agreement study. We analyzed variability in the interpretation of screening mammograms by finding and by difficulty of identification, examining agreement of 119 community radiologists with a consensus-based standard developed by 3 mammography experts.

## METHODS

### Study Population

This study was conducted within 6 breast screening registries of the National Cancer Institute–funded Breast Cancer Surveillance Consortium (BCSC): the Carolina Mammography Registry, the Group Health Surveillance Project in Washington State, the New Hampshire Mammography Network, the San Francisco Mammography Registry, the New Mexico Mammography Project, and the Vermont Breast Cancer Surveillance System [12,13,14]. Radiologists interpreting mammograms at BCSC facilities between January 2005 and December 2006 were invited to participate. In addition, we invited 103 radiologists from outside the BCSC in Oregon, Washington, North Carolina, San Francisco, and New Mexico. Of the 469 invited radiologists, 148 (31.6%) provided informed consent, and 119 (80.4%) of these completed the study. Radiologists received up to 8 *AMA PRA Category 1 Credits*™ for interpreting a test set. Each site received institutional review board approval for study activities. Active or passive consent or waivers are obtained at each site from women undergoing mammography. The identities of women, physicians, and facilities are protected by a federal certificate of confidentiality and other protections. All procedures comply with HIPAA.

### Test Set Development

Test set development is described in detail elsewhere [14]. Briefly, we developed 4 test sets differing by cancer prevalence and case difficulty of 109 mammograms each, sampled from 130 screening mammograms interpreted by BCSC radiologists from 2000 to 2003. Some of the 130 cases were included in 1 or more test sets on the basis of the desired composition of each test set. For example, test sets 1 and 2 contained 13% subtle cases, while test sets 3 and 4 contained 30% subtle cases. To achieve the needed proportions, some of the subtle cases were used in 2 or more of the test sets. We used screening examinations for women aged 40 to 69 years with previous mammograms within the prior 11 to 30 months that could be used for comparison views. We excluded women with a history of breast cancer or breast augmentation. Each test set case consisted of craniocaudal and mediolateral oblique views of each breast (4 views per woman for each screening and comparison examination).

Custom-designed software for viewing images and collecting interpretations was created in collaboration