Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: an eye tracking study

Avi Aizenman
Trafton Drew
Krista A. Ehinger
Dianne Georgian-Smith
Jeremy M. Wolfe

Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: an eye tracking study

Avi Aizenman,^{a,*} Trafton Drew,^b Krista A. Ehinger,^c Dianne Georgian-Smith,^d and Jeremy M. Wolfe^{d,e}

^aUniversity of California, Vision Science Department, Berkeley, California, United States

^bUniversity of Utah, Psychology Department, Salt Lake City, Utah, United States

^cYork University, Centre for Vision Research, Toronto, Ontario, Canada

^dBrigham and Women's Hospital, Surgery Department, Boston, Massachusetts, United States

^eHarvard Medical School, Ophthalmology and Radiology Department, Boston, Massachusetts, United States

Abstract. As a promising imaging modality, digital breast tomosynthesis (DBT) leads to better diagnostic performance than traditional full-field digital mammograms (FFDM) alone. DBT allows different planes of the breast to be visualized, reducing occlusion from overlapping tissue. Although DBT is gaining popularity, best practices for search strategies in this medium are unclear. Eye tracking allowed us to describe search patterns adopted by radiologists searching DBT and FFDM images. Eleven radiologists examined eight DBT and FFDM cases. Observers marked suspicious masses with mouse clicks. Eye position was recorded at 1000 Hz and was coregistered with slice/depth plane as the radiologist scrolled through the DBT images, allowing a 3-D representation of eye position. Hit rate for masses was higher for tomography cases than 2-D cases and DBT led to lower false positive rates. However, search duration was much longer for DBT cases than FFDM. DBT was associated with longer fixations but similar saccadic amplitude compared with FFDM. When comparing radiologists' eye movements to a previous study, which tracked eye movements as radiologists read chest CT, we found DBT viewers did not align with previously identified “driller” or “scanner” strategies, although their search strategy most closely aligns with a type of vigorous drilling strategy. © 2017 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.4.4.045501]

Keywords: perception; eye tracking; tomography; mammography.

Paper 17239R received Aug. 10, 2017; accepted for publication Oct. 2, 2017; published online Oct. 27, 2017.

1 Introduction

In recent years, breast cancer screening has relied upon full-field digital mammography (FFDM) as the main method for detecting breast cancer. FFDM largely replaced earlier film-based methods. In screening, FFDM can help identify cancerous regions before they become symptomatic. With early treatment, clinical interventions are more likely to succeed.¹ Despite the benefits of FFDM, it has received criticism for excessive false positive and false negative findings, around 7.5% to 14% and 30%, respectively.^{2,3} False positive diagnoses typically lead to further testing and can lead to unnecessary treatment. This is particularly true for women with dense breasts (increases in fibro-glandular tissue), who are more susceptible to false positive errors as overlapping dense tissue can imitate the appearance of cancer in traditional radiological images.⁴ In addition, dense tissue can occlude true cancer, resulting in false negative errors which delay possible treatment.

Digital breast tomosynthesis (DBT) is a relatively new imaging modality that has the potential to ameliorate some of the challenges radiologists face when screening for breast cancer.⁵ DBT creates images that are virtual “slices” through the breast. This reduces potential occlusion from overlapping tissue by creating a stack of two-dimensional (2-D) images

that radiologists can view by scrolling in depth. This stack-viewing mode is typical when radiologists are reading other volumetric images, such as CT scans and MR images. The ability to examine a suspicious finding with less obscuration by summation tissues from above and below probably explains the decrease in false positive rates seen with DBT.^{5–8} DBT also improves cancer detection rates⁶ though at a substantial cost in the time required per case.⁸

A typical FFDM exam contains four images, two 2-D images of each breast that the radiologist reads. A DBT exam typically has those four 2-D images to examine (either FFDM or a version synthesized from the DBT data⁹). In addition depending on slice thickness and the thickness of the breast, the DBT exam can have up to 80 more images for the radiologist to examine. Both FFDM and DBT may also include images from one or more prior exams. As FFDM has been the primary modality for screening for breast cancer, reading DBT presents a challenge to radiologists as they are forced to alter their accustomed search strategies. There is extensive literature centered on understanding how radiologists search 2-D medical images, such as FFDM^{10–13} and chest radiographs.^{14–17} Recording the eye movements of radiologists has proven particularly useful as a window into their search strategies. However, despite the increase of volumetric images, such as DBT in the radiology reading room, there is much less work examining the pattern

*Address all correspondence to: Avi Aizenman, E-mail: avigael_aizenman@berkeley.edu

of eye movements of radiologists and best practices for reading volumetric radiological images.

When reading a DBT exam, radiologists are forced to develop approaches to reading a stack of image slices that represent the 3-D breast. Understanding the basic eye movement patterns of radiologists while they read DBT and FFDM can help shed light on how radiologists accomplish this task with the eventual goal of establishing “best practice” approaches to this type of image. Although DBT has been shown to be more accurate than FFDM screening, there has been a lag in clinical usage. One reason for this is that radiologists take at least two times longer to read and interpret a DBT exam as compared to an FFDM exam.^{8,18–22} Eye movement patterns can provide insight into the use of that extra time. There is no evidence-based consensus on the optimal strategy for reading DBT because there is not much evidence to date. Indeed, the only published studies examining eye movements in DBT appear to be Timberg et al.²³ and Jiang et al.²⁴ Timberg et al.²³ eye-tracked observers while they read DBT cases with masses and calcifications inserted. This study tested only four observers and although they analyzed dwell time, Timberg et al. did not specifically look at eye movements or scan path data. Jiang et al. eye-tracked three physicist observers while they read synthesized DBT cases, and found that dwell time was highly correlated with feature values from an adaptive matched filter. Details of the eye movements and scan paths of radiologist observers reading real DBT images were not the topic of these papers.

Drew et al.²⁵ have used eye-tracking data to assess search strategies for radiologists searching for lung nodules in volumetric lung CTs. They found that radiologists fell into two broad categories of search strategy. “Drillers” kept their eye position limited to one quadrant of the lung at a time while scrolling briskly through the stack of image slices before moving on to search another quadrant. In contrast, “scanners” searched broadly across an image before scrolling more slowly through the stack of image slices. Drillers overall seemed to have a higher nodule detection rate and covered more of the volume of the lung in their search when compared with scanners. However, the sample was too small to be sure if this difference is a general rule for search strategy during lung cancer screening or specific to the group sampled in the study.

What strategy or strategies do radiologists use when examining volumes of DBT images? The radiologist observers who participated in the current experiment generally reported using a driller search strategy when reading DBT. However, humans are not very good at consciously monitoring their own eye movements.^{26,27} Accordingly, it would be interesting to know if clinicians are, in fact, following those driller instructions. In the current study, we used eye tracking to quantify and compare the basic pattern of eye movements adopted by radiologists searching DBT versus FFDM, and to determine how DBT search compares to searching chest CT cases.

2 Materials and Methods

2.1 Participants

This study was carried out roughly 1 year after Brigham and Women’s Hospital began routinely using DBT. On average, radiologist observers had more experience with FFDM (6 years) than with DBT (0.62 years). Eight radiologist observers were American Board of Radiology certified, while three were

not. We eye-tracked 11 radiologist observers (average years of age 37 years) as they read DBT and FFDM cases. The radiologists were recruited from Brigham and Women’s Hospital, and the Brigham and Women’s Hospital Institutional Review Board approved all experimental procedures.

2.2 Procedure and Method

Radiologist observers read nine DBT and eight FFDM images (including one practice DBT trial). In an effort to ensure that the cases were equally difficult, the DBT and FFDM cases were taken from the same patient. One image type was mirror reversed to decrease the chances that a radiologist would remember the case from its first appearance. Thus, the same breast might be viewed as a left breast in the FFDM mediolateral-oblique (MLO) view and then mirror-reversed to be the right breast in DBT. In standard practice, a radiologist would have access to both 2-D and 3-D volumetric images for a case. In this experiment, we restricted viewing to just one type of view on each trial to focus on how performance and eye movements differed across images from the same cases. All images were presented in one view, either cranial-caudal (CC) or MLO. For each case, the FFDM and DBT views were both CC or MLO. The order in which the cases were viewed was randomized across radiologist observers. In half of the cases, there was a positive finding, a mass or area of architectural distortion that was sufficiently suspicious to warrant follow-up testing. Positive images could contain more than one suspicious finding. Two cases had 1, one case had 2, and one case had 3. In the other half of the cases—the negative cases, there were no findings sufficiently suspicious to warrant follow-up testing. Calcifications were excluded from both positive and negative cases. Ground truth was determined by Georgian-Smith. He also provided density ratings on the following scales: A—fatty tissue, B—scattered fibroglandular densities, C—heterogeneously dense, and D—extremely dense. Table 1 shows details for each case. The nine DBT cases contained 54 to 78 slices with a 1-mm thickness on average. Contrast levels of all cases were fixed.

Images were displayed on a workstation with a 5 K monitor with minimum luminance of 1.07 candela/m² and maximum

Table 1 Details of cases used.

Case number	Case details	
	Number of masses	Density
1	2	C
2	3	C
3	1	A
4	1	C
5	0	A
6	0	C
7	0	C
8	0	C

Fig. 1 An example of the display shown on each trial of the experiment and examples of positive cases from four different trials. DBT slices shown correspond to the slice where the mass was most conspicuous. Please see the [Appendix](#) to see all cases that were presented to radiologists with suspicious findings labeled.

luminance of 451 candela/m². Thereby fulfilling the mammography quality standards act and program standards for presentation. Ambient light levels were similar to Brigham and Women's Hospital's clinical radiology reading rooms. Observers were given 2 min per case to mark any clinically significant lesions warranting further diagnostic workup. Radiologists were instructed to ignore calcifications and to search for masses or architectural distortions. Radiologists scrolled through the stack of image slices using the up and down arrows on a keyboard. Typically, radiologists use a scroll wheel to move through the stack of image slices. As with a scroll wheel, our readers had freedom over the speed and direction with which they moved through the stack. Radiologists marked any questionable masses or architectural distortions they localized with a mouse click. Prior to the start of the experiment, the experimenter explained the task as well as the navigation system and verified that the radiologist understood both. The experiment began with a positive DBT case that was used as a practice trial to familiarize radiologist observers with the overall experimental setup and allow them to practice navigating through the stack of image slices. This initial trial was not included in analyses. Prior to completing the experiment, radiologist observers were asked to fill out a survey with detailed questions about their experience reading radiological images. Observers were also asked to free write what strategy they thought they were using to read the DBT cases, specifically with regard to their eye movements. Drew et al.²⁵ later classified each radiologists' search strategy.

Figure 1(a) shows the display screen from an individual trial in the experiment. There are a few differences between the

experimental setup used and the workflow found in the typical clinical environment. Radiologists were shown only one view of a case (e.g., CC or MLO) in a trial. They did not have access to a four-view display of the DBT or FFDM cases, nor did they have access to any prior exams or case histories. Radiologists were unable to control image properties, such as zoom and window/level. Figure 1(b) shows examples of positive cases from four separate trials. The [Appendix](#) shows all cases presented to radiologists with masses and architectural distortions labeled.

Radiologists were seated in a darkened room 62 cm away from a 53-cm GE SMD 21500 LCD monochrome monitor with a resolution of 2048 × 2560 pixels. At a 62-cm viewing distance, the 53-cm diagonal of the monitor subtended about 46 deg of visual angle. The observer's head rested on a chin rest to reduce head movement and minimize eye-tracking error. The images were displayed at 1996 × 2457. The experiment was executed using Psychtoolbox and the Eyelinktoolbox in MATLAB.^{28,29}

Eye position was tracked using a desktop-mounted Eyelink 1000 (SR Research, Ontario, Canada) which sampled the *x* and *y* positions of one eye at 1000 Hz. We used a nine-point calibration procedure for each radiologist observer both before and during the experiment. Changes in DBT depth/slice were also recorded, and used offline to generate a depth value (*Z*) at each time sample. This allowed us to recreate eye movement scan paths that included radiologists' movement in depth. The requirements of eye tracking meant that observers could not move closer to the image to scrutinize a suspicious region, nor did they have the normal capabilities to magnify the image, to

Fig. 2 Mean behavioral performance as a function of image modality (DBT or FFDM). (a) Shows sensitivity and (b) shows false positive rate for DBT and FFDM. Error bars shown here and throughout the paper are standard error of the mean.

compare the current image with a previous image from the same patient, or to change the windowing and leveling of the image.

Eye tracking in 3-D volumes of images presents some challenges for analysis. Traditionally, the 2-D search literature categorizes each recorded eye position as part of a fixation or a saccade. In a 3-D stack of images, these categorizations are complicated by radiologists' ability to navigate in depth. For example, radiologists may make smooth pursuit movements while tracking features that change position with depth; these movements cannot technically be categorized as either a fixation or a saccade.³⁰ For present purposes, we identify fixations and saccades using the EyeLink parser, and tracked which depth levels were viewed during each fixation to recreate the 3-D scan paths. We used EyeLink default values for acceleration and velocity thresholds: 8000 deg/s² and 30 deg/s, respectively.

3 Results

3.1 Digital Breast Tomosynthesis Versus Full-Field Digital Mammography

3.1.1 Detection performance

Figure 2 shows the basic sensitivity and specificity data per finding. Note that the term "sensitivity" refers to different quantities in the medical and vision sciences literature. We are using the definition from the medical literature where "sensitivity" refers to the true positive rate. In the vision science literature, "sensitivity" is used as a term for signal detection figures of merit (e.g., d' or the area under a receiver operating characteristic curve). Figure 2(a) shows radiologists had higher sensitivity when using DBT than FFDM images. They were significantly better at detecting masses and architectural distortions in DBT [$t(10) = 2.55$, $p = 0.03$].

Also replicating previous findings,⁵⁻⁸ radiologists had higher specificity with DBT than FFDM. Figure 2(b) plots the false alarm rate (1 – specificity) and shows significantly fewer false positive errors in DBT cases than in FFDM images [$t(10) = 2.75$, $p = 0.02$].

As shown in Fig. 3, DBT cases took significantly longer to search than FFDM [$t(10) = 6.136$, $p < 0.0001$], probably because DBT cases comprise many more images than FFDM. This finding replicates previous studies on search time and DBT.⁸ Radiologists were given 2 min to search either DBT or FFDM cases, but the vast majority of trials (95.5%) were

Fig. 3 Search duration by image modality. DBT takes significantly longer to search than FFDM. Trials that timed at 2 min were excluded.

completed by the observer before the 2-min time point at which point the trial would have automatically ended. Across observers, 7% of trials were ended during the last 10 s of the trial, and trials that timed out after 2 min were excluded. The results could be described as a speed-accuracy trade-off in which the longer time spent with the DBT case produces better performance. However, earlier results indicate that just adding time does not improve performance in non-DBT mammography.³¹

3.1.2 Eye tracking results: fixation duration and saccade amplitude

As shown in Fig. 4, if we compare the standard measures of fixation duration and saccade amplitude between FFDM and DBT, we find that radiologists make significantly longer fixations while searching DBT cases (mean: 405 ms) than FFDM [mean: 296 ms, $t(10) = 6.88$, $p < 0.0001$], whereas saccadic amplitudes were similar between imaging modalities [DBT: 3.61, FFDM: 3.72, $t(10) = 1.282$, $p = 0.22$]. On average, radiologists scrolled through 3.8 image slices during a single fixation while reading DBT. The longer fixation durations in DBT reflect fixation in the XY plane, while observers move in Z (depth).

3.1.3 Coverage analysis

The advent of volumetric images vastly increases the available area that radiologists could search through and assess. What percentage of the available imagery do radiologists actually

Fig. 4 Eye-tracking measures by imaging modality. (a) Shows the distribution of fixation durations, and *T*-tests on average fixation duration found that radiologists have significantly longer fixations while reading DBT. (b) Shows the distribution of saccadic amplitudes, and that the average saccadic amplitude is similar between DBT and FFDM.

cover during their search? Work by Kundel et al.³² estimates the coverage in 2-D images by assuming that a 5-deg-diameter circle, centered at the point of fixation, is processed on each fixation. This estimated radius almost certainly varies based on the complexity of the image and should be empirically verified, but it is a useful starting point in the absence of other data examining this question for DBT and FFDM. As such, we calculated coverage for DBT and FFDM cases while estimating the useful field of view (UFOV) for each fixation to be 1, 2, 3, 4, or 5 degrees visual angle (DVA). DVA is the angle a viewed object subtends at the eye, a measurement that takes into consideration the distance the observer is away from the stimulus as well as the monitor size. A complication for calculating coverage in DBT arises through the volumetric nature of DBT cases. As radiologists fixate in one image level, they are getting some amount of information about the image slices above and below the currently fixated slice as slices are not completely independent. Similar to images in the natural world, pixels in radiological images are highly dependent on neighboring pixels and less dependent on pixels that are farther away. Variability in image slice thickness, shape, and conspicuity as well as density and location of a potential mass make it challenging to estimate what an appropriate coverage window in depth should be. To estimate coverage for DBT cases, we used a matrix of *XY* fixation values, and for each *XY* pair we recorded which depth levels had been visited during the duration of the fixation. For each *XY* fixation pair, we “painted” circles over the DBT image slices at the location that had been fixated for each depth level viewed. We also painted a circle above and below the highest and lowest depth levels visited per fixation to account for the fact that slices are not completely independent and that radiologists may be processing information about slice(s) both above and below the currently viewed slice due to the similarity between neighboring pixels, even in depth. We repeated this procedure for UFOVs 1 to 5 DVA, and divided the fixated tissue by the total tissue to get estimates of coverage for DBT. The results of this coverage analysis can be seen in Fig. 5. It is apparent that the estimate of coverage is higher for FFDM than for DBT regardless of the size of the hypothetical UFOV.

3.1.4 Classifying search strategy

When asked to introspect on what search strategy they used to read DBT, 10/11 of the radiologist observers from the

Fig. 5 Percent coverage for DBT and FFDM. The diameter of the circle used to represent the UFOV around each fixation was systematically varied to be 1 to 5 DVA.

experiment described restricting their eye movements to a region of breast tissue, and scrolling through depth while keeping their eye movements isolated to one region at a time. This strategy is consistent with the drilling strategy described by Drew et al.²⁵ in their 2013 study of eye movements using lung CT images. The 11th observer was unable to articulate what their strategy was in terms of eye movements. Upon further questioning of all the radiologist observers, they described either learning this drilling strategy from their mentors (6/11 radiologist observers) or from DBT training workshops (4/11 radiologist observers).

To verify the extent to which our radiologists were doing what they thought they were doing and what they were taught to do, we compared the 3-D scan paths of our DBT observers to the data from Drew et al.’s study.²⁵ In this experiment, 23 radiologist observers read chest CT cases while their eye movements were tracked. Drew et al.²⁵ found that radiologists fell into two camps: they were either drillers who restricted their eye movements to one quadrant of the *XY* plane while they quickly scrolled through *Z*/depth or scanners who scanned widely over the *XY* plane while moving more slowly in *Z*.

Do the eye movements of the radiologists viewing DBT resemble those of chest CT scanners or drillers or do they look different from either of these? To assess this, we analyzed the DBT data and reanalyzed the chest CT data using a “boxcar” averaging method in which we compared movements in *XY* to

Fig. 6 Number of image depth levels viewed as a function of scan path normalized to distance travelled in 5 s. Chest CT drillers = red, lung CT scanners = green, and DBT = blue. Each point represents a trial. Lung CT observers were classified as drillers or scanners in Drew et al.'s work.²⁵

movements in Z for a succession of 5-s windows, each of which started with a fixation. For each fixation, we measured the length of the scan path between that fixation and the first fixation that occurred at least 5 s later. We also measured the maximum range of slices viewed in that time window. We repeated this for each fixation. This process averages data over windows in time that are a bit longer than 5 s (because the last fixation does not occur at exactly the 5-s mark). Accordingly, the XY scan path and the Z slice measure are corrected by multiplying these measurements by 5 s/actual duration. We removed from the set of fixations those that fell outside the lung or the breast. For instance, if the observer fixated on the controls, it would not make sense to include that long excursion away from the lung or the breast in the calculation of the length of the scan path.

Since the final fixations in a 5-s window occurred at different precise lags from the first fixation in the window, we normalized the scan path lengths by dividing by the actual time between first and last fixations. To determine how radiologists were moving in depth during each 5-s window, we subtracted the smallest depth level from the largest depth level viewed giving the total number of image slices viewed. Different cases have different numbers of slices. It is not immediately clear if the best way to compare movements in depth is to express depth in number of slices or as a percentage of the total stack of images. We will present both analyses.

In Fig. 6, the graph scatterplots normalized travel in XY on the x-axis and the absolute number of slices traversed on the y-axis. Each data point represents one case for one observer. All of the ~5-s intervals from a case are averaged into a single point. Red data points are from chest CT observers classified as drillers by Drew et al.²⁵ Green is chest CT "scanners." This was a smaller group in the Drew et al. paper. Blue data points are from the DBT observers.

Turning first to the previously reported lung CT data, it is clear that, on a trial by trial basis, there is significant overlap between observers classified as drillers and those classified as scanners. It is also clear that the two groups differ in this analysis in the manner that would be predicted. Most drillers show

relatively long movements in Z while moving less dramatically in the XY plane. Scanners, on average, show larger movements in the XY plane. When we look at the DBT results from the same analysis, the data points fall in a completely different part of the space. The scan paths in XY of radiologists reading DBT are markedly longer than the scan paths of drillers or scanners in lung CT. The number of slices traversed is comparable to those of average drillers.

This description of the differences between behavior in the lung CT and DBT situations could be misleading because there are systematic differences between the lung CT and DBT images. On the Z-axis, there are many more image layers in lung CT (121 to 290 slices) than DBT (52 to 79 slices). In XY, the image size in lung CT (500 × 500 pixels or 16.97 DVA) is smaller than in the DBT images (1996 × 2457 pixels or 46 DVA). Accordingly, we have replotted the data using normalized distances in XY and Z. To normalize the scan paths, we began by calculating the maximum and minimum in fixation position in X and Y. This defines a box that contains all the eye movements for a specific trial. We normalized the scan path by dividing the scan path over 5 s by the diagonal of that box. XY scan paths are divided by the length of the maximum possible eye movement in the XY plane. Note that this normalized scan path value can be greater than one because the sum of all saccades in 5 s can be much longer than the single longest distance in the image. The movements in Z were normalized as the fraction of the total stack. The results of these normalizations are shown in Figs. 7(a) and 7(b).

The scatterplots look different from the data that has not been normalized. In lung CT, drillers and scanners continue to produce different if overlapping distributions of trials. Scanners, on average, have longer normalized scan paths in XY. Drillers, on average, make larger movements in Z. A MANOVA over image type (DBT, drillers and scanners) with two dependent variables (normalized XY and percent Z covered) showed a significant main effect of image type $F(2,31) = 22.55$, $p < 0.001$. *Post hoc t*-tests suggest that for the normalized XY scan path data the DBT trials are significantly different from scanners [$t(14) = 3.32$, $p < 0.01$], but not significantly different from drillers [$t(27) = 0.36$, $p = 0.72$], while drillers and scanners are significantly different from each other [$t(21) = 3.69$, $p = 0.001$]. On the normalized Z-axis, DBT observers move through a much larger percentage of the stack of images than do lung CT drillers [$t(27) = 10.85$, $p < 0.0001$] and scanners [$t(14) = 10.37$, $p < 0.001$], and drillers cover a larger percentage of the stack of image slices than do scanners [$t(21) = 2.69$, $p = 0.01$]. How should we characterize the DBT eye movement data? Neither the normalized or unnormalized method of plotting the data can be considered to be unequivocally correct. It may be difficult to directly compare the depth information in DBT and chest CT since they differ in how suspicious features are seen across slices. Moreover, since the DBT images are four times larger than the chest CT cases, it seems misleading to compare the raw XY values. Normalizing eye movements by image size (as in Fig. 7) may be a fairer comparison. As they were taught, DBT viewers do drill through the depth of the breast quite quickly, generally covering more depth levels in a 5-s window than drillers, with somewhat reduced movements in the XY plane that are similar to what is seen with drillers in chest CT. Nevertheless, it is clear from the XY data that viewers do not hold their eyes particularly steady in XY while drilling in Z.

Fig. 7 Percent of the total stack of image depth levels viewed as a function of normalized scan path. (a) Each point represents a trial and (b) each point represents one subject.

Fig. 8 Depth as a function of time in seconds for representative individual trials. Red plots show driller data, green plots show scanner data, and blue plots show DBT data. Each row corresponds to a different radiologists' data. Please see Ref. 33 for graphs of the complete set of data.

To get a better sense of between trial variability within subjects, we created depth by time plots for each trial. A sample of these can be seen in Fig. 8 (the full set is found at Ref. 33). The lung CT data show the mixing of categories that we also see in Figs. 6 and 7. While the data show that the groups differ overall, on any given case a radiologist categorized as a driller might scan and a scanner might drill. The DBT data have a different quality. The overall impression is of many more sawtooth plots as the observer moves back and forth through the stack multiple

times. While they are doing this, they do not move their eyes in XY space as much as the scanners in our previous dataset even if they are not fixated on one spot.

4 Discussion/Conclusion

With advances in technology, the number, the size, and intricacy of medical images have been increasing.³⁴ While there is a significant literature on how radiologists examine 2-D medical images, such as chest x-rays or mammograms, little is known

about how radiologists approach reading a volumetric image such as chest CT and even less is known about how radiologists read DBT. For radiologists who are accustomed to reading static 2-D FFDM images, the depth dimension of DBT significantly changes the process radiologists go through to detect suspicious findings. Because DBT is a relatively new imaging modality, this is one of the first studies to examine basic eye movement patterns as radiologists read DBT.^{23,24} By tracking the eye movements of radiologists while they search DBT and FFDM, we are able to contrast the two modalities. We are also able to compare search strategies in DBT to those used in lung CT.

4.1 *Digital Breast Tomosynthesis Versus Full-Field Digital Mammograms*

Our eye-tracking data illustrate the differences between DBT and FFDM as oculomotor tasks. As measured in the *XY* plane, fixations are longer in DBT because there are times when the observers move in depth while holding fixation in *XY*. Presumably because there is more to look at, DBT search takes roughly twice as long as FFDM. This extra time pays benefits. In a replication of previous experiments,⁵⁻⁸ our experiment found that DBT improved diagnostic accuracy both by decreasing the number of false alarms, and increasing the detection rate for masses. Given that exams with DBT take longer than FFDM exams, it might be proposed that the observed DBT benefit could be a form of a speed accuracy trade-off. As radiologists are spending more time reading DBT, their search is more thorough and complete than when reading FFDMs. However, as noted earlier, work with non-DBT mammography indicates that adding time does not produce improved performance.³¹ Instead, prior research suggests that improvements in accuracy seen in DBT are due to an improved ability to disambiguate overlapping tissue.⁵⁻⁸ Given the earlier evidence that more time does not necessarily help, future research should test whether time spent with DBT can be reduced without a loss of performance.

Our finding that search coverage is significantly higher for FFDM presumably arises from the increased area that radiologists must search in DBT. Given the differences between 2-D and volumetric stacks of images, this comparison should be viewed with caution. Obviously, coverage is well under 100% even with an optimistic 5 deg UFOV. However, these coverage values are merely descriptive. It is not known how much coverage is enough since there will be parts of the image that do not need to be fixated. A quantitative estimate of enough coverage would require a more strongly supported estimate of the UFOV in mammography and some way to determine what areas of the breast need not be fixated because, for example, they are clearly empty of any features that could be suspicious. Future directions may include exploring to what extent the apparently low coverage rate for DBT may be contributing to missed cancer. Comparing DBT and FFDM is somewhat artificial since the 2-D and 3-D volumetric images are typically used together in clinical practice. A target for future research would be to uncover the oculomotor strategies of combined DBT and FFDM with the goal of finding strategies that allow for the benefits of DBT with less of a cost in time.

There are limitations to the current study. There are only eight cases and observers were presented with only one view (CC or MLO) of one breast in each of those cases. Obviously, this misses the richness of the radiologist's examination of

the whole case. This is merely a first exploration of the search strategies used while reading a single stack of DBT images. It gives some insight into observers' search strategies. In future work, there are a host of questions worth asking. Does a clinician's choice of hanging protocol interact with her way of searching individual images? How does the presence of a 2-D image influence search of the DBT images? How are scan paths affected by factors such as breast density? Perhaps most critically, can we use information from eye tracking to reduce errors and/or time per case in mammography? Additionally, in the current study we asked observers to identify masses, and to ignore calcifications. Search for calcifications would typically involve changing magnification, but we chose not to allow zooming in this study as it complicates the eye movement analysis. However, this is a limitation in the current study that future experiments should address.

4.2 *Comparing Search in Digital Breast Tomosynthesis and Chest CT*

Eye movements during DBT do not appear to fall neatly into the scanning and drilling categories previously identified in the context of chest CT.²⁵ Rather DBT appears to be characterized by vigorous drilling, often at quite a rapid rate combined with some quite extensive scanning. It is notable that, although there was a great deal of variability in how these images were searched across both individuals and trials, there was a strong consensus on the reported strategy of how to most effectively search. All but one of our observers described a strategy that sounded quite similar to the drilling strategy outlined by Drew et al.²⁵

When comparing patterns of search in DBT and lung CT, it is important to note how different these image modalities are. The images typically differ in size. Obviously, the types of suspicious findings and tissues differ in important ways. Lung CT images have a greater diversity of structures (e.g., lungs, arteries, and the esophagus) than breast images. Thus, it is unsurprising that there are differences between the oculomotor behavior seen in the current study and in the Drew et al. data. Nevertheless, the driller/scanner distinction is useful starting point for further assessments of how radiologists search in volumetric images.

As noted DBT is quite a new technology. Radiologists reading lung CT for Drew et al. had significantly more experience reading lung CT than the radiologists from the current experiment had reading DBT (18.8 years versus 0.62 years). It will be interesting to see if oculomotor strategies in DBT change as the field develops more experience. Drew et al.'s data suggested that the drilling strategy may be better than scanning while engaged in a lung cancer screening task. In the current study, there were no clear distinctions in how different radiologists searched the DBT images. Thus, at present it is difficult to be certain what the best practices are in DBT. Perhaps different strategies will emerge as radiologists accrue more experience with these sorts of images. Once the field has matured, it may be useful to measure eye movements as trainees view DBT to help them move toward the best practice (or practices if multiple approaches turn out to be equivalent.).

While subjective report of search strategy is an important starting point for understanding the different strategies for approaching a search space, data from the basic science literature make it clear that humans are often quite poor at knowing where they have or have not looked.^{26,27} Thus, eye tracking is

a vital tool for clearly quantifying the differences between different search strategies and their relationship to resultant behavior, particularly for imaging modalities such as DBT. Moreover, given the dissociation between reported and actual eye movements, eye tracking may prove to be a useful pedagogical tool as radiologists learn how to most effectively search through increasingly complex medical images.

Appendix

All positive cases shown to radiologist observers with suspicious findings marked. Image slices for DBT cases correspond to the slice where the suspicious finding is most conspicuous (Figs. 9–13).

Fig. 9 Positive DBT case 1 shown to radiologist observers. Image slices for DBT cases correspond to the slice where the mass is most conspicuous. This case contained two positive findings.

Fig. 10 Positive DBT case 2 shown to radiologist observers. Image slices for DBT cases correspond to the slice where the mass is most conspicuous. This case contained three positive findings.

Fig. 11 Positive DBT cases 3 and 4 shown to radiologist observers. Image slices for DBT cases correspond to the slice where the mass is most conspicuous. Each of these cases contained one positive finding.

Fig. 12 Positive FFDM cases 1 and 2 shown to radiologist observers. Case 1 contained one positive finding, while case 2 contained three positive findings.

Fig. 13 Positive FFDM cases 3 and 4 shown to radiologist observers. Each of these cases contained one positive finding.

Disclosures

No competing conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

This work was supported by 1F32EB011959-01 from NIH awarded to T.D., R01 EY017001 from NEI and CA207490 from the NCI awarded to J.M.W.

References

1. L. Tabár et al., "Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades," *Radiology* **260**, 658–663 (2011).
2. S. L. Jackson et al., "Are radiologists' goals for mammography accuracy consistent with published recommendations?" *Acad. Radiol.* **19**, 289–295 (2011).
3. S. R. Hoff et al., "Breast cancer: missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography—results from a retrospective review," *Radiology* **264**, 378–386 (2012).
4. P. A. Carney et al., "Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography," *Ann. Intern. Med.* **138**(3), 168–175 (2003).
5. S. Vedantham et al., "Digital breast tomosynthesis: state of the art," *Radiology* **277**, 663–684 (2015).
6. S. M. Friedewald et al., "Breast cancer screening using tomosynthesis in combination with digital mammography," *J. Am. Med. Assoc.* **311**, 2499–2507 (2014).
7. B. M. Haas et al., "Comparison of tomosynthesis plus digital mammography and digital mammography alone for breast cancer screening," *Radiology* **269**, 694–700 (2013).
8. P. Skaane et al., "Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program," *Radiology* **267**(1), 47–56 (2013).
9. S. P. Zuckerman et al., "Implementation of synthesized two-dimensional mammography in a population-based digital breast tomosynthesis screening program," *Radiology* **281**(3), 730–736 (2016).
10. E. A. Krupinski and R. M. Nishikawa, "Comparison of eye position versus computer identified microcalcification clusters on mammograms," *Med. Phys.* **24**, 17–23 (1997).
11. H. L. Kundel et al., "Holistic component of image perception in mammogram interpretation: gaze-tracking study," *Radiology* **242**, 396–402 (2007).
12. C. Mello-Thoms, "Perception of breast cancer: eye-position analysis of mammogram interpretation," *Acad. Radiol.* **10**(1), 4–12 (2003).
13. C. Mello-Thoms et al., "Effects of lesion conspicuity on visual search in mammogram reading," *Acad. Radiol.* **12**(7), 830–840 (2005).
14. K. S. Berbaum et al., "Role of faulty visual search in the satisfaction of search effect in chest radiography," *Acad. Radiol.* **5**, 9–19 (1998).
15. S. M. Ellis et al., "Thin-section CT of the lungs: eye-tracking analysis of the visual approach to reading tiled and stacked display formats," *Eur. J. Radiol.* **59**, 257–264 (2006).
16. H. L. Kundel, C. F. Nodine, and D. Carmody, "Search, recognition, and decision making in lung nodule detection," *Invest. Radiol.* **12**(5), 431–431 (1977).
17. H. L. Kundel, C. F. Nodine, and D. Carmody, "Visual scanning, pattern recognition and decision-making in pulmonary nodule detection," *Invest. Radiol.* **13**, 175–181 (1978).
18. D. Bernardi et al., "Application of breast tomosynthesis in screening: incremental effect on mammography acquisition and reading time," *Br. J. Radiol.* **85**(1020), e1174–e1178 (2012).
19. P. A. Dang et al., "Addition of tomosynthesis to conventional digital mammography: effect on image interpretation time of screening examinations," *Radiology* **270**(1), 49–56 (2014).
20. D. Gur et al., "Digital breast tomosynthesis: observer performance study," *Am. J. Roentgenol.* **193**(2), 586–591 (2009).
21. M. G. Wallis et al., "Two-view and single-view tomosynthesis versus full-field digital mammography: high-resolution x-ray imaging observer study," *Radiology* **262**, 788–796 (2012).
22. M. L. Zuley et al., "Time to diagnosis and performance levels during repeat interpretations of digital breast tomosynthesis: preliminary observations," *Acad. Radiol.* **17**, 450–455 (2010).
23. P. Timberg et al., "Investigation of viewing procedures for interpretation of breast tomosynthesis image volumes: a detection-task study with eye tracking," *Eur. Radiol.* **23**, 997–1005 (2013).
24. Z. Jiang, M. Das, and H. Gifford, "Analyzing visual-search observers using eye-tracking data for digital breast tomosynthesis images," *J. Opt. Soc. Am. A* **34**, 838–845 (2017).
25. T. Drew et al., "Scanners and drillers: characterizing expert visual search through volumetric images," *J. Vision* **13**(10), 3 (2013).
26. A. D. Clarke et al., "People are unable to recognize or report on their own eye movements," *Q. J. Exp. Psychol.* **70**, 2251–2270 (2016).
27. M. L. Võ, A. M. Aizenman, and J. M. Wolfe, "You think you know where you looked? You better look again," *J. Exp. Psychol.* **42**, 1477–1481 (2016).
28. D. H. Brainard, "The psychophysics toolbox," *Spat. Vision* **10**, 433–436 (1997).
29. F. W. Cornelissen, E. M. Peters, and J. Palmer, "The eyelink toolbox: eye tracking with MATLAB and the psychophysics toolbox," *Behav. Res. Methods Instruments Comput.* **34**, 613–617 (2002).
30. A. C. Venjakob and C. R. Mello-Thoms, "Review of prospects and challenges of eye tracking in volumetric imaging," *J. Med. Imaging* **3**(1), 011002 (2016).
31. C. F. Nodine et al., "How experience and training influence mammography expertise," *Acad. Radiol.* **6**(10), 575–585 (1999).
32. H. L. Kundel et al., "Searching for lung nodules: a comparison of human-performance with random and systematic scanning models," *Invest. Radiol.* **22** (5), 417–422 (1987).
33. A. Aizenman et al., Supplementary Graphs, <http://search.bwh.harvard.edu/new/pubs/DBTdepthxtimePlots.pdf> (2017).
34. K. P. Andriole et al., "Optimizing analysis, visualization, and navigation of large image data sets: one 5000-section CT scan can ruin your whole day," *Radiology* **259**, 346–362 (2011).

Biographies for the authors are not available.