Original Investigation

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

Diana L. Miglioretti, PhD, Laura Ichikawa, MS, Robert A. Smith, PhD, Diana S. M. Buist, PhD, Patricia A. Carney, PhD, Berta Geller, EdD, Barbara Monsees, MD, Tracy Onega, PhD, Robert Rosenberg, MD, Edward A. Sickles, MD, Bonnie C. Yankaskas, PhD, Karla Kerlikowske, MD

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 0 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.

Key Words: Screening mammography; interpretive performance; test sets.

Acad Radiol 2017; ■:■■–■■

From the 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 (D.L.M.); Kaiser Permanente Washington Health Research Institute, Seattle, Washington (D.L.M., L.I., D.SM.B.); Cancer Control Department, American Cancer Society, Atlanta, Georgia (R.A.S.); Departments of Family Medicine and Public Health and Preventive Medicine, School of Medicine: Mail Code FM, Oregon Health & Science University, Portland, Oregon (P.A.C.); Office of Health Promotion Research, Department of Family Medicine, University of Vermont, Burlington, Vermont (B.G.); Mallinckrodt Institute of Radiology, Washington University, St. Louis, Missouri (B.M.); Department of Community and Family Medicine, Dartmouth Medical School, HB 7927—Community & Family Medicine, Lebanon, New Hampshire (T.O.); University of New Mexico-HSC and Radiology Associates of Albuquerque, Albuquerque, New Mexico (R.R.); Department of Radiology, University of California, San Francisco Medical Center, San Francisco, California (E.A.S.); Department of Radiology, University of North Carolina, Chapel Hill, North Carolina (B.C.Y.); Departments of Medicine and Epidemiology and Biostatistics (K.K.); General Internal Medicine Section, Department of Veterans Affairs, University of California, San Francisco, California (K.K.). Received November 21, 2016; revised March 16, 2017; accepted March 17, 2017. 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 (U01CA63740, U01CA86076, U01CA86082, U01CA70013, U01CA69976, U01CA63731, U01CA70040). 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 United States. For a full description of these sources, please see http://www.bcsc-research.org/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. **Address correspondence to:** D.L.M. e-mail: dmglioretti@ucdavis.edu

http://dx.doi.org/10.1016/j.acra.2017.03.016

INTRODUCTION

The interpretive performance of screening mammography varies extensively among US radiologists (1,2). Given US radiologists 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 or images) and variability in test set composition, performance measures evaluated, and statistical approaches used (7–9). In a study of 27 US 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: 0.16, 0.62), but no evidence of correlation between clinical and test set sensitivity (–0.18, 95% Bayesian credible interval: –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.

MATERIALS 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 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 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 to 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, 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