Original Investigation

Strategies for Decreasing Screening Mammography Recall Rates While Maintaining Performance Metrics

Lisa A. Mullen, MD, Babita Panigrahi, MD, Jacqueline Hollada, MPH, Benita Panigrahi, MD, Eniola T. Falomo, MD, Susan C. Harvey, MD

Rationale and Objective: This study aimed to determine the impact of interventions designed to reduce screening mammography recall rates on screening performance metrics.

Materials and Methods: We assessed baseline performance for full-field digital mammography (FFDM) and digital breast tomosynthesis mammography (DBT) for a 3-year period before intervention. The first intervention sought to increase awareness of recalls from screening mammography. Breast imagers discussed their perceptions regarding screening recalls and were required to review their own recalled cases, including outcomes of diagnostic evaluation and biopsy. The second intervention implemented consensus double reading of all recalls, requiring two radiologists to agree if recall was necessary. Recall rates, cancer detection rates, and positive predictive value 1 (PPV1) were compared before and after each intervention.

Results: The baseline recall rate, cancer detection rate, and PPV1 were 11.1%, 3.8/1000, and 3.4%, respectively, for FFDM, and 7.6%, 4.8/1000, and 6.0%, respectively, for DBT. Recall rates decreased significantly to 9.2% for FFDM and to 6.6% for DBT after the first intervention promoting awareness, as well as to 9.9% for FFDM after the second intervention implementing group consensus. PPV1 increased significantly to 5.7% for FFDM and to 9.0% for DBT after the second intervention. Cancer detection rate did not significantly change with the implementation of these interventions. An average of 2.3 minutes was spent consulting for each recall.

Conclusion: Reduction in recall rates is desirable, provided performance metrics remain favorable. Our interventions improved performance and could be implemented in other breast imaging settings.

Key Words: Breast cancer; mammographic screening; digital breast tomosynthesis; recall from screening.

INTRODUCTION

Breast cancer is the most common cancer affecting women worldwide, with approximately 500,000 annual deaths due to disease-specific mortality. An estimated 246,660 new cases of invasive breast cancer will be diagnosed in the United States alone, with 40,450 deaths from the disease in 2016 (1,2). Early detection with screening mammography leads to improved survival and less aggressive treatment, decreasing mortality from breast cancer by an estimated 19%–40% depending on age and breast tissue density (3–5).

Screening mammography benefits women with earlier detection, but less desirable outcomes may occur if a woman is recalled for additional evaluation and the result is ultimately benign. After a recall from screening mammography, a benign interpretation may occur with additional imaging alone. However, in some cases, large core needle biopsy or even surgery may be required to prove benignity. False-positive screening recalls and the resulting additional evaluation may cause significant psychological distress to women (6–9).

There are significant financial costs associated with false-positive recalls in the United States. One report suggests an annual cost of $1.6 billion dollars in the United States for recall utilization resulting from mammographic screening (10). It is challenging to estimate the actual total cost of each recall from screening. Work by Bonafede et al. calculates an average weighted cost per recall of $1082 for health insurance companies (11). The patient cost burden is also substantial, averaging $138 for all recalled patients, and $449 for those undergoing a biopsy (12). These numbers do not include indirect costs to the patient or employer. Resource utilization must also be considered, including physician time, equipment utilization, technologist and support staff time, clinic/facility time, and even operative time.

Acad Radiol 2017; ■:■■–■■

From the The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University School of Medicine, 601 North Caroline St Suite 4120 E, Baltimore, MD 21287 (L.A.M., B.P., E.T.F., S.C.H.); Department of Health, Behavior, and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (J.H.); The Earth Institute, Columbia University, New York, New York (B.P.). Received January 20, 2017; revised April 14, 2017; accepted June 8, 2017. **Address correspondence to:** S.C.H. e-mail: sharvey7@jhmi.edu

http://dx.doi.org/10.1016/j.acra.2017.06.009

Reducing the recall rate from screening mammography while maintaining performance metrics could significantly decrease direct and indirect costs. If one considers a population of 25,000 women undergoing annual screening mammography and the average US recall rate of 10%, 2500 women will be recalled for additional imaging. If the recall rate could be reduced to 5%, 1250 fewer women would be recalled. Considering the average cost per recall of $1082 (cost to the health insurance company), the cost of recall could be reduced by $1,352,500 annually. Decreasing the recall rate would also decrease other associated costs, including the costs to patients.

To safely and effectively improve value, performance metrics, including cancer detection rate and positive predictive value (PPV), must be maintained so that the quality and safety of the screening program is preserved. There are reports that low recall rates and high performance metrics are achieved in European countries (13).

European literature suggests that double reading of screening mammograms by at least two radiologists is an effective method to decrease recall rates without sacrificing cancer detection rates (14–16). Such practices are routine in many countries, including Great Britain (13). Although cost concerns regarding radiologists' time in the United States may deter groups from routinely double reading screening mammograms, targeted utilization of double reading potential recalls may make the practice more feasible.

Our study describes a method with the potential to achieve lower recall rates from screening mammography while maintaining performance metrics. We hypothesize that using review of one's own screening recalls as well as requiring consensus of two readers for potential recalls can safely decrease recall rates. We additionally propose that these interventions can be implemented and scaled to both community and academic settings to improve performance. Our pilot data suggest that these simple, efficient interventions can decrease recall rates. This may lead to significant resource savings and health-care cost reduction, and decrease unnecessary distress to women.

MATERIALS AND METHODS

This is an institutional review board (IRB) approved and Health Insurance Portability and Accountability Act compliant study. Between January 3, 2012 and April 3, 2016, data pertaining to mammography screening examinations, both standard two-dimensional (2D) full-field digital mammography (FFDM) and three-dimensional (3D) digital breast tomosynthesis (DBT), were obtained from Magview (version 3, Burtonsville, MD), a software system which tracks, reports and records mammography outcomes. Screening mammograms from three outpatient sites of an academic breast imaging division were included in the analysis. A total of 10 radiologists, all breast imaging specialists, participated in the study. The breast imagers had an average of 10.4 years of experience (range 1–22 years). The data collected included the date of the screening mammogram, reviewing physician, and the Breast Imaging Reporting and Data System (BI-RADS) assessment.

All screening examinations with a BI-RADS 0 assessment, indicating that the patient was recalled for additional imaging, were identified. The case outcomes were analyzed to determine the final BI-RADS assessment for the recalled patients, identifying all follow-up imaging, procedures, and surgeries performed on each patient. All BI-RADS 0 examinations that were recalled for technical reasons were excluded.

Screening mammography performance metrics obtained for each period of the study included the recall rate (screening callback studies divided by total number of screening examinations), cancer detection rate (true positive screening examinations per 1000 screening examinations), and PPV1 (positive predictive value 1, ie, the percentage of positive screening examinations resulting in a tissue diagnosis of cancer), broken down by FFDM and DBT. The date range of January 3, 2012 to February 2, 2015 was considered the baseline period, before any interventions occurred.

The first intervention was designed to improve imager awareness of group and individual performance, and occurred over a 7-month period from February 3, 2015 to September 3, 2015. At the beginning of this phase, each breast imager compared his or her individual performance metrics to that of the group. The group discussed perceptions of recall and performance, including the most frequent individual reasons for recall and individual fears that prompt recall. The physicians set a goal to reduce both the division's recall rate and each physician's recall rate to 5%, while monitoring cancer detection rate and PPV to be certain that the lower recall rate was safe for patients. The rate of 5% was targeted from the successful performance of mammography screening programs in Europe (13). The next phase of the first intervention required each radiologist to review the imaging and reports of his or her personal screening recalls, followed by the imaging and reports from the subsequent diagnostic evaluation or biopsy for each recalled patient. This review was performed on a weekly basis.

The second intervention occurred over a 7-month period from September 4, 2015 to April 3, 2016. During this phase, all potential recalls underwent consensus double read. A second reader determined whether or not he/she agreed with the recall suggested by the first reader. If the second reader agreed with the primary reader, the patient was recalled. If the second reader did not feel that the finding warranted recall, a third reader was asked to provide the tie-breaking decision for recall vs no recall. All second and third reader reviews, as well as release of the final report, were required to be completed within 24 hours of the first reader interpretation, so there was no significant delay in releasing the final result to the provider and the patient. Data were collected for each recalled case, using an integrated electronic form with required fields, so that all data were consistent. The collected data included the names of all readers, patient demographic data including age, type of mammogram (FFDM or DBT), description of the finding(s), breast tissue density, and availability of comparison studies. The amount of time for the consensus review was also recorded, which included time for a third reader consultation, if needed.