Title:

Does breast imaging experience during residency translate into improved initial performance in digital breast tomosynthesis?

Authors:

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

ABSTRACT

Rationale and Objectives: To determine the initial digital breast tomosynthesis (DBT) performance of radiology trainees with varying degrees of prior breast imaging experience.

Material and Methods: To test trainee performance with DBT, we performed a reader study following IRB approval. Two medical students, twenty radiology residents, four non-breast imaging fellows, three breast imaging fellows, and three fellowship-trained breast imagers reviewed 60 unilateral DBT studies (craniocaudal and medio-lateral oblique views). Trainees had no prior DBT experience. Each reader recorded a final Breast Imaging-Reporting and Data System (BI-RADS) assessment for each case. The fellowship-trained breast imager consensus
interpretations were used to establish the ground truth. Area under the Receiver Operating Characteristic curve (AUC), sensitivity, and specificity were calculated. For analysis, first through third year residents were classified as junior trainees and fourth year residents plus non-breast imaging fellows were classified as senior trainees.

Results: The AUCs were .569 for medical students, .721 for junior trainees, .701 for senior trainees, and .792 for breast imaging fellows. The junior and senior trainee AUCs were equivalent (p<.01) using a two one-sided test for equivalence with a significance threshold of 0.1. The sensitivities and specificities were highest for breast imaging fellows (.778 and .815 respectively), but similar for junior (.631 and .714 respectively) and senior trainees (.678 and .661 respectively).

Conclusion: Initial performance with DBT among radiology residents and non-breast imaging fellows is independent of years of training. Radiology educators should consider these findings when developing educational materials.
INTRODUCTION

Digital breast tomosynthesis (DBT) is a promising new imaging modality that has the potential to transform breast cancer screening practices [1,2]. Several large studies have demonstrated that compared to digital mammography alone, the addition of DBT results in decreased screening recall rates and increased cancer detection rates despite fewer biopsies [3-7]. These advantages are due to the ability of DBT to decrease tissue superimposition which allows the radiologist to more easily distinguish normal overlapping tissue from abnormal tissue that appears conspicuous [8]. Importantly, this increase in cancer detection is almost exclusively for invasive ductal carcinoma and not ductal carcinoma in situ [3,6]. These promising results have led to the steadily growing adoption of DBT into breast imaging practices.

The incorporation of DBT into routine clinical use means that the next generation of radiologists should be trained in this new imaging modality. Graduating radiology residents may join academic or private practices which already incorporate breast tomosynthesis into routine clinical practice. Furthermore, the American College of Radiology and the Society of Breast Imaging guidelines now recommend that residents should be familiar with the “principles, methods, strengths, and pitfalls of digital breast tomosynthesis” [9]. However, there is no published literature that has evaluated the best means of incorporating DBT into residency education. Although residents may be familiar with mammography and cross-sectional imaging, it is currently unknown if these skills translate to an understanding of how to interpret DBT studies.
The first step towards teaching DBT to radiology residents is to develop an understanding of their baseline abilities. Although it may be assumed that radiology residents with more breast imaging and general radiology experience would initially perform better with DBT, this hypothesis has not yet been tested. The goal of this study is to test the initial DBT interpretive skills for radiology residents with varying degrees of breast imaging training.

METHODS
Reader Study
In order to assess the performance of radiology residents in DBT we performed a reader study. The study was approved by the Institutional Review Board on June 8th, 2012. Two medical students, twenty radiology residents, four non-breast imaging fellows, and three breast imaging fellows from two academic institutions were recruited to act as readers. All trainees, with the exception of the medical students who acted as controls, had varying degrees of experience interpreting breast imaging studies. The medical students had all spent a month on a radiology elective, including several days on breast imaging, and were given additional instructions on the classic appearance of cancers. None of the trainee readers had any prior experience with DBT. Prior to the start of the study, all readers were given a short primer on the technique of viewing DBT studies and the similarities between DBT and mammography. Each reader reviewed 60 DBT studies at a clinical workstation, consisting of DBT craniocaudal and medio-lateral oblique views of a single breast. In addition, three fellowship trained breast imagers with 7-20 years of experience reading 2D mammography, certified to interpret DBT studies, and with DBT experience from prior clinical studies reviewed the same studies to establish the gold standard.
Readers recorded a final Breast Imaging-Reporting and Data System (BI-RADS) assessment for each case. The cases included for analysis were obtained from a previous research study which included a mix of normal, benign, and malignant cases, although the readers were unaware of the distribution.

Data Analysis

The BI-RADS interpretations for each expert were binarized so that a BI-RADS assessment of 1 or 2 was considered negative and 3, 4, or 5 was considered positive. Consensus interpretations from the three fellowship-trained breast imagers were used to establish the ground truth (i.e., if two readers marked a case as positive, then it was considered positive, even if the third reader marked it as negative). Area under the Receiver Operating Characteristic curve (AUC), sensitivity, and specificity were calculated to assess the performance of each trainee reader, using the expert consensus interpretation as the ground truth. BI-RADS assessments by the trainee readers were used as the response variable to calculate the AUCs. By utilizing expert consensus interpretations rather than pathology results as the ground truth, we were able to test the interpretative skills of the readers to assess which errors are due to a lack of expertise (i.e., difference between the reader and the experts) rather than imperfections in the imaging modality (i.e., errors that any reader regardless of experience would make). The convention of using reader consensus interpretations has been utilized previously [10,11].

We divided the readers into four groups. The first group was comprised of medical students and therefore represented the pre-residency experience (0 weeks of breast imaging). The second group was junior trainees which were radiology residents in years one through three (mean 5.4
and range 1-8 weeks of breast imaging). The third group was senior trainees which were radiology residents in year four and non-breast imaging radiology fellows (mean 10.7 and range 8-16 weeks of breast imaging). The last group consisted of trainees who had completed a radiology residency and were currently enrolled in a breast imaging fellowship (mean 33.3 and range 20-40 weeks of breast imaging). Comparisons were made between the four groups: medical students, junior trainees, senior trainees, and breast imaging fellows. A t-test was used to test for differences between groups. A two one-sided tests approach, with a difference between AUCs of 0.1 considered significant, was used to test for equivalence between the groups. This test compares two groups and a p-value of less than 0.05 indicates that the AUCs are equivalent at the threshold of 0.1 considered significant.

RESULTS

The consensus expert breast imager interpretations resulted in 55.0% (33/60) of cases deemed positive (BI-RADS 3, 4, or 5). In those positive cases, the experts reached agreement on morphology (i.e. two or more readers assigned the same type of an abnormality) for 16 masses, 2 calcifications, 4 asymmetries, and 4 architectural distortions. There was no agreement regarding morphology in 8 cases. In comparison, the distribution of positive interpretations was 38.3% (46/120) for medical students, 47.6% (314/660) for junior trainees, 52.6% (410/780) for senior trainees, and 51.1% (92/180) for breast imaging fellows. The breakdown of the AUC, sensitivity, and specificity for each group is shown in Table 1. Plots of the AUC, sensitivity, and specificity for all groups are shown in Figure 1. In addition, a plot of the AUC by weeks of training is shown in Figure 2.
In general, medical students performed worse than all other groups. The medical student AUC was significantly worse than that of junior trainees (p=.02), senior trainees (p=.03), and breast imaging fellows (p=.01). This difference is mostly explained by a lower sensitivity, however, the differences in sensitivity and specificity between the medical students and other groups did not reach statistical significance (p>=.10 for all tests).

Junior and senior trainees had very similar performance. The difference between the AUCs of junior (.721) and senior (.701) trainees did not reach statistical significance (p=.51). This corresponded to a slightly better, but not significant (p=.19), sensitivity for senior trainees (.678 vs .631) and a slightly better, but not significant (p=.77), specificity for junior trainees (.714 vs .661). Based on the similarity in AUC between the junior and senior trainees, a two one-sided tests was performed which demonstrated equivalence (p<.01) between the two groups. This means that the AUC of the junior and senior trainees was statistically identical when a value of .1 was considered significantly different.

In contrast, breast imaging fellows generally performed better than the other groups. The mean AUC of breast imaging fellows (.792) was slightly better than junior trainees (p=.08) and senior trainees (p=.05). However, while the sensitivity for breast imaging fellows was significantly better than junior (p<.01) and senior trainees (p<.01), their specificity was not significantly better than junior (p=.17) or senior trainees (p=.09).
DISCUSSION

Our study provides some insights into the performance of radiology trainees when initially exposed to digital breast tomosynthesis. Previous work with mammography has demonstrated that with increased experience radiology trainees perform better by utilizing more efficient visual search strategies which allow for fewer false positive interpretations[12,13]. In contrast, our study unexpectedly demonstrated that the initial performance of residents was equivalent regardless of the number of years of training, despite the strong similarity between DBT and mammography. Radiology residents are expected to get progressively more competent in the interpretation of breast imaging studies over the course of their residency; however, experience in breast imaging and other areas of radiology do not appear to translate to initial DBT performance.

These findings have implications for radiology educators who are introducing DBT into residency programs. As radiology residents transition from junior to senior residents, their skill sets improve and they are generally entrusted with a greater deal of autonomy (e.g., providing preliminary interpretations to referring clinicians or guiding technologists to acquire additional views). However, the results of our study suggest that when DBT is introduced, senior residents will require as much teaching as junior residents and thus should not be expected to provide more accurate interpretations than their junior colleagues until dedicated DBT-instruction has been implemented. In addition, DBT teaching materials designed by educators should be targeted toward a baseline level of understanding and not tiered towards the level of residency training. An interesting follow up study would test the performance of trainees over time to see if learning for DBT is faster for trainees with more experience in mammography. Do senior
residents improve faster because they are able to apply their general breast imaging experience once they become comfortable with the basics of DBT or do junior residents improve faster because they are not hindered with preconceived notions of how abnormalities traditionally manifest on 2D imaging.

The number of medical students (n=2) and breast imaging fellows (n=3) included in the study was small, but they were included to provide upper and lower boundaries for performance expectations. Medical students acted as medically literate controls. Medical students would be expected to perform the worst while breast imaging fellows would be expected to perform the best. If the performance of all readers, medical students through breast imaging fellows, was identical, then it would indicate that DBT is such a novel modality that radiology experience was meaningless. Instead, we were able to demonstrate that performance with DBT is better among radiology residents than controls, but the effect is the same regardless of the number of residency years completed. Only those with no radiology, and thus no breast imaging experience (i.e., medical students), performed worse and only those with significantly more breast imaging experience (i.e., breast imaging fellows) performed better.

Although we recognize that not all radiology residents of a similar year will have had the same number of weeks dedicated to a breast imaging rotation, we chose to focus our analysis based on the number of years of training, as opposed to the number of weeks. We did this because the skill level expected from a particular radiology resident is typically based on their year of training and it is a common way residents are categorized. Furthermore, we chose to focus on the year of training since clinical duties are typically sorted in this way when residents are on
service. Different clinical responsibilities may be assigned to residents based on their year of training, not on the number of weeks one has spent on a particular service. Finally, we felt it more appropriate to sort by year of training rather than weeks on service as we presume a more senior resident would likely have been exposed to a greater number of breast imaging lectures and case conferences and thus have a greater fund of knowledge than a more junior resident even if both had spent the same total number of weeks rotating through breast imaging. The conclusions however, are similar regardless of how our study categorized trainees, as shown in Figure 2.

Our study provides a snapshot into the performance of radiology trainees with a new imaging modality, which to our knowledge is the first study to look specifically at trainee performance with DBT. Due to the increased interpretation time for DBT, the total number of cases interpreted by each trainee (n=60) in our study was smaller than some similar mammography studies (range: 50-150); however, we included more readers (n=29) than in comparable mammography studies (range:10-19) to compensate for this difference [10,11,13-15]. This allows us to evaluate 1740 data points (60 studies times 29 readers), which provides similar, if not better, statistical power than prior work evaluating resident performance in mammography. While it is possible that some differences between junior and senior residents would be found in a larger study, our study strongly indicates that such difference would be very small.

There are some limitations to our study. Readers were only provided unilateral views of the breast and no 2D mammograms were available for comparison. We realize this does not reflect the real world practice as DBT is interpreted in conjunction with 2D mammography and the lack
of bilateral imaging hinders one’s ability to assess for symmetry, but these limitations should affect the performance of both trainees and experts similarly. The study was designed to isolate DBT interpretation skills and therefore only DBT images were provided for review. In addition, there were a higher percentage of actionable cases (BI-RADS 3, 4, or 5) than in general population which may have resulted in improved reader performance. The expert breast imagers all have extensive experience with breast imaging, but since DBT is such a new modality their experience with DBT is less, although they had all participated in prior DBT reader studies. We used consensus expert interpretations to compensate for individual reader inexperience. Finally, since this is a new modality, there may be a learning curve that developed over the course of the study for the readers. Future work looking at performance over time may be able to better ascertain the slope of the learning curve.

Our study demonstrates that the initial performance with DBT for radiology residents and non-breast imaging fellows is independent of years of breast imaging experience and exposure to 2D mammography. Radiology educators should keep these findings in mind when they are developing educational materials for residents and also when determining the degree of autonomy and level of expectation from trainees new to DBT. To our knowledge, there is currently no published literature evaluating DBT and radiology residency education. As DBT continues to be adopted into clinical practice, the findings from this study support the need to incorporate dedicated DBT training into the breast imaging rotation of radiology residency programs and can provide a foundation upon which future educational research efforts can be based.
REFERENCES

Table 1. Area under the Receiver Operating Characteristic Curve (AUC), sensitivity, and specificity with standard deviations for trainees utilizing expert breast imager consensus interpretations as the gold standard.

<table>
<thead>
<tr>
<th>Training Level</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Students (n=2)</td>
<td>.569 ± .040</td>
<td>.439 ± .150</td>
<td>.685 ± .131</td>
</tr>
<tr>
<td>Junior Trainees* (n=11)</td>
<td>.721 ± .072</td>
<td>.631 ± .142</td>
<td>.714 ± .155</td>
</tr>
<tr>
<td>Senior Trainees¥ (n=13)</td>
<td>.701 ± .072</td>
<td>.678 ± .105</td>
<td>.661 ± .188</td>
</tr>
<tr>
<td>Breast Imaging Fellows (n=3)</td>
<td>.792 ± .061</td>
<td>.778 ± .017</td>
<td>.815 ± .134</td>
</tr>
</tbody>
</table>

* Junior trainees refer to radiology residents in years one, two, and three
¥ Senior trainees refer to radiology residents in year four and non-breast imaging fellows.
Figure 1. Plots of the Area under the Receiver Operative Curve (AUC, A), sensitivity (B), and specificity (C). Dots refer to the performance of individual readers in each group. The bars refer to the standard deviation for the group. The lines connect the means of the groups.
Figure 2. Area under the Receiver Operating Characteristic Curve (AUC) versus the number of weeks of training.