What eye tracking can tell us about how radiologists use automated breast ultrasound

Jeremy M. Wolfe[a,b,*] **Wanyi Lyu**[a] **Jeffrey Dong**[c] **and Chia-Chien Wu**[a,b]

[a]Brigham and Women's Hospital, Boston, Massachusetts, United States
[b]Harvard Medical School, Boston, Massachusetts, United States
[c]Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States

Abstract

Purpose: Automated breast ultrasound (ABUS) presents three-dimensional (3D) representations of the breast in the form of stacks of coronal and transverse plane images. ABUS is especially useful for the assessment of dense breasts. Here, we present the first eye tracking data showing how radiologists search and evaluate ABUS cases.

Approach: Twelve readers evaluated single-breast cases in 20-min sessions. Positive findings were present in 56% of the evaluated cases. Eye position and the currently visible coronal and transverse slice were tracked, allowing for reconstruction of 3D "scanpaths."

Results: Individual readers had consistent search strategies. Most readers had strategies that involved examination of all available images. Overall accuracy was 0.74 (sensitivity = 0.66 and specificity = 0.84). The 20 false negative errors across all readers can be classified using Kundel's (1978) taxonomy: 17 are "decision" errors (readers found the target but misclassified it as normal or benign). There was one recognition error and two "search" errors. This is an unusually high proportion of decision errors. Readers spent essentially the same proportion of time viewing coronal and transverse images, regardless of whether the case was positive or negative, correct or incorrect. Readers tended to use a "scanner" strategy when viewing coronal images and a "driller" strategy when viewing transverse images.

Conclusions: These results suggest that ABUS errors are more likely to be errors of interpretation than of search. Further research could determine if readers' exploration of all images is useful or if, in some negative cases, search of transverse images is redundant following a search of coronal images.

 [DOI: 10.1117/1.JMI.9.4.045502]

Keywords: breast ultrasound; eye tracking; medical error; visual search; mammography.

Paper 22068GR received Mar. 18, 2022; accepted for publication Jul. 8, 2022; published online Jul. 26, 2022.

1 Introduction

From the vantage point of the study of medical image perception, it is useful to distinguish between two tasks involved in the analysis of images of the breast. First, there is a visual search task in which potential abnormalities must be found. Second, there is a recognition task in which the nature of some suspicious features must be identified. Breast ultrasound is an established imaging modality, routinely used for both search and recognition tasks. In breast cancer screening, x-ray mammography is the standard starting point. Traditionally, this has involved radiologists and/or technologists examining two-dimensional (2D) mammograms. More recently, digital breast tomography (DBT) has been used to create stacks of virtual slices through the breast, yielding a three-dimensional (3D) volume of image data.[1,2] In the United States, ultrasound is widely used as a follow-up exam after a mammogram or MRI to diagnose suspicious

*Address all correspondence to Jeremy M. Wolfe, jwolfe@bwh.harvard.edu

2329-4302/2022/$28.00

masses in the breast.[3] It can also be useful, as a screening modality, in the initial search for abnormalities, especially for women with dense breasts.[4,5] Masses that are hard to see in x-ray or DBT images can be more detectable in ultrasound simply because of the differences in the imaging physics involved.

Currently, most initial ultrasound exams are performed by a sonographer moving an ultrasound transducer over the breast (hand-held ultrasound, HHUS). This creates a dynamic view of a slice into the breast in multiple imaging planes. However, HHUS has some drawbacks. For example, imaging acquisition and interpretation are operator-dependent. As a result, it is not easily reproducible, and although HHUS lets the operator explore the entirety of the breast tissue, it may not create a standardized, storable 3D dataset. Automated breast ultrasound (ABUS) has been developed in an effort to address these issues.[6] The ABUS version that we use here is the general electric (GE) Invenia ABUS system. It uses a large, reversed-curve transducer. The acquisition of ABUS images resembles image acquisition in x-ray mammography. The ABUS detector is pressed in contact with the breast while the ultrasound transducer moves inside the detector (transducer frequency range: 6 to 15 MHz transducer aperture length: 15.3 cm). Typically, three acquisitions are enough to image the entire breast. From the resulting 3D volume of the data, it is possible to reconstruct images in various planes. Normally, in ABUS, images of the breast are generated as a stack of coronal images that allows a reader to view the breast from the anterior (the nipple) back to the posterior (the chest wall). These images are synced to transverse images, which are the slices showing axial planes from the skin surface into the depth of the breast. The transverse images resemble the images created by more classic HHUS systems (see Fig. 1).

By placing the crosshairs on a region of interest (ROI) – e.g., on the finding in the right coronal image in Fig. 1 – the transverse images are aligned, and the corresponding location is marked, so a feature, imaged in the coronal plane, can be examined by "multiplanar correlation." The ability to scroll through the images allows the reader to view the transverse images in a manner similar to the methods they would use with HHUS.

When added to routine mammography, ABUS can improve diagnostic accuracy, assessment of lesion location, call-back rate, and confidence in call-back, especially for women with dense breasts.[7,8] However, it adds significantly to the time required for each case. Skaane et al.[9]

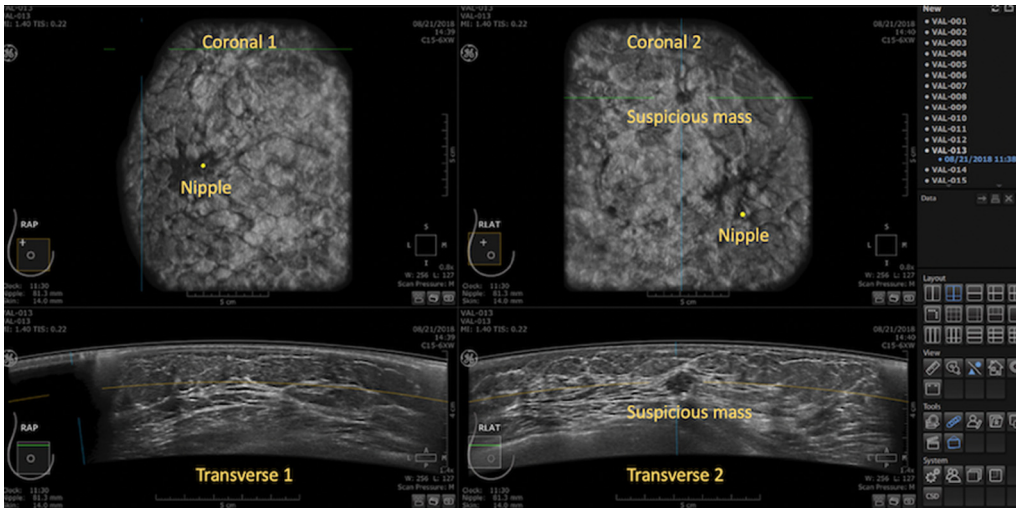

Fig. 1 One of the several available ABUS imagery "hanging protocols." This layout shows two coronal and two transverse images. Here are two sets of images of the right breast. Each coronal image set consists of a stack of slices from nipple to ribcage. Below each coronal image is the synced transverse image that allows the reader to examine slices showing axial planes from the skin to the ribs. Masses can appear as hypoechoic (black and round) or as architectural distortions in these images. Thus, in the right coronal image, there is a suspicious lesion. By placing the cursor at that location, the same lesion can be examined in the synced transverse image below the coronal image.

reported about 9 min per case. Other estimates range from 3 to 10 min.[6] The technology is relatively new, so there is only a modest body of clinical wisdom on the best way to use it. Like many modern image modalities, ABUS generates a very substantial volume of imagery, so readers could spend an arbitrarily long time on any case.[10] It would be valuable to know how experts spend their time when viewing ABUS. Such information could be a useful starting line for assessing how experts should spend their time on ABUS.

Efforts to study medical imaging professionals' search and recognition behavior have been underway since the 1960s when Tuddenham and Calvert[11] had radiologist readers point a spotlight to indicate where they were looking and recorded the movements of that spot on photographic film. The effort moved forward dramatically when Kundel et al.[12] brought eye-tracking technology to the task. Since then, eye tracking equipment has improved and has become an increasingly practical tool in medical image perception research (for recent reviews, see Refs. 13 and 14). Modern eye trackers can collect useful information from a reader, viewing the image while in a relatively natural position, seated in front of a workstation, without the need to restrain the head or to wear the eye tracking equipment.

One of the most important uses of eye tracking in medical image perception is that it can provide insight into the sources of error in search. Kundel et al.[12] proposed a tripartite division of false negative ("miss") errors that remains useful. There are search errors in which the eyes never fixate within a small "Region of Interest" (ROI) around the target. There are recognition errors in which the eyes land in the target ROI but stay only briefly; half a second has been used as a threshold in studies of the lung and one second for the breast. That half-second or one-second threshold is somewhat arbitrary, but it is designed to capture the situations in which observers fail to recognize an important item, even though its image fell on the fovea, the center of gaze. If the eyes spend more than this threshold amount of time scrutinizing a target, errors are considered to be decision errors. These are cases in which the readers recognize that a stimulus is of interest but fail to categorize it as a target. False positive errors are also often accompanied by extensive scrutiny. In this case, a stimulus is miscategorized as a target even though it is not. For 2D mammography, Krupinski[15] reported that errors by experts could be categorized as 25% search, 25% recognition, and 50% decision errors. In previous work, we replicated this result in a larger dataset.[16] Using eye-tracking metrics, Carrigan[17] reported on the distribution of these errors in sonographers, searching for signs of cancer in breast images. Krupinski's less experienced, novice observers produced 29% search, 42% recognition, and 29% decision errors.

As new imaging modalities appear, new viewing behaviors emerge. Classic eye tracking data involved eye movements in the XY plane over 2D images. Modern modalities create 3D volumes of data in which the observer's search path must be tracked over X, Y, and Z planes. Drew et al.[18] tracked the eyes of radiologists as they searched for lung nodules in chest CT imagery. By registering those XY data to corresponding slices in the CT image stack, 3D scan paths could be reconstructed. Drew et al. identified two search strategies, reporting that individual readers could be classified as "drillers" or "scanners." Drillers tended to hold the eyes relatively steady in XY while quickly moving through the lung in the Z direction. Then, they would move to a new XY position and "drill" through Z again. In contrast, scanners moved slowly along the Z axis while scanning more extensively in the current XY plane. When Aizenman et al.[19] conducted a similar study with the 3D stacks of images created by DBT; they found that behavior did not fit neatly into driller/scanner categories. Their observers engaged in rapid and repeated drilling while also scanning quite widely. Anecdotally, observers often reported that they believed that they were drilling, when asked after the experiment. They were less aware of their scanning behavior [c.f. Ref. 20]. This illustrates why introspection is not a substitute for objective experimentation.

In the current paper, we extend this line of investigation to ABUS. Specifically, we examine three questions:

1. How do readers divide their time between coronal and transverse views?
2. What can eye movement data reveal about sources of error in ABUS search tasks?
3. What do 3D drilling and scanning eye movement patterns look like in ABUS displays with multiple 3D stacks?

This study should be considered to be an initial exploration of a potentially large data space.

Table 1 Characteristics of the 12 readers in this study.

Observer	Profession	Speciality	ABR certified	ABUS trained	Years reading ABUS	Years reading mammo	Year reading DBT
1	Radiologist	Breast	Yes	Yes	1	15	6
2	Radiologist	Breast	Yes	Yes	1.5	19	2.5
3	Radiologist	Breast	Yes	Yes	10	35	17
4	Sonographer	Breast	No	Yes	5.5	0	0
5	Radiologist	Breast	Yes	Yes	2	25	8
6	Radiologist	Breast	Yes	Yes	1	13	6
7	Radiologist	Breast	Yes	Yes	5	8	8
8	Radiologist	Breast	Yes	Yes	4	4	4
9	Radiologist	Breast	Yes	Yes	1.5	27	10
10	Radiologist	Breast	Yes	Yes	2.5	22	8
11	Radiologist	Breast	Yes	Yes	12	13	5
12	Radiologist	Breast	Yes	Yes	4	25	10

ABR, American Board of Radiology.

2 Methods

2.1 *Observers*

A total of 15 observers participated in this study. Three observers did not have ABUS training and therefore were excluded from the analysis reported here. Table 1 describes the final 12 observers. All specialized in reading breasts and had ABUS training. Eleven were radiologists and one was a sonographer with several years' experience with ABUS. Ten observers were tested at the ABUS Optimization Meeting in Charlotte, North Carolina, United States (September 2019). Two observers were tested at the Radiology Society of North America meeting in Chicago, Illinois, United States (December 2019). The observers were given informed consent as approved by the Brigham and Women's Hospital IRB (IRB approval #2007P000646).

2.2 *Cases*

A total of 18 cases (10 abnormal and 8 normal) were used, and all were deidentified. Images were read on an Invenia ABUS workstation provided by GE Healthcare from their training cases. Ground truth was based on GE Healthcare information. Table 2 lists the 18 cases that were used in this study. Because we were testing readers for a fixed amount of time and different readers took significantly different amounts of time per case, the number of cases read by an individual varies widely. The number of readers who examined each case is listed in the "instances" column in Table 2. The number of cases read by each reader is given in Table 3.

As noted, we collected data in two different venues. During the first data collection, cases were selected randomly for each observer. No observer was able to finish all of the cases. In the second data collection, cases were presented to observers in the same order so that we could meaningfully compare eye-movement patterns between observers for the same set of cases, but this turned out to include only two of the readers whose data are included here. Overall, our readers completed a total of 102 instances of these 18 cases. A total of 62 of these were target present, with the other instances being normal. Abnormal cases had a range of findings from fibroadenoma to atypical papillary lesion (Table 2).

Table 2 Cases used.

Case	Breast	Diagnosis	Instances
VAL-002	LT	FA	10
VAL-003	RT	Normal	1
VAL-004	LT	IDC	7
VAL-005	RT	IDC	7
VAL-006	LT	Normal	1
VAL-007	RT	Normal	2
VAL-008	LT	IDC	6
VAL-009	RT	IDC	9
VAL-010	LT	Normal	1
VAL-011	RT	Normal	12
VAL-012	LT	Normal	9
VAL-013	RT	PAPILARY	8
VAL-014	LT	Normal	6
VAL-015	RT	Normal/FC	5
VAL-016	LT	FA and IDC	10
VAL-017	RT	FA	4
VAL-018	RT	Cyst	1
VAL-019	LT	Normal	3

FA, fibroadenoma, normal; IDC, invasive ductal carcinoma; FC, normal fibrocystic changes; papillary, atypical papillary lesion.
Note: FA cases were considered "normal" when the task was to search for cancer and "abnormal" when readers were asked if there were clinically significant findings that needed further examination (see procedure). The "instance" column lists the number of times each case is represented in the dataset.

Table 3 The number of cases read by each observer, divided by response type based on observers' final reports on the questionnaire.

Observer	1	2	3	4	5	6	7	8	9	10	11	12	Grand total	Percentage
True negative	2	2	5	6	4	3	2	3	1	2	3	4	37	84.1%
False positive	0	0	0	1	0	2	0	1	2	1	0	0	7	15.9%
True positive	2	2	2	6	3	7	3	7	2	1	1	2	38	65.5%
False negative	1	0	5	3	4	1	1	0	1	1	1	2	20	34.5%
Total cases	5	4	12	16	11	13	6	11	6	5	5	8	102	
Accuracy	0.80	1.00	0.58	0.75	0.64	0.77	0.83	0.91	0.50	0.60	0.80	0.75	0.74	
Time per case	2.9	4.78	1.58	1.14	1.65	0.88	2.78	1.76	3.29	3.85	3.07	1.56		

3 Procedure

The experiment started with a five-point eye tracking calibration procedure. Average fixation error during the calibration was <1- deg visual angle. After the calibration, observers were asked to read ABUS cases for 20 min. To simulate the natural reading scenario, the observers were allowed to read the cases at their own pace within the time window. The 20-min reading time was based on a pilot session done by one ABUS expert before the experiment. Eye movements were recorded using an SMI RED 250 mobile eye tracker with a sampling frequency of 250 Hz. The distance between the readers' eyes and the screen display was ~60 cm, and the eye tracker was attached underneath the display. After finishing each case, readers were asked to rate each case on an online questionnaire opened separately on a webpage. An instruction on the questionnaire told readers which case they should select to read next. Readers could freely move between the ABUS system and the online questionnaire. Before reading the experimental cases, each reader read one practice case to familiarize themselves with the ABUS hanging protocol used in the study and with the rating questionnaire. Next, readers were instructed to go to the questionnaire, which indicated the order in which to select cases. Because we were conducting a "natural history" study, to assess normal reading behavior, no specific instructions on how to use the ABUS system or restrictions were given. Readers could select any of the available image acquisitions or layouts and scroll through them with their mouse. In fact, readers rarely departed from the hanging protocol shown in Fig. 1. After viewing each case, readers answered a set of questions on the questionnaire. In the first venue, the readers were asked

1. Is the case abnormal (100-point scale: 0 = completely normal, 100 = definitely abnormal)?
2. What breast imaging reporting and data system (BIRADS) score would you give to this patient (0 to 6)?

An abnormality case was defined as a case that shows any clinically significant findings. In the second venue, the questions were modified, and the readers were asked

1. Does the case have cancer (100-point scale: 0 = completely normal, 100 = definitely cancerous)?
2. Would you call the patient back (yes, no)?
3. What BIRADS score would you give this patient (0 to 6).

The first question was changed based on the feedback received in the first venue because asking whether a case is abnormal or not gave ambiguous and inconsistent answers among readers. Some might interpret the question as asking if signs of cancer were present, whereas others may decide that abnormality comprised all clinically significant findings. In practice, this meant that the correct answer to case #2 (fibroadenoma, normal) was "abnormal" in the first venue and "not cancer, normal" in the second. This is accounted for in the results, tabulated below. The next case number was given to the reader as soon as they had answered the set of questions for the previous case. No feedback was given to the reader during the session. In addition to eye movements, we also recorded videos of the workstation screen during the experiment.

3.1 *Data Analysis*

Given the fixed 20-min recording, the total amount of eye tracking data from each observer is, unsurprisingly, quite similar, though the time to complete each case varies substantially. A number of trials completed ranges from 4 to 16, making comparisons between observers of limited value.

As is the case with most studies of medical image perception, certain compromises were needed. Of the many available display layouts/hanging protocols, we only looked at reader behavior in the most used default display layout. In fact, readers spent 89% of their time viewing this layout shown in Fig. 1 with two coronal views with their linked transverse views. One of the challenges in the current study is that the ABUS station is not specifically designed to connect with an eye tracker, nor does it automatically output information about the slice/depth in the coronal or transverse image stacks. The recorded video has on-screen values that can be used to calculate the currently present slice/depth. We extracted these values using optical character

recognition software (Tesseract). Thus, we have x, y, and the inferred z coordinates for the scanpaths of the readers. Due to insufficient data, we removed eye movement data when readers were not using the default layout. All results reported here are based on the remaining data.

4 Results

Figure 2 shows the scanpath for one trial. This illustrates ~5 min of eye position recording for one observer viewing the case shown in Fig. 1.

Similar images for all 102 cases are available online.[21] These figures are intended to give a general impression of the movements of the eyes. Eye position is plotted by a dot every 200 ms. Larger movements of the eyes (>100 pixels) are plotted as line segments connecting the starting and ending positions. Notice that these are not exactly the same as conventionally calculated saccades because it is not entirely clear how to define saccades when readers are scrolling in depth as they move their eyes. Similarly, the definition of a fixation is not as clear as it would be in a 2D image. Accordingly, we show descriptive eye position information as an approximation to standard 2D fixations and saccades. The colors of dots and lines code the time, starting with yellowish-green (Fig. 2, #1) and moving through green (#2) to light blue (#3) to dark blue and purple (#4).

Red boxes show the positions of coronal (upper boxes) and transverse (lower) images. When observers are not fixating within one of these four regions, eye position dots and lines are drawn in white. Recall that these boxes do not show left and right breasts, but they show the different image acquisitions from the right breast, in this case. In the current study, three acquisitions of the breast were provided. Readers could display two of the three acquisitions at a time and could replace either of those two with the third one by clicking the acquisition boxes at the lower left of the screen while assessing the case. Thus, a cluster of fixations on, for instance, the right coronal field, could represent fixations when viewing more than one acquisition. The long white saccades to the lower left of this image generally indicate fixation on the thumbnails of those three acquisitions as one acquisition was chosen to replace another. Examination of the scanpath on the right shows that the light blue traces in the coronal field are followed by blue traces in the transverse field. After that, the purple coronal and transverse traces probably reflect the examination of the third set of images.

While observers are examining the images, they can also move in depth in either the coronal view (Fig. 2, #5) or the transverse view (#6). These movements are illustrated by graphs of depth

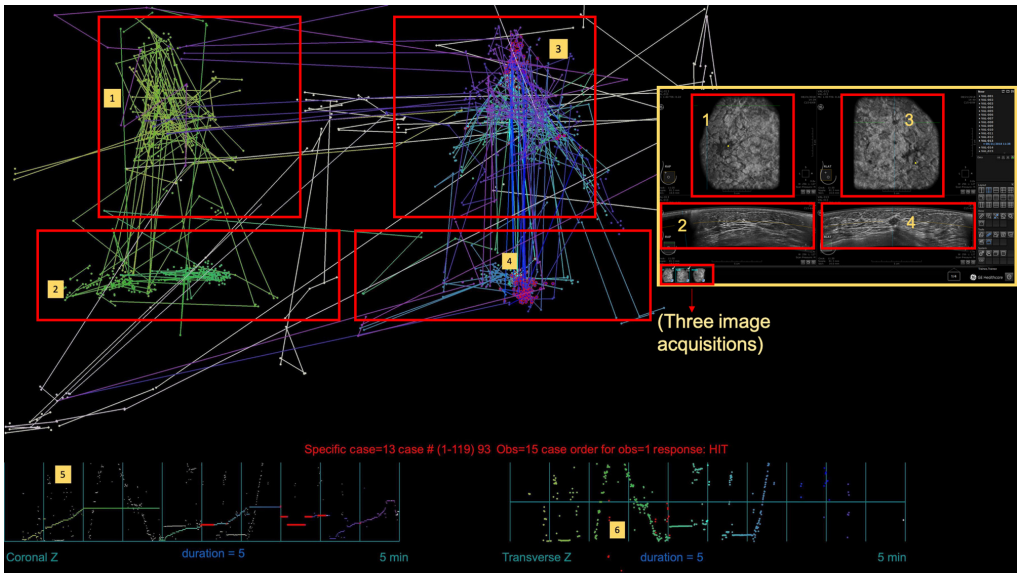

Fig. 2 Scanpath for one observer, viewing the case shown in Fig. 1 (here, shown in the inset to the right).

at the bottom of the figure. The *x*-axis is time (marked in 30-s intervals), and the *Y*-axis shows depth or slice. When the eyes fall within an ROI around the target lesion, the eye position dots are marked with a reddish outline as can be seen in #3 and #4. Notice that this corresponds to the "suspicious mass," marked in Fig. 1 and labeled in the inset in Fig. 2. The color scheme in the depth graphs (#5 and #6) follows that of the *XY* eye position graphs above them.

Given this information, it is possible to reconstruct the narrative of how this case was examined. The reader started in the coronal image on the left (#1). As can be seen in the Coronal Z graph (#5), at the bottom of Fig. 2, the reader moved through the depth of the coronal image from nipple to chest wall over the course of a minute while fixating on the left coronal image region. Next, they moved to the corresponding transverse image (#2). Note that, when they fixated on the transverse image, they stopped moving in coronal *Z* and moved through the transverse images indicated by the up/down movement in the third- and fourth-time bins in the transverse *Z* graph (#6). Rotational movements that are possible in the ABUS display are not represented here.

Having found nothing in the left image, the observer moved to the second coronal (#3). Here, they found something suspicious. At this point, it may be that, after looking at the images in the right coronal (#3) and right transverse (#4), the observer went to the lower left to load the images from the third acquisition and examined them as well. The three epochs of movement in depth in coronal (#5) and transverse (#6) images strongly suggest that all three acquisitions were examined even if it was possible to locate the target with fewer acquisitions. The reader scrutinized the representation of the suspicious item in the transverse representation (4, 6) and correctly decided that this was a positive ("HIT") case. Similar stories can be derived from each of the scanpaths in the dataset. As discussed below, the key takeaway is that readers had quite standardized ways of evaluating cases.

Table 3 summarizes each observer's performance. As noted, because observers read for a roughly fixed period of 20 min at their own pace, the number of cases read by each observer varied. The response types (true negative, false positive, etc.) are based on whether the case contained a search target (abnormal clinical finding/cancerous target depending on the question) and the readers' rating result (0 to 50 = target absent; 51 to 100 = target present). Thus, e.g., a rating of, say, 25 for a target present case would produce a false negative error.

It is important to note that the small number of cases per observer renders analysis of individual accuracy uninteresting, and the overall accuracy is unlikely to be representative of error rates in a clinical setting. Overall, there are 20 or more examples each of true positives, false negative errors (misses), and true negatives. False positive errors (false alarms) are rarer.

The false negative/miss rate (34.5%) is relatively high, although, again, both the number of cases and observers are too small for that figure to be especially informative. It does, however, provide a usefully large sample of errors to analyze. In the classic analysis of eye movements over 2D mammograms or chest CTs, false negative errors are divided into "search," "recognition," and "decision errors."[12] Search errors are said to occur when the eyes never fixate within an ROI around the target. Recognition errors are those in which the eyes fixate on the target ROI but only briefly (<500 ms for chest, <1000 ms for breast). A longer cumulative period of fixation on the target defines a decision error in which the observer scrutinizes the correct location but fails to report the target. Using the standard definitions of error types as proposed for the breast, at least 17 of 20 false negative errors (85%) would be classified as decision errors (over 1-s fixation time). There was one recognition error and two search errors with no fixations on the target. This distribution is quite different from error distributions in 2D x-ray mammography.[15]

Why the large proportion of decision errors? One possibility is that, in ABUS imagery, features that may be clinically significant are easy to spot. Masses often appear as hypoechoic (round and black) in the coronal views, and even relative novices can spot them (see Fig. 1). Once spotted, these hypoechoic findings must be interpreted. That proved to be more of a problem for the readers viewing these images. This could reflect the relative novelty of ABUS imaging. Readers may simply have less experience interpreting these images. A portion of a typical scanpath pattern for this type of error is shown in Fig. 3(a). The figure shows a small patch of the larger scanpath. As can be seen, the reader found something suspicious in the coronal image. They moved down to the transverse to examine the same finding, but they decided, incorrectly, that this particular hypoechoic finding was benign.

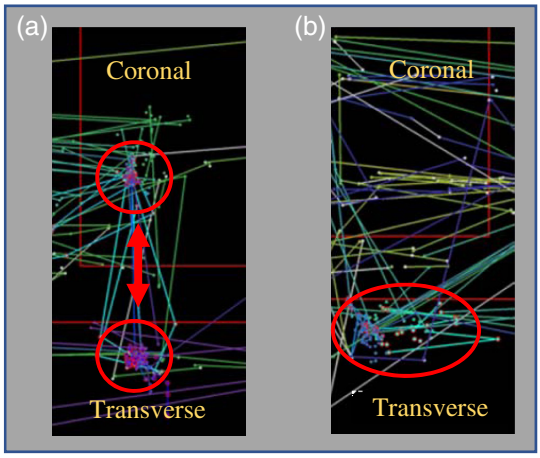

Fig. 3 Enlarged views of two scanpaths illustrate two types of false negative errors. Each dot represents a single fixation during the search. The red outlined dots represent the fixations on the target. (a) A clear decision error with extended scrutiny of the target. (b) A possible recognition error with the only fleeting fixation on the target.

It is also possible that the decision errors represent ambiguity in the images. It could be that malignant and benign anomalies in the breast give rise to similar signs in the ABUS image and that performance is limited by the images rather than by the training of the experts. It would require more research to determine if the high rate of decision errors in this dataset is an attribute of ABUS images, of the readers, or merely of this relatively small study. For the present, it is simply worth noting that readers made errors when they came to the wrong conclusion about correctly scrutinized spots in the ABUS images.

4.1 *Where Do Observers Spend their Time?*

Figure 4 shows the cumulative time spent in each image field as a function of the response type. The first point to notice is that readers appear to spend more time on average looking at the coronal images (26.5 s) than at the transverse images (18.5 s). There was no systematic preference for the left or right field, so they were combined in a three-way ANOVA with factors of

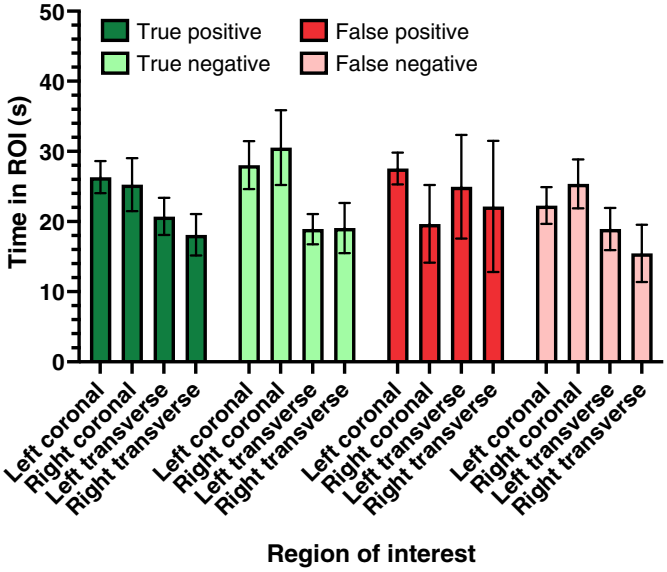

Fig. 4 Cumulative time in regions of interest, defined by image field, as a function of response type. Error bars show one standard error of the mean.

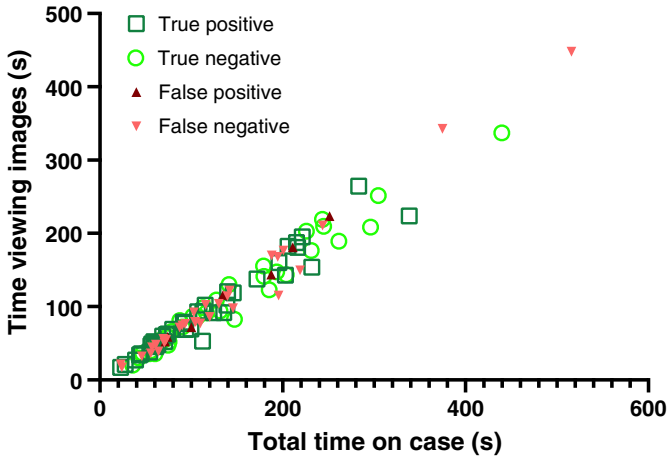

Fig. 5 Time viewing images as a function of total time on a case. Each datapoint represents one case.

image type (coronal versus transverse), target presence, and accuracy. The effect of image type is significant [$F(1.98) = 11.05$, $p = 0.0012$, partial $\eta^2 = 0.10$]. No other main effects or interactions are significant (all $p > 0.2$, all partial $\eta^2 < 0.02$). The most interesting aspect of this analysis is not the smaller proportion of time in the transverse images; it is that the readers spend about the same time in an image, whether or not there was a target present. This did not need to be the case. If observers had looked at the coronal images and only used the transverse images extensively if they found something of interest in the coronals, then we would have expected to see higher transverse usage in the target present (true positive) cases compared with the target-absent (true negative) cases. However, that is not what the data show.

The data suggest that observers used a quite stereotyped strategy, often similar to Fig. 2 in which each field was visited in sequence. This consistency can also be seen in Fig. 5. It shows the total time spent in all four of the image fields (the ROI) compared with the total time for the entire case. This allows us to ask what percentage of the time in a case is spent viewing other parts of the interface (e.g., switching from one image acquisition to another). Each datapoint in Fig. 5 shows one of the 102 cases, color-coded by response type. It is clear that the proportion of time spent in the image ROIs is very consistent, averaging 80%. There is no difference between different trial types (range 78% to 81%, no pairwise comparison approaches significance, all $p > 0.48$).

Across all 102 cases, several generalizations are possible. Notably, all of the scanpaths for an observer, regardless of the case type, tend to look broadly similar to each other. Moreover, observers tend to look through all of the images that are available in the case. Recall that there are typically three acquisitions – three different views – for each single-breast case. In many cases, as in Fig. 2, it is quite clear that the observer loaded all three image acquisitions and examined their coronal and transverse images in order. This can be seen quite clearly in Fig. 6 which reproduces just the movements in depth in the coronal and transverse views for one observer viewing one case. As in Fig. 2, the X axis is time with 30-s intervals marked, and the Y axis shows depth or slice. The observer starts by moving through the depth of one coronal image. They then move to the transverse image and look through that reconstruction. This is repeated for a second coronal image (blue), its transverse, and finally for a third coronal and transverse image (purple).

As noted, observers tend to have quite consistent eye movement strategies in this task. It would be easier to tell if two scanpaths came from the same individual than to tell if two scanpaths came from the same type of trial.

4.2 *Drilling and Scanning*

As mentioned above, in earlier work examining the eye movements of radiologists viewing lung CT, we found that readers tended to adopt either a "drilling" or "scanning" strategy.[18] With

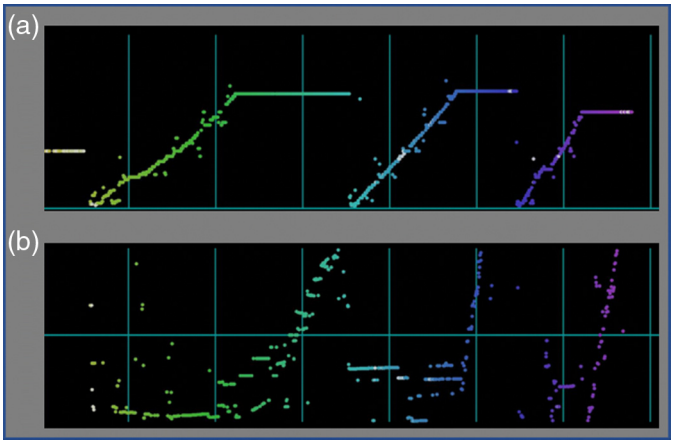

Fig. 6 Movements in depth (slice) for one case across time. (a) The coronal slices and (b) the transverse. This is a fairly typical, if particularly clear, example of an observer looking through all of the available images.

ABUS displays, it is somewhat more difficult to put a "driller" or "scanner" label on a reader because there are two different types of images. A reader might scan one and drill another. Nevertheless, we can see the difference between drilling and scanning behavior in these data. Figure 6 is a clear example of scanner behavior. The reader moved relatively slowly and steadily through the transverse and, especially, through the coronal images. They were scanning fairly widely in the *XY* plane of the image as they were slowly moving through in depth. In contrast, Fig. 7 shows drilling behavior. The clouds of dots represent multiple rapid back and forth movements through the stack of images over the same 30- to 60-s period in which the reader in Fig. 6 moved only once, in one direction.

The results in Figs. 6 and 7 were selected as particularly clear examples. Overall, it appears that there is more scanning behavior through the coronal images than the transverse. In an effort to examine this question in a more quantitative manner, we filtered the data to find all pairs of eye positions, separated by 200 ms that had both endpoints in the same field. From these pairs, we can calculate the extent of movement in the *XY* plane (*dXY*: distance in *XY* plane, measured in pixels) and in *Z* (*dZ*: distance in *Z*, measured in slices). Thus, for all appropriate 200-ms time frames, we compared the movement in the *XY* plane with the number of slices visited in *Z*. The 200-ms time frame was selected as it is sufficiently long to obtain meaningful information while producing a large sample of datapoints.

There are over 59,000 such pairs of *dXY* and *dZ* values. The distributions of *dXY* and *dZ* values are strongly positively skewed. Most values are small with a few large movements.

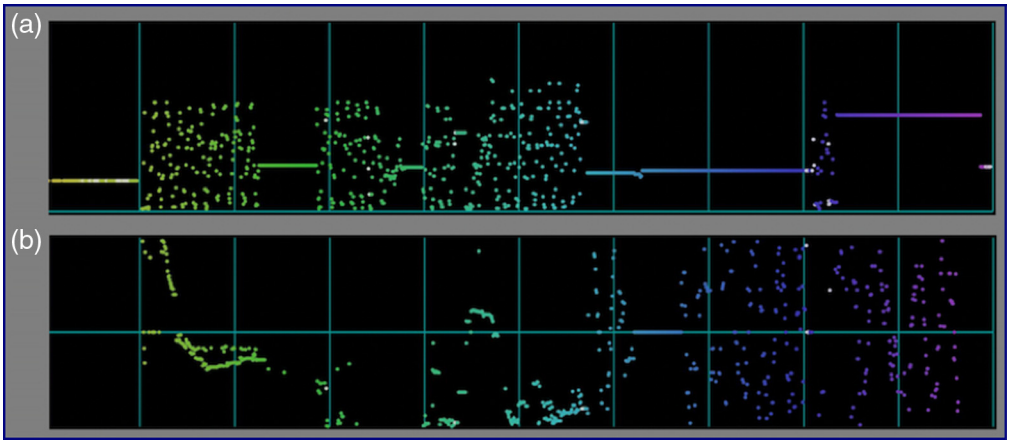

Fig. 7 Drilling behavior is shown by the clouds of dots that represent rapid movement up and down through the stack: (a) the coronal images and (b) the transverse.

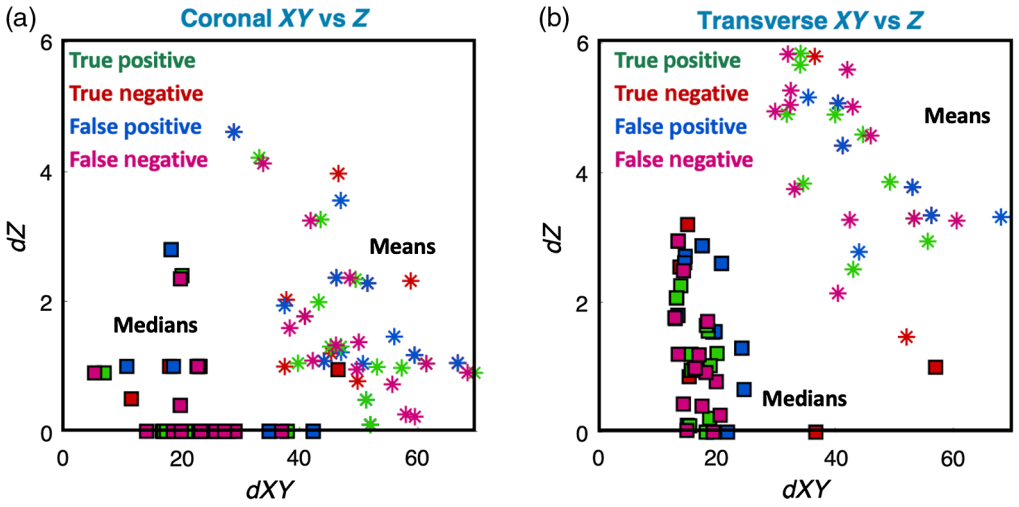

Fig. 8 Movement in the Z plane (dZ, slices) as a function of the movement in the XY plane (dXY, in pixels) for (a) coronal and (b) transverse images. Squares show medians for each observer. Stars show means. Each color shows one type of trial: green – true positive, red – true negative, blue – false positive, magenta – false negative. Large dZ movements with small dXY indicate drilling (e.g., many of the transverse medians). The near-zero coronal medians and low means suggest scanning behavior.

Some of these large movements are likely to be measurement errors. To deal with these outliers, the top 2% of values were removed from each distribution before analysis.

Figure 8 shows mean (stars) and median (squares) values in plots of dXY against dZ for coronal and transverse images for each observer for each type of trial (target present/absent X correct/incorrect).

The strong positive skew in the underlying data explains why the mean values are much greater than medians. The median values may be more informative in this case. For coronal images, the median dZ value is actually zero, with no motion in depth, for most of the observers. This is consistent with a bias toward scanning in the coronal image. More than half of the time, readers are scanning around in the XY plane while not moving in Z. For the transverse images, dXY movements are surprisingly consistent across readers (15 to 20 pixels in the medians). The dZ movements in depth are larger than what is seen in the coronal images and more variable. This more closely resembles a drilling strategy in which observers place the eyes at a relatively stable XY location while moving through the stack in the Z direction. Taken together, these patterns would be consistent with a strategy of using the coronal image to find a target, somewhere in XY, and then focusing on that XY location while looking at the item in multiple slices in the Z direction. One might think that this would mean that evidence for drilling would be stronger in true positive than in true negative cases. However, this does not appear to be the case. The different colors in Fig. 8 denote different types of trials. This measure reveals no obvious differences between target present vs target absent or correct vs incorrect cases. This can be seen as evidence for a quite stereotyped pattern of behavior, used by readers across all cases. Readers seem to scan the coronal for the most suspicious features and then drill down on that feature in the transverse image.

5 General Discussion

To summarize the findings, readers examine ABUS images systematically. Each reader may have their own specific variant, but a good generalization would be that our typical reader would start with the left coronal field and move to the left transverse field, then to the right coronal, and to the right transverse. At this point, given that there was a third acquisition available, readers would typically load it and examine those coronal and transverse images, in turn. This is eminently sensible and thorough. For any one reader, this general pattern is much the same,

regardless of whether the target is present or not. When readers fail to report a target (~34%), that target has usually been scrutinized. Most of the false negative errors in this dataset qualify as "decision errors" in the Kundel taxonomy. With regard to their search through the 3D volumes of images, most readers cannot be neatly categorized as "drillers" or "scanners."[18] Overall, more scanning behaviors are observed in the coronal images and more drilling in the transverse.

At the outset, the time required for ABUS exams was raised as a problem that might discourage the use of ABUS in standard clinical practice. The eye tracking results presented here suggest one potential intervention. The typical behavior of looking through all of the available images may be prudent, but it could also be unnecessary in a significant fraction of cases, especially in a screening setting. Specifically, it would be worth knowing if true negative responses could be reliably based on coronal images alone. The coronal and transverse images are different representations of the same data. The transverse images are useful in determining whether something suspicious, detected in the coronal image, is malignant or not. However, if there is nothing of note in the coronals, it may be unnecessary to examine the transverse images at all. This is akin to noting that, if the cookie jar is empty when viewed from the side, there is little point in examining it again from the top. It is important to note that the current results do not tell us, one way or the other, whether time spent on the transverse images when nothing is seen in the coronals is well-spent time. To assess this, further experimentation is required. In a screening setting, in which positive cases are rare, substantial time could be saved if even a modest percentage of cases could be dismissed without recourse to the transverse images.

The pattern of false negative errors is also of interest in this dataset. The sample of cases and readers is too small to make statements about ABUS accuracy in general. What is noteworthy here is that a very substantial majority of false negative errors appeared to be "decision" errors, by the Kundel et al.[12] classification. Readers scrutinized the suspicious feature, but declared it to be normal, perhaps a cyst, rather than a malignancy. In this, our ABUS data differ from the analysis of errors in other imaging modalities in which errors tend to be more evenly divided among search, recognition, and decision error categories. In this dataset, there were 58 positive cases, and readers searched successfully for 54 of those, that is, they scrutinized the right spot on the image. If the 17 decision errors could have been avoided, sensitivity (here defined as the true positive rate) would rise from 64% to 93%. Eliminating all decision errors is unlikely because there will always be stimuli that are simply ambiguous. However, the large proportion of decision errors in this dataset argues for potential benefits from further education or, perhaps, assistance from an AI system.

5.1 *Limitations*

The present study has multiple limitations that make it an initial foray into the understanding of search behavior in ABUS and, by no means, the final word. Recruiting readers in medical image perception research is challenging, and the current study is no exception. Conducting our eye tracking study on the sidelines of radiology conferences allowed us to recruit readers, but it also limited the number of cases that could be tested. Of course, the COVID pandemic rendered it essentially impossible to collect further data of the sort shown here from 2020 to 2022. In addition, the prevalence of disease in our set of images is much higher than it would be in normal clinical screening settings, and this is known to impact readers' decision criteria.[22]

It would be interesting to compare the performance of sonographers and radiologists, especially because it is frequently the sonographers who are performing the initial search task [Carrigan et al.[17] #14865]. In this study, with just one sonographer, we cannot say anything on that topic.

Clinical workstations are not designed to output all of the data that would be useful for experiments of this sort. Clearer statements about the relative virtues of drilling versus scanning would require more data. Similarly, more data would be required to be convinced that the pattern of errors in ABUS is different than that in x-ray-based mammography. This study was limited to one of the multiple available "hanging protocols." To optimize the use of ABUS, it would be valuable to compare across display strategies. It is never going to be practical to collect fine-grained eye-tracking data on a large number of radiologists while they read a large number of cases and while they use a variety of hanging protocols in a neatly counter-balanced

experimental design. However, significant progress could be made if workstations routinely produced a detailed running record of the display and the reader's activity. The workstation, of course, "knows" which stacks of images are on the screen and which slice is currently visible, but that knowledge is not saved into a datafile that is accessible to researchers. Mouse tracking can be used like eye tracking in some cases. For example, if the task required the use of a magnification tool, a type of scanpath could be obtained by tracking the deployment of that tool – again, if the information was available from the workstation. The workstation software is proprietary, but manufacturers would find that medical image perception researchers would be eager to collaborate if those data were available.

6 Conclusion

The current dataset shows that eye tracking can reveal much about the reader's strategies and outcomes in medical image perception. It is worth noting that simply asking readers what they were doing with their eyes is unlikely to be useful. People are surprisingly bad at monitoring where they have looked, even over short periods of time,[20,23] let alone over a period of several minutes while performing a demanding task. In the present study, we learn that readers have idiosyncratic, but typically very systematic, approaches to reading these images. When they commit false negative errors, it is typically because they incorrectly classified what they found, not because they failed to find the target. Further research will be required to determine if the systematic approaches can be optimized to save time and if the incorrect classifications could be reduced.

Disclosures

The authors declare no conflict of interest. This research was approved by IRB of Partners (now Mass General Brigham), Protocol No. 2007P000646.

Acknowledgments

This work was funded by GE Precision Healthcare LLC and the National Institutes of Health, Grant No. CA207490.

References

1. A. Chong et al., "Digital breast tomosynthesis: concepts and clinical practice," *Radiology* **292**(1), 180760 (2019).
2. S. Vedantham et al., "Digital breast tomosynthesis: state of the art," *Radiology* **277**(3), 663–684 (2015).
3. C. Sohn, *Breast Ultrasound: A Systematic Approach to Technique and Image Interpretation*, Thieme, Stuttgart, New York (1999).
4. M. Nothacker et al., "Early detection of breast cancer: benefits and risks of supplemental breast ultrasound in asymptomatic women with mammographically dense breast tissue. A systematic review," *BMC Cancer* **9**(1), 335 (2009).
5. S. H. Kim, H. H. Kim, and W. K. Moon, "Automated breast ultrasound screening for dense breasts," *Korean J. Radiol.* **21**(1), 15–24 (2020).
6. S. S. Kaplan, "Automated whole breast ultrasound," *Radiol. Clin. North Am.* **52**(3), 539–546 (2014).
7. R. Rella et al., "Automated breast ultrasonography (ABUS) in the screening and diagnostic setting: indications and practical use," *Acad. Radiol.* **25**(11), 1457–1470 (2018).
8. R. R. Guo et al., "Ultrasound imaging technologies for breast cancer detection and management: a review," *Ultrasound Med. Biol.* **44**(1), 37–70 (2018).
9. P. Skaane et al., "Interpretation of automated breast ultrasound (ABUS) with and without knowledge of mammography: a reader performance study," *Acta Radiol.* **56**(4), 404–412 (2015).

10. 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**(2), 346–362 (2011).
11. W. J. Tuddenham and W. P. Calvert, "Visual search patterns in roentgen diagnosis," *Radiology* **76**, 255–256 (1961).
12. H. L. Kundel, C. F. Nodine, and D. Carmody, "Visual scanning, pattern recognition and decision-making in pulmonary nodule detection," *Invest. Radiol.* **13**(3), 175–181 (1978).
13. T. T. Brunye et al., "A review of eye tracking for understanding and improving diagnostic interpretation in medical trainees and experts," *Cogn. Res.: Principles Implications* **4**(1), 1–16 (2019).
14. Z. Gandomkar and C. Mello-Thoms, "Visual search in breast imaging," *Br. J. Radiol.* **92**(1102), 20190057 (2019).
15. E. A. Krupinski, "Visual scanning patterns of radiologists searching mammograms," *Acad. Radiol.* **3**(2), 137–144 (1996).
16. J. M. Wolfe et al., "What do experts look at and what do experts find when reading mammograms?" *J. Med. Imaging* **8**(4), 045501 (2021).
17. A. J. Carrigan et al., "A 'snapshot' of the visual search behaviours of medical sonographers," *Australas. J. Ultrasound Med.* **18**(2), 70–77 (2015).
18. T. Drew et al., "Scanners and drillers: characterizing expert visual search through volumetric images," *J. Vis.* **13**(10), 3 (2013).
19. A. Aizenman et al., "Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: an eye tracking study," *J. Med. Imaging* **4**(4), 045501 (2017).
20. M. L. Vo, A. M. Aizenman, and J. M. Wolfe, "You think you know where you looked? You better look again," *J. Exp. Psychol: Hum. Percept. Perform.* **42**(10), 1477–1481 (2016).
21. J. M. Wolfe, W. Lyu, and C. Wu, "Automated Breast UltraSound (ABUS) Eye-Tracking study," What eye tracking can tell us about how radiologists use automated breast ultrasound, https://osf.io/m9rcn/files/ (accessed 22 February 2021).
22. K. K. Evans, R. L. Birdwell, and J. M. Wolfe, "If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening," *PLoS ONE* **8**(5), e64366 (2013).
23. E. M. Kok et al., "Even if I showed you where you looked, remembering where you just looked is hard," *J. Vis.* **17**(12), 2–2 (2017).

Jeremy M. Wolfe is a professor of ophthalmology and radiology at Harvard Medical School. He runs the Visual Attention Lab at Brigham and Women's Hospital. With expertise in vision and visual attention, his research focuses on visual search with a particular interest in socially important search tasks in areas such as medical image perception. He is a founding editor of *Cognitive Research: Principles and Implications* and an elected member of the American Academy of Arts and Sciences.

Wanyi Lyu received her BS degree in neuroscience from Bates College in 2019. From 2019 to 2022, she worked as a research assistant in the Visual Attention Lab headed by Jeremy Wolfe at Mass General Brigham Hospital. She is currently a graduate student at York University, Toronto.

Jeffrey Dong is a resident physician at the Department of Medicine, Beth Israel Deaconess Medical Center. He received his MD degree from Harvard Medical School and his BS degree in electrical engineering and computer science from the University of California, Berkeley.

Chia-Chien Wu is a senior human factors engineer at Cepheid, Sunnyvale, where he conducts usability and human factors study for medical diagnostic devices. He was an instructor and research fellow at Brigham and Women's Hospital and Harvard Medical School where he conducted a study focusing on human perception, visual search, and eye movements in medical image perception.