Radiology Trainee Performance in Digital Breast Tomosynthesis: Relationship Between Difficulty and Error-Making Patterns

Lars J. Grimm, MD, MHS[a], Jing Zhang, PhD[a,b], Joseph Y. Lo, PhD[a,c,d], Karen S. Johnson, MD[a], Sujata V. Ghate, MD[a], Ruth Walsh, MD[a], Maciej A. Mazurowski, PhD[a,c]

Abstract

Purpose: The aim of this study was to better understand the relationship between digital breast tomosynthesis (DBT) difficulty and radiology trainee performance.

Methods: Twenty-seven radiology residents and fellows and three expert breast imagers reviewed 60 DBT studies consisting of unilateral craniocaudal and medial lateral oblique views. Trainees had no prior DBT experience. All readers provided difficulty ratings and final BI-RADS® scores. Expert breast imager consensus interpretations were used to determine the ground truth. Trainee sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) were calculated for low- and high-difficulty subsets of cases as assessed by each trainee him or herself (self-assessed difficulty) and consensus expert-assessed difficulty.

Results: For self-assessed difficulty, the trainee AUC was 0.696 for high-difficulty and 0.704 for low-difficulty cases ($P = .753$). Trainee sensitivity was 0.776 for high-difficulty and 0.538 for low-difficulty cases ($P < .001$). Trainee specificity was 0.558 for high-difficulty and 0.810 for low-difficulty cases ($P < .001$). For expert-assessed difficulty, the trainee AUC was 0.645 for high-difficulty and 0.816 for low-difficulty cases ($P < .001$). Trainee sensitivity was 0.612 for high-difficulty and .784 for low-difficulty cases ($P < .001$). Trainee specificity was 0.654 for high-difficulty and 0.765 for low-difficulty cases ($P = .021$).

Conclusions: Cases deemed difficult by experts were associated with decreases in trainee AUC, sensitivity, and specificity. In contrast, for self-assessed more difficult cases, the trainee AUC was unchanged because of increased sensitivity and compensatory decreased specificity. Educators should incorporate these findings when developing educational materials to teach interpretation of DBT.

Key Words: tomosynthesis, error-making patterns, difficulty, breast cancer, residents

J Am Coll Radiol 2016;13:198-202.

INTRODUCTION

Digital breast tomosynthesis (DBT), also referred to as 3-D mammography, is a new breast cancer screening modality that is being rapidly adopted into clinical practice to supplement traditional 2-D mammography [1-3]. Although similar in principle to mammography, DBT requires more complicated image acquisition and postprocessing techniques but allows a radiologist to scroll through the breast slice by slice to remove tissue superimposition [4,5]. These theoretical benefits have been validated in several large prospective screening trials, which have demonstrated that DBT performed in conjunction with mammography leads to reduced screening recall rates and increased invasive cancer detection rates compared with mammography alone [6-8]. These benefits have been shown to apply to all women regardless of breast density [9]. Although the use of DBT is still an area of active research, its use in clinical practice is growing quickly, and radiologists who interpret breast imaging studies are expected to confidently interpret DBT in conjunction with mammography.

[a]Department of Radiology, Duke University Medical Center, Durham, North Carolina.

[b]Department of Computer Science, Lamar University, Beaumont, Texas.

[c]Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina.

[d]Department of Biomedical Engineering, Duke University, Durham, North Carolina.

Corresponding author and reprints: Lars J. Grimm, MD, MHS, Duke University Medical Center, Department of Radiology, Box 3808, Durham, NC 27710; e-mail: lars.grimm@duke.edu.

During the course of this study, Maciej A. Mazurowski, PhD, had an RSNA Education Scholar Grant. The authors have no conflicts of interest related to the material discussed in this article.

1546-1440/15/$36.00 ■ http://dx.doi.org/10.1016/j.jacr.2015.09.025

To ensure that radiology residents graduate with a working knowledge of the next generation of imaging techniques, the ACR and the Society of Breast Imaging recommend that all residents be familiar with DBT [10]. However, incorporating DBT into existing educational training is a unique challenge, in part because this is also a new imaging modality for radiology faculty members. DBT is similar to mammography in principle and relies on the same BI-RADS® lexicon, but many of the rules for mammography do not apply to DBT. For example, readers cannot safely assume that all fat-containing masses or circumscribed, lobulated masses on DBT are benign or probably benign [7,11]. Knowledge of mammography and cross-sectional imaging will provide the basis on which residents approach DBT. Previous work has shown that there are unique error-making patterns associated with difficulty in mammography [12], but there is currently a paucity of published literature regarding resident error-making patterns and best educational practices for DBT.

The purpose of this study was to better understand radiology trainee error-making patterns by investigating the relationship between error making and difficulty for DBT. Specifically, the concepts of self-assessed and expert-assessed difficulty for individual DBT cases were investigated. Understanding the influence of DBT difficulty on the behavior of radiology trainees is the first step toward building better educational approaches to adequately prepare trainees to interpret DBT studies upon residency and fellowship graduation.

METHODS

Readers

After obtaining institutional review board approval, we performed a reader study for radiology trainees in DBT. There were 27 readers, including radiology residents at various levels of training (first year, n = 5; second year, n = 3; third year, n = 3; fourth year, n = 9), non-breast imaging fellows (n = 4), and breast imaging fellows (n = 3), from two academic institutions. The readers had varying degrees of breast imaging experience, but no prior DBT experience, as DBT had not yet been incorporated into the clinical workflow at the institutions at the time of this study. Before the start of the study, all readers were given a short primer on DBT acquisition techniques, viewing protocols, and similarities with mammography. Three fellowship-trained breast imagers with 7 to 20 years of breast imaging experience, certified to interpret DBT studies, and with DBT experience from multiple prior clinical investigations also participated as readers.

DBT Studies

Each reader reviewed 60 DBT studies on a research workstation in a radiology reading room with a medical-grade monitor consisting of craniocaudal and mediolateral oblique views of a single breast. A graphical user interface that was similar to a clinical setup was developed for use in the study that allowed cases to be anonymized, images to be annotated, and follow-up questions to be presented to the reader. Readers were able to scroll, pan, and zoom the images independently. The cases included normal, benign, and malignant studies collected from a previous research study. The interpretations of the expert readers were used as the gold standard, which resulted in 33 positive and 27 negative studies.

Reader Interpretations

After reviewing each case, the trainee and expert readers provided difficulty assessments on a scale ranging from 1 to 5, with 5 considered the most difficult. The trainee and expert readers were not given specific criteria regarding the difficulty assessment but were instead asked to give overall personal assessments of case difficulty. The trainee and expert readers then reported final BI-RADS scores of 1 to 5. A BI-RADS score of 1 or 2 was considered negative and a score of 3, 4, or 5 was considered positive because a score of 1 or 2 is considered to represent a 0% likelihood of malignancy by the reader and a score of 3, 4, or 5 indicates that further workup is needed, either additional imaging surveillance or biopsy. The ground truth was established from a consensus interpretation of the three expert readers: if two or three of the experts deemed a case to be positive, it was considered positive. Using reader consensus interpretations instead of pathology results to test reader performance allowed us to test how closely the performance of the trainees mirrored that of consensus expert readers [12,13].

Assessing Difficulty

The cases were divided into low and high difficulty for each individual trainee and collectively for the experts. For each trainee, a specific threshold between low and high difficulty was applied on the difficulty scores provided by the trainee to allow as even a split between low and high difficulty as possible. As a result, for one trainee, a difficulty value of 3 might represent a low-difficulty case if that trainee mostly reported difficulty scores of 4 and 5, but for another trainee, a difficulty value of 3 might represent a high-difficulty case if that trainee mostly reported difficulty scores of 1 and 2. This approach allowed us to normalize difficulty levels among individuals so that