Digital Breast Tomosynthesis: A Pilot Observer Study

OBJECTIVE. The objective of our study was to assess ergonomic and diagnostic performance–related issues associated with the interpretation of digital breast tomosynthesis–generated examinations.

MATERIALS AND METHODS. Thirty selected cases were read under three different display conditions by nine experienced radiologists in a fully crossed, mode-balanced observer performance study. The reading modes included full-field digital mammography (FFDM) alone, the 11 low-dose projections acquired for the reconstruction of tomosynthesis images, and the reconstructed digital breast tomosynthesis examination. Observers rated cases under the free-response receiver operating characteristic, as well as a screening paradigm, and provided subjective assessments of the relative diagnostic value of the two digital breast tomosynthesis–based image sets as compared with FFDM. The time to review and diagnose each case was also evaluated.

RESULTS. Observer performance measures were not statistically significant ($p > 0.05$) primarily because of the small sample size in this pilot study, suggesting that showing significant improvements in diagnosis, if any, will require a larger study. Several radiologists did perceive the digital breast tomosynthesis image set and the projection series to be better than FFDM ($p < 0.05$) for diagnosing this specific case set. The time to review, interpret, and rate the examinations was significantly different for the techniques in question ($p < 0.05$).

CONCLUSION. Tomosynthesis-based breast imaging may have great potential, but much work is needed before its optimal role in the clinical environment is known.

Recent advances in digital imaging, including full-field digital mammography (FFDM), have enabled us to revisit tomosynthesis in a way that is practical and may actually be relatively easily implemented on several digital systems being used in radiology in general and for breast imaging in particular [1–4]. Although current interest is primarily in using this approach for breast imaging, tomosynthesis is relevant to several procedures, such as chest imaging. Digital breast tomosynthesis is of great interest in screening and diagnostic breast imaging for several reasons including, but not limited to, the possibility of reducing recall rates in screening mammography; improving detection of abnormalities in women with dense breast tissue; improving diagnosis of benign findings, thereby reducing the number of negative biopsies; and assessing therapeutic efficacy.

Several studies have begun to address technical, ergonomic, and performance issues associated with this technology and its application to breast imaging, but to date there are no conclusive results in any of these aspects [5, 6]. Because the display environment is likely to be an important factor in our ability to incorporate this approach into the necessarily efficient routine practices of clinical breast imaging, we have embarked on a comprehensive project to assess display-related issues in a series of observer performance studies. This article describes our general methodology in this regard and presents the results of a preliminary (i.e., pilot) multimode observer performance study.

Keywords: breast cancer, breast screening, digital breast tomosynthesis, full-field digital mammography, observer performance study, tomosynthesis

DOI:10.2214/AJR.07.2841

Received July 9, 2007; accepted after revision October 6, 2007.

This work was supported in part by grant BCTR0600733 to the University of Pittsburgh from the Susan B. Komen foundation.

1Department of Radiology, Imaging Research, University of Pittsburgh, F.A.R.P. Bldg., 3362 Fifth Ave., Pittsburgh, PA 15213. Address correspondence to D. Gur (gurd@upmc.edu).

2Department of Radiology, Magee-Women’s Hospital, Pittsburgh, PA.

AJR 2008; 190:865–869
0361–803X/08/1904–865
© American Roentgen Ray Society
with each view displayed on our custom-designed workstation. The digital mammography images used in this study consisted of FFDM images and tomosynthesis images acquired at our institution and other institutions for other research purposes. The acquisition of all examinations was performed under institutional review board (IRB)-approved protocols that included a signed informed consent by all participants. IRB approval at this institution was obtained for this specific observer performance study as well.

During acquisition of tomosynthesis images, the breast is compressed in a conventional manner, and the X-ray tube moves along a limited arc allowing 11 low-dose images to be acquired rather than the single image acquired during a conventional FFDM examination [2]. Eleven projection images, referred to as “frames,” are acquired with the system in question (Genesis Tomosynthesis System, Hologic). All acquisitions were performed at combined per-examination doses (for all 11 projections) that are comparable to an FFDM examination, and the average mid breast dose was approximately 2 mGy per view. After acquisition, the data from the projection images, or frames, are used to reconstruct between 60 and 80 parallel slices (i.e., the 3D digital breast tomosynthesis data set) depending on the thickness of the compressed breast. The reconstructed 3D data are referred to as the “digital breast tomosynthesis image sets” or “digital breast mass images.”

In this study, readers were asked to detect and rate masses and microcalcification clusters using FFDM images (mode 1), the 11 frames (mode 2), and the reconstructed digital breast tomosynthesis images (mode 3) on our specially designed workstation. A management program determined the reading sessions for individual observers and the order of displayed cases during a session. Display modes were counterbalanced; hence, three readers initially read the FFDM examinations, three readers began with the frames, and three readers began with the digital breast tomosynthesis images. Each reader interpreted all 30 examinations under one mode in one session.

Cases

FFDM and digital breast tomosynthesis examinations performed on 30 women were interpreted in this study. Each examination consisted of two views of one breast, either the right CC and right MLO or the left CC and left MLO. Each examination was reviewed twice by an experienced observer with the knowledge of the verified truth to determine the presence or absence of masses, microcalcification clusters, or both and the location of the depicted abnormalities. All documents and all three display modes were available during these reviews. Five examinations were negative and depicted no abnormality. Table 1 summarizes the distribution of verified and visualized abnormalities in the data set by the type of image and whether the abnormality was associated with malignant findings (pathology). Note that the number of examinations with single versus multiple abnormalities are not included in Table 1.

On the FFDM images, a total of 23 masses and 15 microcalcification clusters were depicted in 25 examinations. Twelve examinations depicted only one abnormality, whereas the remaining 13 depicted multiple abnormalities. Eleven masses and seven microcalcification clusters were associated with malignancy, and the remaining 12 masses and eight microcalcification clusters were benign.

On the frame images, a total of 27 masses and 14 microcalcification clusters were depicted. Nine examinations depicted only one abnormality, whereas the remaining 16 depicted multiple abnormalities. Eleven of the depicted masses and seven microcalcification clusters were associated with malignancy, and the remaining 16 masses and seven microcalcification clusters were benign.

On the digital breast tomosynthesis images, a total of 28 masses and 14 microcalcification clusters were depicted. Nine examinations depicted only one abnormality, whereas the remaining 16 depicted multiple abnormalities. Eleven of the depicted masses and seven microcalcification clusters were associated with malignancy, and the remaining 17 masses and seven microcalcification clusters were benign.

Most of the abnormalities were visible on both the CC and MLO views. The center coordinate of each depicted abnormality was marked and saved in a reference file, which we refer to as the “truth file.” We found that one of the original FFDM examinations of a woman with dense breast tissue depicting a mass and a cluster that later proved to be malignant was of poor quality. Although the abnormalities were visible on both views during retrospective review, the cluster was extremely subtle in terms of its appearance on the mammograms. In addition, because most of the cases had undergone an FFDM examination that resulted in a recall recommendation before the digital breast tomosynthesis procedure, the data set may be biased in favor of the FFDM mode.

Selection of Observers and Prestudy Training

Nine board-certified radiologists with varying experience ranging from 5 to 35 years of reading mammography were selected for the study. Observers were unaware of the specific aims of the study and received an “Instructions to Observers” document to review before beginning the study. The document included a general overview of the study setup, a clear definition of the abnormalities in question, the process for reviewing and rating examinations during a session, how to rate (or not) certain abnormalities such as asymmetric density, and how to independently rate each view of an examination when appropriate. The readers were not made aware of the specifics of each mode until the time the reader would start that mode. Before the start of each mode, each observer was given a specific example and an interactive training session to familiarize him- or herself with the workstation functionality under the study conditions and the computerized scoring form. Observers were given an opportunity to ask questions and a staff member was available during the session to answer questions not related to the actual diagnosis. Because all of our clinical mammography operations are performed using FFDM (> 80,000 procedures per year) and readings are done on soft display, all observers were quite familiar with the use of the workstations.

Performance of the Study

In this study radiologists were asked to independently review and rate each examination for the presence or absence of the abnormalities in question under three reading conditions. The workstation (Dual Core AMD Opteron, Processor 270, 2 GHz and 6.00 GB of RAM) operates under Microsoft Windows Server 2003. The workstation display consists of two high-resolution (2,048 x 2,560), 8-bit gray-scale portrait monitors at a

### TABLE 1: Number of Visible Abnormalities by the Type of Images Displayed

<table>
<thead>
<tr>
<th>Visible Abnormality</th>
<th>FFDM (Total)</th>
<th>FFDM (Malignant %)</th>
<th>Frames (Total)</th>
<th>Frames (Malignant %)</th>
<th>Digital Tomosynthesis (Total)</th>
<th>Digital Tomosynthesis (Malignant %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass</td>
<td>23</td>
<td>11 (48)</td>
<td>27</td>
<td>11 (41)</td>
<td>28</td>
<td>11 (39)</td>
</tr>
<tr>
<td>Calcification cluster</td>
<td>15</td>
<td>7 (47)</td>
<td>14</td>
<td>7 (50)</td>
<td>14</td>
<td>7 (50)</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>18 (47)</td>
<td>41</td>
<td>18 (44)</td>
<td>42</td>
<td>18 (43)</td>
</tr>
</tbody>
</table>
nominal brightness of 80 foot-lamberts: two
Dome C5i flat-panel monitors (Planar Systems)
for image display.

The acquisition resolution for the FFDM images
was 70-µm per pixel. For the digital breast tomo-
synthesis images, reconstruction was at 1-mm
spacing (between slices) and in-plane pixel size
was approximately 120 µm. Because the projection
images, or frames, are acquired at a low dose, pixel
averaging was performed (“binning” 2 × 2), and the
effective in-plane pixel size was 140 µm.

Images could be magnified to full acquisition
resolution by a free-moving magnification box or
quadrant panning, and the frames and digital
breast tomosynthesis images could be sequentially
(serially) displayed as a continuing loop (“movie”)
or as one image at a time controlled manually at
the reader’s discretion and preference of image
display rate. All display function features were
mouse-driven. This workstation has been tested
extensively and used in other observer performance
studies [7].

During interpretation sessions, no comparison
examinations or other clinical information about
the patients were provided. The CC and MLO
views were displayed simultaneously on the left
and right monitors, respectively. The reading
sessions lasted approximately 5 months, with a
minimum required time of 1 month between
sessions. For all three modes, when the observer
detected a suspicious region, he or she moved the
cursor to the center of the suspected region and
clicked the left mouse button to mark the region.
Then a series of questions and rating scales were
prompted in the same order. The rating process
was as follows: After the observer marked the
suspicious region, the type of abnormality in
question was identified; two “semicontinuous”
(0–100) rating scales (sliders) for the likelihood of
the presence (or absence) of an abnormality and
the likelihood of the abnormality in question
being malignant, if actually present, appeared and
were rated. For each marked abnormality the
observer was asked if the same abnormality was
depicted on the ipsilateral view (yes or no) and if
the reader answered yes, he or she was asked to
mark the location and to independently rate (image
based) the presence and malignancy likelihoods
as depicted on the corresponding view. After com-
pleting the rating of one abnormality, the reader
could mark and rate additional abnormalities
shown on the same examination as deemed
appropriate. If no abnormality was detected on
an examination, the reader could just click on the
“Done” button at the bottom of the display.

After rating all suspected abnormalities, the
observer was asked to provide his or her recom-
mendation as if the examination in question was a
first screening examination by using BI-RADS
(i.e., 0 for recall, 1 for negative, and 2 for benign
findings). When this task was completed in mode
1 (FFDM), the next examination appeared on the
workstation. However, in the other two modes
(frames and digital breast tomosynthesis) the last
task before moving to the next case was to rate the
examination using a five-category rating scale as
significantly better, somewhat better, comparable,
somewhat worse, or significantly worse than a
high-quality FFDM examination for interpreting
the examination in question. At any time during
the interpretation of an examination the observer
could edit, remove, or add marks as deemed
appropriate. The total time that a screen was
displayed for each examination was automatically
recorded and serves as an estimate of the time the
observer spent viewing, interpreting, and rating
the examination.

We recognize that the extensive reporting
structure in this small, pilot study introduced a
substantial complexity, but we deliberately chose
this reporting structure primarily for two reasons.
First, this reporting structure allows different
analyses to be performed and the results can be
compared. Second, this was a pilot study in
preparation for a larger one and the multiplicity of
the reporting of different aspects related to this
general problem will allow us to better plan future
studies in this area [8].

Data Analyses

Target definition to determine if a mark was a
true- or false-positive was performed for the
different modes as follows. For the FFDM mode a
circular “acceptance target” for any distance less
than 200 pixels in diameter (on the display)
between the centers of the marked abnormality in
the truth file and the actual mark was established.
In the frames mode, a similar circular target was
established but marks on different frames within a
circle that moved along the imaging arc was
established so that marks on different frames
could be counted correctly when accounting for
the different imaging views (a donut-shaped
acceptance target). For the digital breast tomo-
synthesis mode, a cylindric target with 200 pixels
in diameter and 21 (center ± 10) slices deep was
established as the acceptance target. However,
because of the difference in targets that we
investigated, the sensitivity of the acceptance
target size in all modes on the study results by
systematically increasing or decreasing the target
in all dimensions. When observers marked an
abnormality within the acceptance target, it was
considered a correct response. If a mark was
outside the target, it was considered a false-
positive identification.

Other Data Analyses

The time (in minutes) to review, interpret, and
rate the examinations was averaged for each reader
over all examinations; each mode over all readers
and examinations; and disease status (i.e., malignant
or nonmalignant) over all readers, modes, and
examinations. Times exceeding 8 minutes were
excluded from the analyses on the basis of the
assumption that these excessively long times were
the result of interruptions during the session. As a
result, 3% of all examinations (26 of 810) were
excluded from the time analyses. The mean time
for each reader, mode, and disease status was
compared by using the method of generalized
estimating equations [9], where the mean time
outcome variable was replicated over cases and
the independent variables included mode, reader,
and disease status. Significance levels were com-
puted using Proc GENMOD software (version 9.1,
SAS Institute), which takes into consideration
correlations arising because the same examina-
tions were scored in each mode. A p value of < 0.05
was considered to be statistically significant.

We computed the frequency and proportion of
each subjective rating for each reader over all
examinations and for each mode (frames and digital
breast tomosynthesis) over all readers and exami-
nations. To test whether different radiologists
perceived the tomosynthesis-based image set and
the projection series (i.e., frames) to be better than
FFDM, we combined the categories somewhat
ter better and significantly better into the category
good. Then, a one-sample test for a binomial
proportion assuming normal approximation was
performed for each reader in each mode comparing
the proportion of examinations rated as better to the
probability that the examinations are the same (p =
0.5) for interpreting an examination. A one-sided
p value of < 0.05 was considered to be statistically
significant for each comparison and was calculated
using Proc FREQ software (SAS Institute).

The frequency and proportion of examinations
recalled based on the BI-RADS ratings were com-
puted over all examinations for each reader in each
mode and over all examinations and readers for each
mode for malignant and nonmalignant examinations.
The proportion of examinations recalled across
modes was compared for both malignant and non-
malignant examinations by using a repeated logistic
regression model. For this analysis, the BI-RADS
ratings 1 and 2 were combined to represent those
examinations not recalled; therefore, the binary
outcome variable (recalled or not recalled) was
replicated over cases and the independent variables
included reader and mode. Estimation was done
by using a generalized estimation equation ap-
proach [9], and significance levels were computed
using Proc GENMOD software, which takes into
consideration, correlations arising because the same examinations were scored in each mode. A p value of < 0.05 was considered significant.

The observer performance values (i.e., figure-of-merit [FOM]) for the three modes and nine readers by examination for the detection and classification of either a mass or microcalcification cluster were compared. We performed this analysis by using the jack-knife free-response receiver operating characteristic (JAFROC) method and software of Chakraborty and Berbaum [10] and Chakraborty [11], which is a parametric method combining elements of free-response receiver operating characteristic and the multiple-reader, multiple-case method of Dorfman et al. [12]. All true-positive ratings (i.e., ratings associated with a marked abnormality within the acceptance target as defined) and all false-positive ratings (i.e., ratings associated with a marked abnormality outside the target) were used for each examination.

Results

The mean time spent in reviewing, interpreting, and rating the examinations varied for different readers and modes. For different readers the individual mean times ranged from 0.94 ± 0.67 minutes to 3.78 ± 1.82 minutes. The mean times in minutes over all readers and examinations were 1.58 ± 1.07, 2.03 ± 1.18, and 2.72 ± 1.44 minutes for the FFDM, frames, and digital breast tomosynthesis, respectively. For disease status, the mean times in minutes to review, interpret, and rate the examinations over all readers, modes, and examinations were 2.03 ± 1.31 minutes for nonmalignant examinations and 2.52 ± 1.37 minutes for malignant examinations. The mean time spent in reviewing, interpreting, and rating the examinations was found to be significantly different for different readers (p = 0.0009) and modes (p < 0.0001), but not for disease status (p > 0.05) by using a generalized estimating approach where mean time over all examinations was compared for modes, readers, and disease status. We note that some of the time spent was associated with the nature of the study that required independent reporting of each examination as both a screening one and a diagnostic one.

Table 2 summarizes the distribution of the subjective category ratings for the two modes, frames, and digital breast tomosynthesis, over all readers and examinations. Recognizing that this is a subjective, hypothetic assessment, none of the nine readers perceived the frames or the digital breast tomosynthesis reconstructed image sets to be significantly worse than FFDM images in any one of the 30 examinations (0/30). Using the one-sample binomial test normal approximation for each reader, three of nine readers perceived the frames to be significantly better than the FFDM examination (p < 0.0001, p = 0.005, and p = 0.0053), and six of nine readers perceived digital breast tomosynthesis to be significantly better than the FFDM examination (four p values < 0.0001, p = 0.0053, and p = 0.0142).

Table 3 summarizes the overall detection rates for each mode and the overall proportion of nonmalignant examinations recalled for each mode. The proportion of malignant examinations recalled (i.e., detection rate) varied for different readers for the frames ranging from 55% (six of 11) to 100%. For the digital breast tomosynthesis examinations, one reader detected 82% (nine of 11) of the examinations depicting malignant abnormalities, four readers detected 91% (10 of 11), and the remaining four readers detected 100% (11 of 11). For the FFDM examinations, one reader detected 73% (eight of 11) of the malignant examinations, six readers detected 91% (10 of 11), and the remaining two readers detected 100% (11 of 11). The results were not significant (p > 0.05) as to the effect of different reading modes on the proportion of either malignant and nonmalignant examinations recalled.

The mean FOM and 95% confidence limits were 0.56 (0.41, 0.69), 0.62 (0.55, 0.68), and 0.60 (0.50, 0.70) for modes frames, digital breast tomosynthesis, and FFDM, respectively. The results were not significant (p > 0.05) as to the effect of different reading modes on observer performance. Variation in acceptance target sizes affected somewhat the measured performance and computed FOM, but all trends and statistical tests resulted in similar findings.

Discussion

It is clear to all involved in this area that visualization tools will need to be developed to allow an efficient clinically acceptable assessment of the multiple images generated by this unique imaging approach. Currently, extensive work is being done in this area to address both the viewing of masses and microcalcification clusters and ours is but a limited preliminary study attempting to assess several related issues. The use of CAD [5, 13, 14] may also be a factor in this setting.

## Table 2: Distribution of Subjective Ratings for Frames and Digital Tomosynthesis Reconstructed Image Sets, as Compared with Full-Field Digital Mammography Over All Readers

<table>
<thead>
<tr>
<th>Rating</th>
<th>Frames</th>
<th>Digital Breast Tomosynthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Percent</td>
</tr>
<tr>
<td>Significantly worse</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Somewhat worse</td>
<td>12</td>
<td>4.4</td>
</tr>
<tr>
<td>Comparable</td>
<td>110</td>
<td>40.7</td>
</tr>
<tr>
<td>Somewhat better</td>
<td>131</td>
<td>48.5</td>
</tr>
<tr>
<td>Significantly better</td>
<td>17</td>
<td>6.3</td>
</tr>
</tbody>
</table>

## Table 3: Proportion of Examinations Over All Readers That Were Recalled by Observers

<table>
<thead>
<tr>
<th>Mode</th>
<th>Nonmalignant Examinations</th>
<th>Malignant Examinations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Percent</td>
</tr>
<tr>
<td>Frames</td>
<td>111</td>
<td>64.9</td>
</tr>
<tr>
<td>Digital tomosynthesis</td>
<td>107</td>
<td>62.6</td>
</tr>
<tr>
<td>Full-field digital mammography</td>
<td>110</td>
<td>64.3</td>
</tr>
</tbody>
</table>
Digital Breast Tomosynthesis

in particular as related to the detection of microcalcification clusters on digital breast tomosynthesis examinations that may include as many as 320 slices (4 views x 80 slices). Hence, CAD and other efficiency-enhancing visualization tools will be of particular importance if tomosynthesis-generated images are to be routinely used in the screening environment.

Our experience to date indicates that the appropriate, accurate, and efficient use of tomosynthesis will necessitate substantial training not only in the appearance of different abnormalities but also in the widely varying appearances of normal tissues leading to negative findings. As evidenced from this and other studies [15], even those who are familiar with the procedure and have substantial experience in viewing tomosynthesis examinations may have difficulty with a fraction of the actually negative cases, thereby generating recommendations for recall in “other sites.” For this study, clearly the case selection was heavily enriched with positive examinations by design and this affected the overall recall rate in the study including the recall rate in the actually negative cases (34%). However, both the lack of substantial training and the absence of prior examinations may have contributed to the exceptionally high recall rate as well. The amount of training required for efficient and accurate use of tomosynthesis for breast imaging has yet to be determined and is beyond the scope of this very preliminary project.

Noticeably in this study, observers did not perform as well when interpreting examinations in the frames mode of viewing. A reason for this may be the fact that each of the frames is acquired at a small fraction of radiation dose (~ 10% of FFDM); therefore, more noise is seen in the image making it more difficult to read, in particular as related to microcalcifications. The reason we found no significant difference in the performance value of the frames mode despite the lower absolute FOM may be primarily due to the small sample size of the study, which also precluded any rigorous analysis of subsets of examinations. A larger sample size will be needed for future studies assessing improvement, or decline, in actual diagnosis for either of the tomosynthesis modes.

Other important issues, such as the need for two-view tomosynthesis procedures versus one (e.g., MLO only) (Rafferty EA et al., presented at the 2004 and 2006 Radiological Society of North America meetings), the optimal acquisition techniques such as the optimal arc, the dose per projection image, and the number of projection images, will eventually be addressed. Because of the large number of possible approaches to acquisition, reconstruction, and display of digital breast tomosynthesis examinations, the results of this pilot study may be limited to but one specific approach.

In summary, in the digital era, tomosynthesis may have great potential in screening and diagnostic breast imaging practices and other procedures, and initial results are certainly encouraging. However, further work is needed before this imaging approach finds its optimal role in the clinical environment. Our preliminary study adds but one piece to this very important puzzle.

Acknowledgments

We thank Hologic, Inc., for providing some of the cases used in this study and John Drescher and Glenn Maitz for their diligent work on this project.

References