Scanners and drillers: Characterizing expert visual search through volumetric images

Trafton Drew
Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA

Melissa Le-Hoa Vo
Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA

Alex Olwal
MIT Media Lab, Cambridge, MA, USA

Francine Jacobson
Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA

Steven E. Seltzer
Department of Radiology, Brigham and Women's Hospital, Boston, MA, USA

Jeremy M. Wolfe
Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA

Modern imaging methods like computed tomography (CT) generate 3-D volumes of image data. How do radiologists search through such images? Are certain strategies more efficient? Although there is a large literature devoted to understanding search in 2-D, relatively little is known about search in volumetric space. In recent years, with the ever-increasing popularity of volumetric medical imaging, this question has taken on increased importance as we try to understand, and ultimately reduce, errors in diagnostic radiology. In the current study, we asked 24 radiologists to search chest CTs for lung nodules that could indicate lung cancer. To search, radiologists scrolled up and down through a "stack" of 2-D chest CT "slices." At each moment, we tracked eye movements in the 2-D image plane and coregistered eye position with the current slice. We used these data to create a 3-D representation of the eye movements through the image volume. Radiologists tended to follow one of two dominant search strategies: "drilling" and "scanning." Drillers restrict eye movements to a small region of the lung while quickly scrolling through depth. Scanners move more slowly through depth and search an entire level of the lung before moving on to the next level in depth. Driller performance was superior to the scanners on a variety of metrics, including lung nodule detection rate, percentage of the lung covered, and the percentage of search errors where a nodule was never fixated.

Introduction

Rapid technology improvements have led to massive increases in the size and complexity of medical images (Andriole et al., 2011). What would have been a single chest radiograph has become a chest CT (computed tomography) scan with, perhaps, 1,000 "slices" through the body. These high-resolution images allow an observer to detect abnormalities that might not have been seen using older technology. For instance, in a recent national lung cancer screening trial comparing cancer-screening performance using either low-dose chest CT scans or chest radiographs, chest CT led to significantly reduced mortality (Aberle et al., 2011). By all accounts, this finding as well as the increased access to chest CT technology means that the number of chest CT scans read by radiologists is bound to increase dramatically in the coming years.

Still, while chest CT is undoubtedly a better lung imaging modality, we know very little about how

Citation: Drew, T., Vo, M. L.-H., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision*, *13*(10):3, 1–13, http://www.journalofvision.org/content/13/10/3, doi:10.1167/13.10.3.

doi: 10.1167/13.10.3 Received May 24, 2013; published August 6, 2013 ISSN 1534-7362 © 2013 ARVO

search is accomplished in this type of volumetric image. Understanding the patterns of search may inform best practices in training future radiologists. It may also allow researchers to design more effective computer aided detection (CAD) algorithms. The most common method of viewing volumetric images such as CT scans is a "stack viewing" mode in which observers navigate through stacks of 2-D images, typically oriented as if the observer is looking up from below (the axial projection). Scrolling through depth while viewing 2-D orthogonal slices is a common form of search in radiology, including virtually all CT and magnetic resonance (MR) images. Thus, understanding how search is accomplished in chest CT scans will be important for both specific information about this task and more general information regarding search strategies through volumetric images.

Although there is an existing literature devoted to understanding search strategies in 2-D medical images such as chest radiographs (Berbaum et al., 1998; Ellis et al., 2006; Kundel, Nodine, & Carmody, 1977, 1978; Kundel, Nodine, & Krupinski, 1989; Kundel, Nodine, Thickman, & Toto, 1987; Kundel, Nodine, & Toto, 1991; Manning, Barker-Mill, Donovan, & Crawford, 2006) and mammograms (Krupinski, 1996; Krupinski & Nishikawa, 1997; Kundel, Nodine, Conant, & Weinstein, 2007; Mello-Thoms, Dunn, Nodine, & Kundel, 2001; Mello-Thoms, Dunn, Nodine, Kundel, & Weinstein, 2002), much less is known about how search is accomplished through 3-D chest CT scans, or about 3-D search strategies more generally. A limited number of studies have examined visual search while moving through depth outside of the medical setting (Smith et al., 2008; Smith, Hood, & Gilchrist, 2010; Solman, Cheyne, & Smilek, 2012; Solman, Wu, Cheyne, & Smilek, 2013). Some studies have used eye-tracking to examine search strategy in CT colonography (Phillips et al., 2008) and stroke diagnosis in head CT scans (Cooper, Gale, Darker, Toms, & Saada, 2009; Cooper et al., 2010). These studies have predominantly focused on difference between 2-D and 3-D search, as well as the role of expertise. Thus far, these studies of search through volumetric images seems to conform to the general findings in the 2-D medical image literature (e.g., Manning, Ethell, Donovan, & Crawford, 2006): Experts tend to be much more efficient in their eye-movement patterns, while novices seem to follow a haphazard pattern of search.

Detecting lung nodules in chest CT scans is a difficult task. The radiologist is asked to detect small spherical nodules while ignoring features such as blood vessels that can also appear circular in a 2-D slice. When moving in depth, nodules will tend to "pop" in and out of visibility while vessels, which are essentially tubes, persist across many slices (Seltzer et al., 1995). Experienced radiologists tend to quickly scroll up and down through the lung as they search for nodules, hoping to catch the pop of a nodule quickly flitting in and out of view. Informal conversations with radiologists suggested that there was a great deal of variability in how radiologists are taught to search lungs and in the strategies that they employ during the screening process. Here, we use eye-tracking data to visualize search strategies and to assess which strategy led to the better outcome as measured by the true positive detection rate. Recent research has shown that chest CT plays an important role in the detection of small pulmonary nodules (Aberle et al., 2011). In the current study, we aimed to gain a better understanding of the strategies used to find these nodules and to assess why errors are made. Our ultimate goal is to use this information to improve radiologists' performance.

Materials and methods

In the current study, we monitored the eye position of 25 radiologists as each searched through chest CT cases. The goals of this study were to describe the 3-D scan paths, to estimate the cumulative volume of the lung that was examined, and to measure detection performance. Eye-tracker calibration for one radiologist was poor, and this person's data were excluded from all subsequent analyses. The radiologists' task was to detect as many nodules as possible in a single chest CT scan during a 3-min time period. Fifteen radiologist examiners at the American Board of Radiology annual meeting (ABR) searched through five chest CT scans. Ten radiologists from Brigham and Women's Hospital (BWH) searched three cases in addition to the same five examined by the participants at the ABR. For present purposes, we will focus on the five cases that both groups examined. Area of expertise and experience with chest CT cases varied widely across our radiologist observers. While all observers were familiar with the lung nodule detection task, the number of chest CT cases read per week varied from a high of 200 to a low of 0.

Participants scrolled through the stacks of images using the keyboard's up and down arrow keys and were asked to click on the center of mass of any detected nodule using the mouse. In clinical practice, radiologists typically use the mouse wheel to navigate through stacks of images, but this varies across viewers and viewing systems. Prior to beginning the experiment, the experimenter explained the navigation system and ensured that the radiologist understood this method of navigation. While the radiologists may have been slightly less familiar with this method of navigation, we

Case number	Slices	Slice thickness	Nodules
1	133	2.5 mm	7
2	290	2 mm	23
3	121	2.5 mm	4
4	134	2.5 mm	8
5	260	2 mm	10

Table 1. Case description.

have no reason to suspect that it would dramatically alter their overall search strategy and informal discussions several of the radiologists corroborate this assertion.

Our cases were drawn from the Lung Image Database Consortium (LIDC: Armato et al., 2007). To describe their cases, the LIDC method was to identify as many nodules as possible in each lung scan without requiring consensus between observers. Four radiologists read over 1,000 chest CT scans, marking all nodules they detected that were greater than 3 mm in diameter. The LIDC listed 2,669 locations of nodules that were marked by at least one radiologist. However, only 938 of these nodules were marked by all four of the radiologists who read all of these cases. Thus, there is no gold-standard truth associated with these nodules. Even under ideal conditions, nodule detection is a very difficult task and the low level of agreement on these nodules is a testament to this fact. In the five cases used here, a total of 52 nodules were marked by at least one LIDC radiologist. The five cases had 4, 7, 8, 10, and 23 nodules, respectively. It is possible that some of these were not true nodules and that some nodules were not marked. Our measures of accuracy, given below, must be understood in that light. This uncertainty is unavoidable if real chest CT images are used.

In clinical practice, radiologists can make adjustments to image position, scale, brightness, and contrast. To avoid additional factors that could affect performance in the study, we used the same fixed preset parameters for all subjects and CT stacks. For the present study, window and level were set to 1500 and −700 Hounsfield Units, respectively, values that are commonly used for this task in the radiology clinic. Table 1 shows number of slices per case and the slice thickness.

Radiologists were seated in a darkened room 62 cm away from a 20 in. monitor. To minimize eye-tracker error, the radiologists sat with their head loosely restrained by a chin rest and were asked to minimize movement as much as possible. The images were displayed at 512 × 512 pixels. Due to differences in the display equipment at ABR and BWH, the image was projected at a slightly smaller size at BWH (subtending 14.25° visual angle, DVA) than at ABR (16.97 DVA). The experiment was run using Psychtoolbox and the Eyelinktoolbox for MATLAB (Brainard, 1997; Cornelissen, Peters, & Palmer, 2002; Pelli, 1997).

True positives were defined as mouse clicks that occurred within 30 pixels of the center of mass of the nodule location as defined in the LIDC database and within two slices to either side of the central location of the nodule in depth. As noted, the LIDC database is a valuable but imperfect resource for the location and size of nodules. This makes it difficult to evaluate the meaning of marks that do not fall on listed nodules. It would be tempting to consider these to be false positive errors. However, there were a number of locations in our sample that were marked as nodules by the majority of our radiologists even though they were unmarked in the LIDC database. These seem likely to be nodules. In fact, it was clear from post-hoc consultation with radiologists that there were a number of suspicious nodules that were not marked by the LIDC database. As a result, marks that fell into the false-positive category may be made up of both false positive and some true positives that were omitted from the database. This makes it very difficult to compute common metrics of overall performance such as d′ or area under curve. As a result, for present purposes we will focus on the increased rate of true positives.

The nodule locations that were obtained from the LIDC database are locations that were marked as nodules greater than 3 mm by at least one of the four radiologists. The database is made up of over 1,000 chest CT scans, so it is not surprising that there may be some nodules that were missed by all four radiologists. Given this uncertainty, our analysis will focus on true positive and false negatives as defined relative to the LIDC database. That is, an observer could find or miss an LIDC nodule but we are agnostic about marks on other locations in the case.

Eye-tracking parameters

Eye tracking was carried out using a desktop mounted Eyelink 1000 (SR Research, Ontario, Canada) which sampled the *x* and *y* position of the eye at 500 Hz. We calibrated the eye tracker using a nine-point calibration procedure for each radiologist prior to each case. Offline, we coregistered the eye-position information with the position in depth, which allowed us to recreate three-dimensional eye-movement scan paths. In the majority of the 2-D search literature, eye-movement data are categorized as either fixations or saccades depending on eye position velocity and acceleration (for a review see Rayner, 2009). These categorizations are complicated

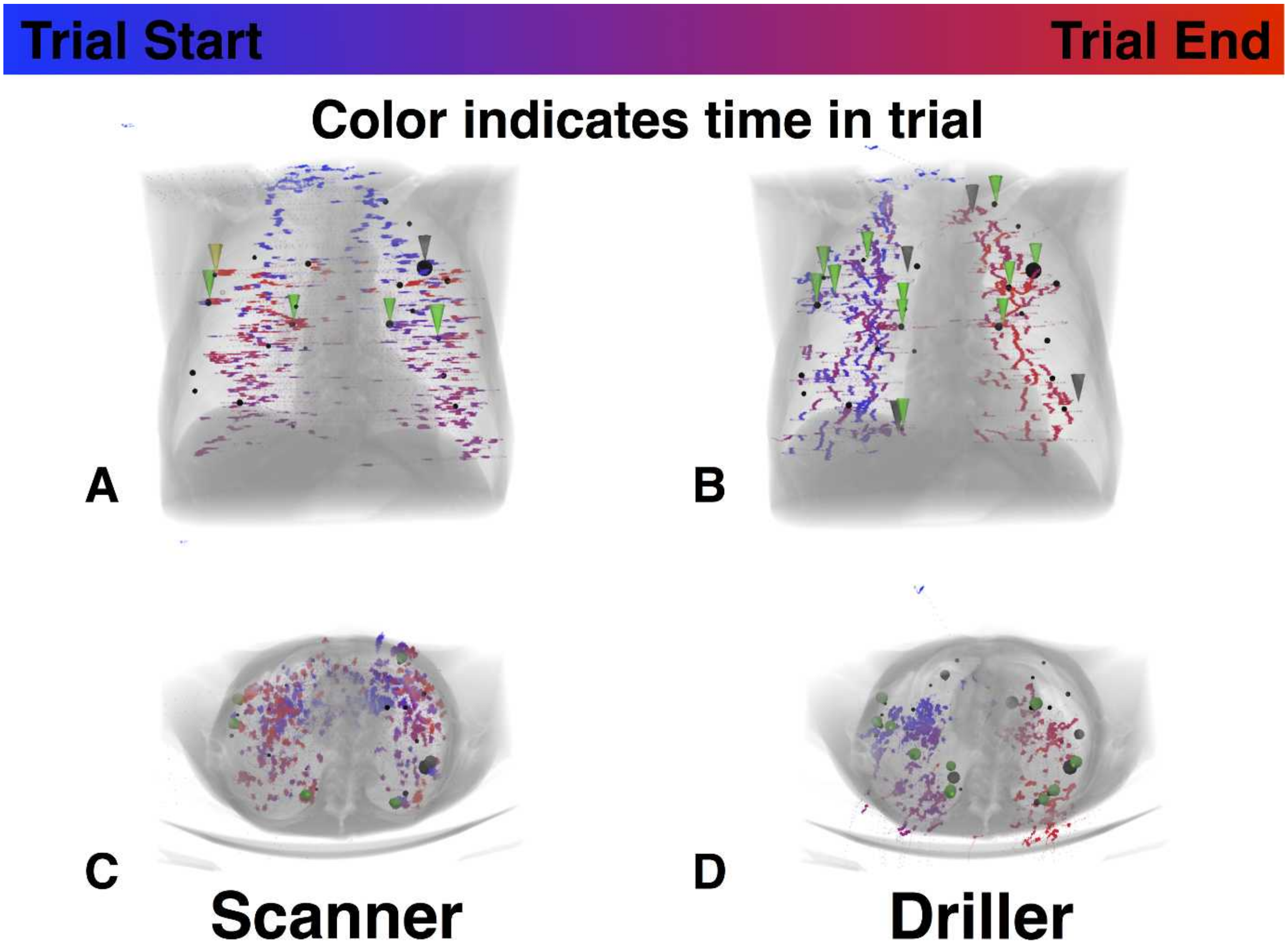

Figure 1. Examples of two observers' eye-movement traces from the coronal (A and B) and axial (C and D) perspective. Color of the symbols connotes time in trial, going from blue at the beginning of the trial to red at the end. Black spheres denote location of nodules according to the LIDC database. Green triangles indicate true positives. Black triangles: false positive marks. These figures are best viewed as movies, examples of which are available in the Supplemental Data.

by the ability to navigate in depth. If eye position remains relatively fixed in the x and y plane while the observer scrolls in z, depth, this is neither a conventional fixation nor a saccade. Here, we identify fixations and saccade amplitude without taking changes in depth into account. For our coverage and false negative categorization data, we focused on eye-position samples rather than the more traditional fixation or saccade dichotomy. A taxonomy of eye movements through a volume would be a good project for future research.

To support data exploration and analysis, we developed an interactive visualization tool that combines volumetric rendering of the anatomy combined with 3-D rendered gaze traces, LIDC nodule locations, and the locations marked by the radiologists. The tool allows the user to switch between different rendering styles (points, lines, glyphs), animations, color mappings, cut planes, and opacity transfer functions for the graphic processing unit-based (GPU-based) volume-rendering. These visualizations combine 3-D gaze plots with 3-D medical imagery and can be viewed stereoscopically through various means. We have primarily used our system with a ViewSonic V3D231 display, which enables stereoscopic viewing with passive polarization glasses through an embedded overlay pattern with lines of alternating polarization for a resolution of 1920 × 540 pixels. Our software is based on C++ and we use the Visualization Toolkit (Visualization Toolkit, http://vtk.org, January, 2013), as it both provides numerous useful visualization techniques, as well as GPU-accelerated volume rendering of the 16-bit imagery in our DICOM (Digital Imaging and Communications in Medicine) image stacks. Examples of these eye movement records can be downloaded from the Supplemental Data.

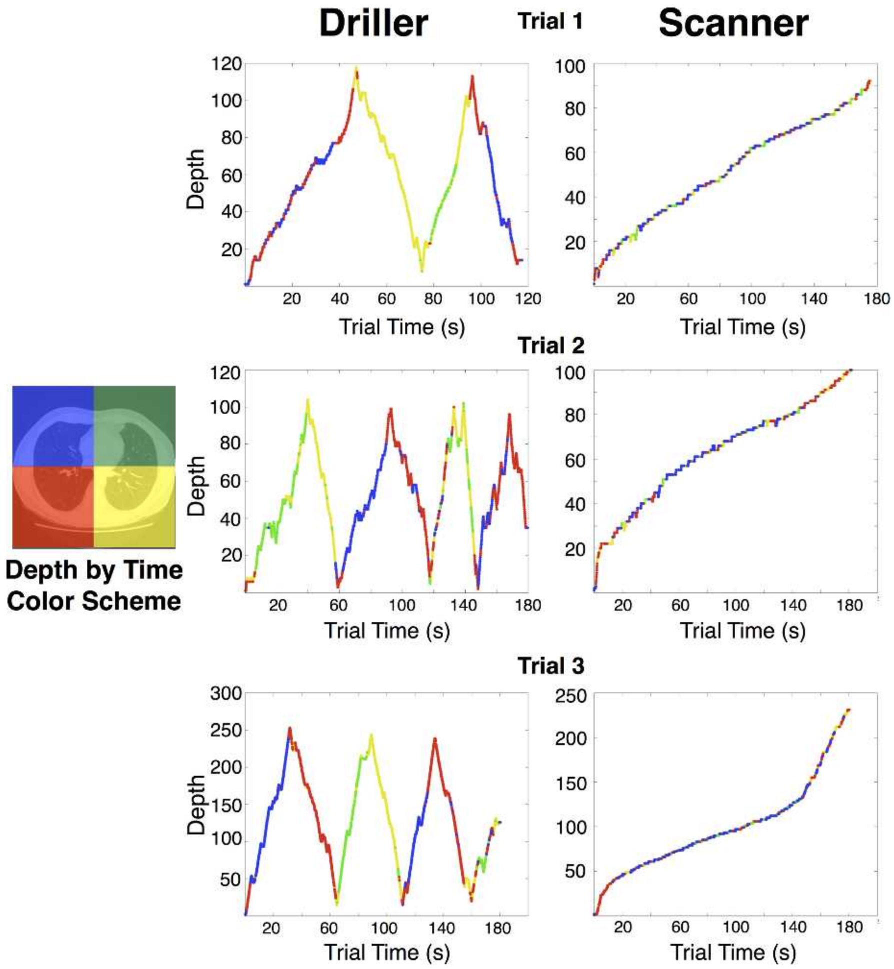

Figure 2. Examples of depth by time plots for two radiologists. Color indicates what quadrant of the image the eyes are at during that time. The radiologist on the left used drilling strategy. The radiologist on the right side used the scanning strategy.

Results and discussion

3-D reconstruction of volumetric data

To investigate the search patterns of the radiologists, we coregistered the x and y position of the eye with the slice in the CT stack to produce scan paths in x, y, and z. At 500 Hz, 3 min of scanning yields 90,000 samples per trial for each individual. We plotted these data in 3-D space. Figure 1 shows two views of the same case as viewed by two radiologists. The LIDC nodules are shown using black spheres, with size based on measured nodule radius. The plotted gaze data is color mapped based on trial time: from blue at the beginning of the trial to red at the end. Locations where the radiologist clicked to mark a nodule are indicated with a cone, with its tip pointing to the mouse cursor's location. The cones are colored green if they coincide with the ground truth data (true positives) or dark gray if the location was not labeled as a nodule by the LIDC radiologists (nominal false positives).

Scanner and drillers

Plotting the data in this manner illustrates the two different strategies adopted by different radiologists when searching through a volumetric image. Radiologists tended to consistently adopt one of these strategies. Some radiologists, here designated "scanners," tended to search throughout a given slice in depth before moving to a new depth. Others, here designated "drillers," held their eyes relatively still in x and y, limiting search to, a single quadrant of the lung while quickly scrolling—drilling—through slices in depth. Figure 1A and B show one example of each type. As color tracks time in these plots, scanners' graphs show a clear connection between color and depth in the lung. In Figure 1A, a scanner shows blue markings at the top of the lung and slowly progresses to red as the radiologist slowly moved toward the bottom of the lung. In Figure 1B, a driller shows long strings of similar colors that extend through depth as the radiologist drills through the lung. Views of the eye-movement data from the axial perspective are also useful in categorizing search behavior into one of these two groups. Drillers (as in Figure 1D) tend to show strong organization of color from this perspective since they tend to keep the eyes in roughly constant *xy* positions for extended periods of time as they move through depth. On the other hand, scanner eye movements from the axial perspective have little organizational structure: Scanning tends to lead to little connection between *xy* position and time once the *z* dimension is collapsed (Figure 1C). These strategies are not absolute. For example, a scanner might scan down through the lung and then back up rather than being restricted to a single pass. Given the three-dimensional quality of this data, it is best viewed as movies. These can be downloaded from the Supplemental Data.

Depth-by-time plots

While the 3-D reconstructions are a rich way to view the data, in an effort to derive a metric that would distinguish drillers and scanners, we simplified the volumetric reconstruction of the eye-movement data by reducing the *xy* position to a measure of the quadrant of the lung, currently fixated. Each 2-D slice of the lung was divided into four equally sized quadrants, an imperfect but adequate approximation of the anatomical quadrants of the lung.

n Figure 2, these four quadrant values are coded as different colors. This allows a 2-D representation of the 3-D scan path with time on the *x* axis, depth/slice on the *y* axis, and eye position in the plane coarsely coded into the four colors. Figure 2 shows these plots for each of the five cases for each of two observers, one driller and one scanner.

Quantifying search strategy

We have thus far discussed only qualitative measures of search strategy, noting that (a) there appear to be two distinct strategies for searching through the lung in this task, and (b) the strategies appear broadly consistent within, but not between, radiologist observers. Using these representations of the data, we tentatively divided our population of radiologists into two groups: 19 drillers and five scanners, based on the appearance of the color-coded depth by time plots (Figures 2 and 3) discussed previously. We hypothesized that these two groups would differ on simple descriptive statistics of eye-movement patterns.

During informal conversations with radiologists, one distinct difference between the driller and scanner strategy is the area on each depth that the radiologist actually searches. Whereas scanners try to search the entire lung area at the currently viewed level, drillers typically hold fixation on subsections (often quadrants) of the lung as they scroll through depth, making multiple passes (or drills) through the lung one quadrant at a time. Given this description, mean saccadic amplitude in the 2-D plane should be longer for scanners than for drillers, since drillers essentially confine search to a smaller region of *xy* space. Another simple metric that may be indicative of different search strategies is the number of fixation clusters. Fixation clusters are defined as successive fixations that fall within a single interest area. We defined four large interest areas that encompassed the four quadrants of the lung. As in the depth by time plots described previously, these interest areas are broad approximations based on using the center of lung image to divide each image into four equal quadrants. Given these large interest areas, any time that a radiologist had two or more successive fixations in a single quadrant, it was counted as one quadrant fixation cluster. Therefore according to our operational definition of drillers and scanners, drillers should have fewer quadrant fixation clusters per trial because they spend long periods of time within a single quadrant without leaving while a scanner may quickly visit all four quadrants during this time. As some radiologists spent less time searching than others, we then divided the number of quadrant fixation clusters by the amount of time spent searching. This yields a measure of the number of quadrant fixation clusters made per second. We used these two measures to create a scatterplot where each point represents one radiologist as a function of saccade amplitude and number of fixation clusters (Figure 3A). This method of displaying the data results in a small group of radiologists with a high quadrant fixation cluster rate and long saccadic amplitude grouped toward the upper right of Figure 3A. To better appreciate this difference we normalized both saccade amplitude and quadrant fixation cluster rate across our population from zero to one, then added the two measures to create a single index. This index can be thought of as the tendency to make large or cross-quadrant eye movements. The index has been sorted from lowest to highest in Figure 3B. In this representation, the distinction between our previously defined scanners and driller is quite clear. The five radiologists that we had previously categorized as scanners are also highest on this derived measure of eye-movement behavior.

Descriptive statistics based on saccadic amplitude and number of quadrant fixation clusters further support the rough categorization made on the basis of graphical representations of the data. Both saccadic

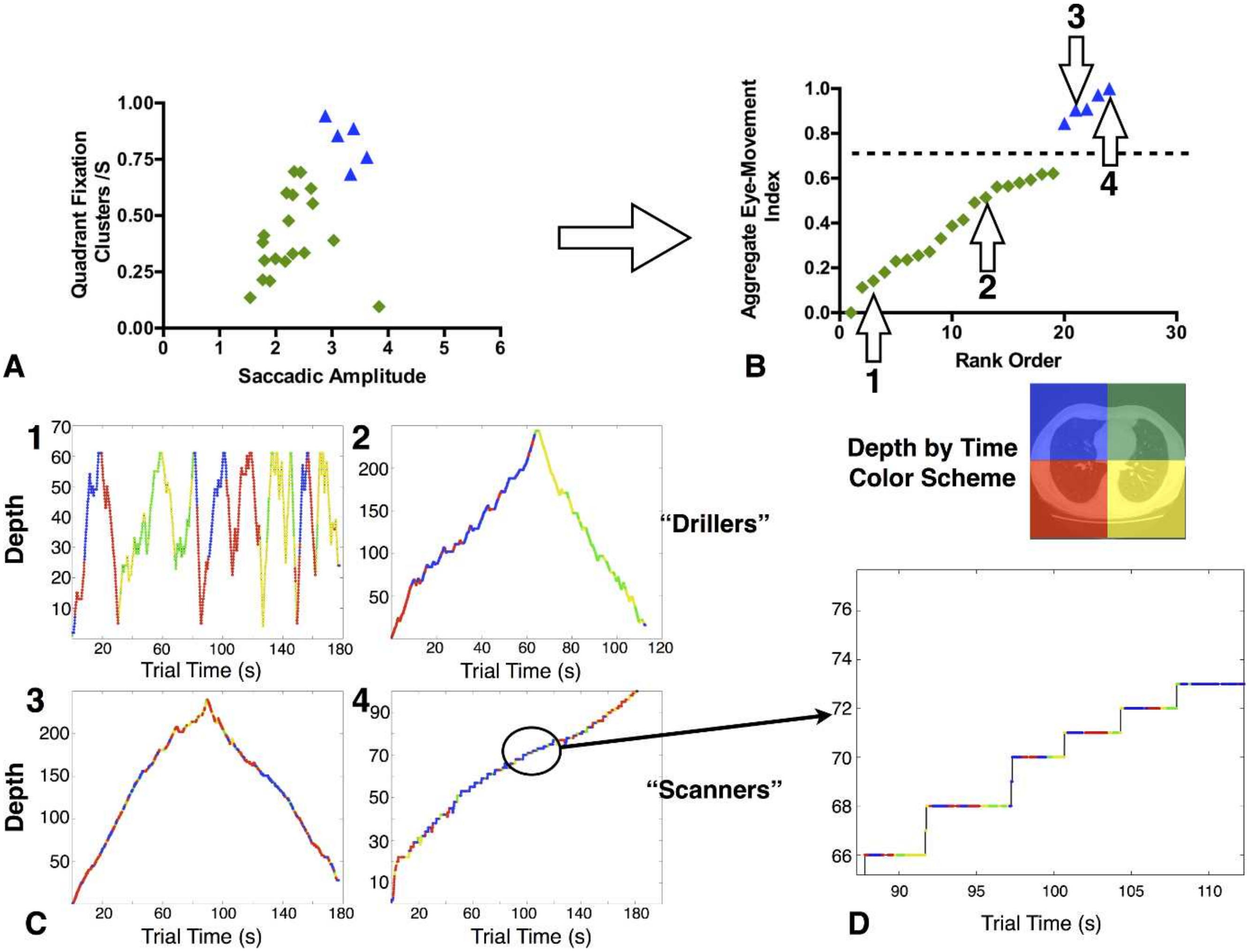

Figure 3. Defining drillers and scanners. (A) Scatterplot of quadrant fixation clusters/s by saccadic amplitude. The measures were normalized and then aggregated. (B) Resultant metric is then plotted with rank order on the *x* axis. (C) Four examples of individual trials from different radiologists are plotted. Numbers indicate the point associated with depth by time plot below. (D) Zoomed portion of a scanner's depth by time plot. Each quadrant is visited on each level prior to moving to a new depth.

amplitude, $t(22) = 4.76$, $p < 0.001$, and mean number of quadrant fixation clusters, $t(22) = 5.32$, $p < 0.001$, are larger for scanners than drillers (see Figure 3A). Although there is a large amount of variability in both measures within our two categories of search strategy, combining these two measures together appears to be a promising, simple way to categorize search strategy. In future work, we hope to test the validity of this method by asking observers to adopt either a scanning or drilling search strategy, then determining whether these measures are still an effective way to categorize these two distinct search strategies.

Nodule detection performance

Does it matter if radiologists are adopting a driller or scanner strategy during lung nodule detection tasks on CT scans? If one strategy is more effective for this task than the other, we should expect that it would lead to a higher true positive rate. This is a very difficult task, made more difficult given the strict 3-min time constraint employed in this experimental version. In pilot testing, it was clear that nodule detection rate for these stimuli was high when the radiologists were not given a time limit, but they also tended to search each case for a much longer time than cases are typically searched in the clinic. For additional discussion of this issue, see the Coverage section of the results.

In this sample, the drilling strategy led to a marginally significant overall increase in the number of locations marked, drillers: 10.3 marks per case, scanners: 7.9; $t(22) = 1.98$, $p = 0.060$. More important than the total number of responses, the true positive rate was higher for drillers than scanners, drillers: 60%

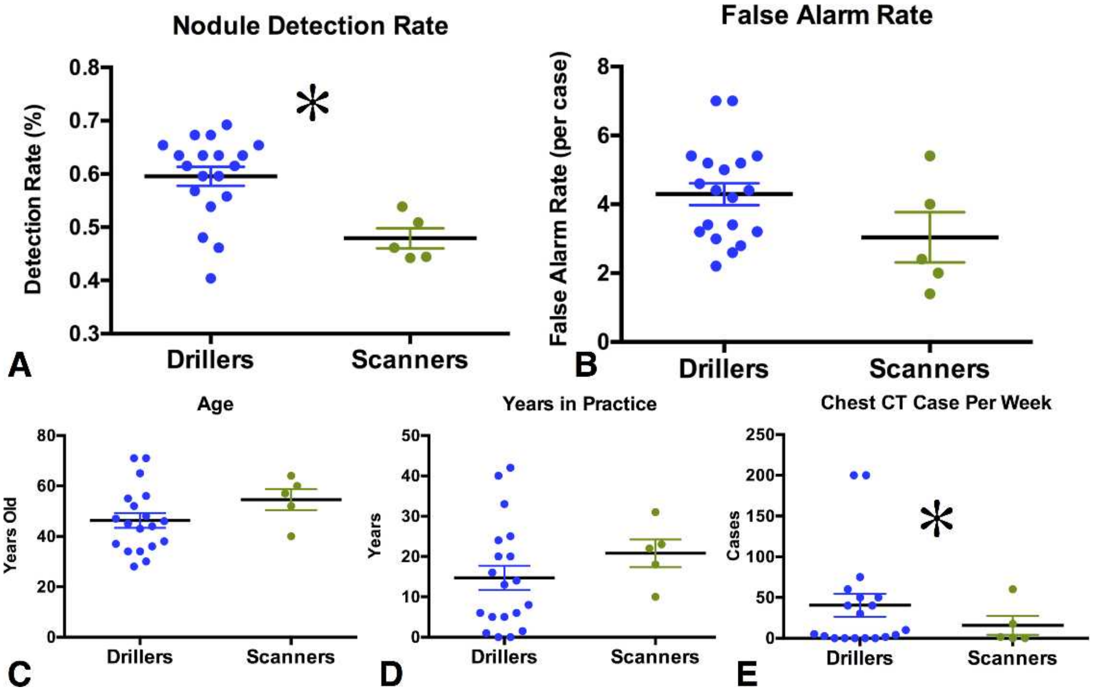

Figure 4. Behavioral performance and experience as a function of search strategy. Error bars here and throughout the paper represent standard error of the mean. Asterisks denote statistically significant differences ($p < 0.05$).

of all available nodules versus scanners: 48%; $t(22) = 3.9$, $p < 0.001$. The number of false alarms per case did not differ significantly, drillers 4.3; scanners: 3; $t(22) = 1.42$, $p = 0.17$.

There are a number of reasons why drillers might have found more lung nodules. It is possible that drilling is simply a better strategy for searching through volumetric images. This strategy seems to do a better job of taking advantage of the ability to scroll quickly through the image. However, if this is a real advantage it is almost certainly dependent upon the task at hand. Scrolling through depth is thought to allow radiologists to discriminate more effectively between lung nodules and lung vessels given the different 3-D shapes of these objects (Seltzer et al., 1995). The ability to see nodules pop in and out of existence is one of the reasons that radiologist began to examine chest CT scans using stack viewing mode.

Several covarying factors make it unwise to argue too strongly that the driller method is inherently superior. For example, our drillers tended have more experience searching through chest CTs (see Figure 4). The average number of chest CT scans read per week by scanners was lower (3.8) than drillers, 43.6; $t(22) = 2.75$, $p < 0.05$. In fact, while 40% of our scanners reported reading zero chest CT cases per week, only 21% of drillers responded in this manner. Obviously, this complicates the interpretation of the finding that drillers appear to be better at nodule detection than scanners. It is possible that the difference between experienced and less experienced readers would be just as great if all of them were scanners or all were drillers. Another possibility is that radiologists with more experience with chest CT eventually learn that drilling is a superior strategy and adopt this strategy as a result. Although the scanners did not read as many chest CT cases per week as the drillers, the area of expertise of these radiologists suggests that the observed performance decrement associated with this group was not driven by a lack of experience with chest CT images: Of the five scanners, two specialized in thoracic imaging, with the others specializing in emergency room radiology, mammography, and neuroradiology. In a post-hoc exploratory analysis, we compared nodule detection rate from the five drillers who reported reading zero cases per week (56% correct) and the scanners (48%). With only five radiologists in each group, the comparison is underpowered and the difference did not reach statistical significance, $t(8) = 1.71$, $p = 0.13$, but appears to follow the overall pattern that drilling leads to higher detection rates even when controlling for experience. Clearly, future research will be necessary to determine if drilling is truly a better strategy for this task. If it is, we may be able to improve the performance of scanners by encouraging them to using the drilling strategy.

There were nonsignificant age and experience trends. Scanners were older, 54.2 compared to 46.4: $t(22) = 1.4$, $p =$ n.s., and had more years in practice, 19.8 compared to 15.0: $t(22) = 0.9$, $p =$ n.s., than the drillers. These factors might warrant future investigation as well.

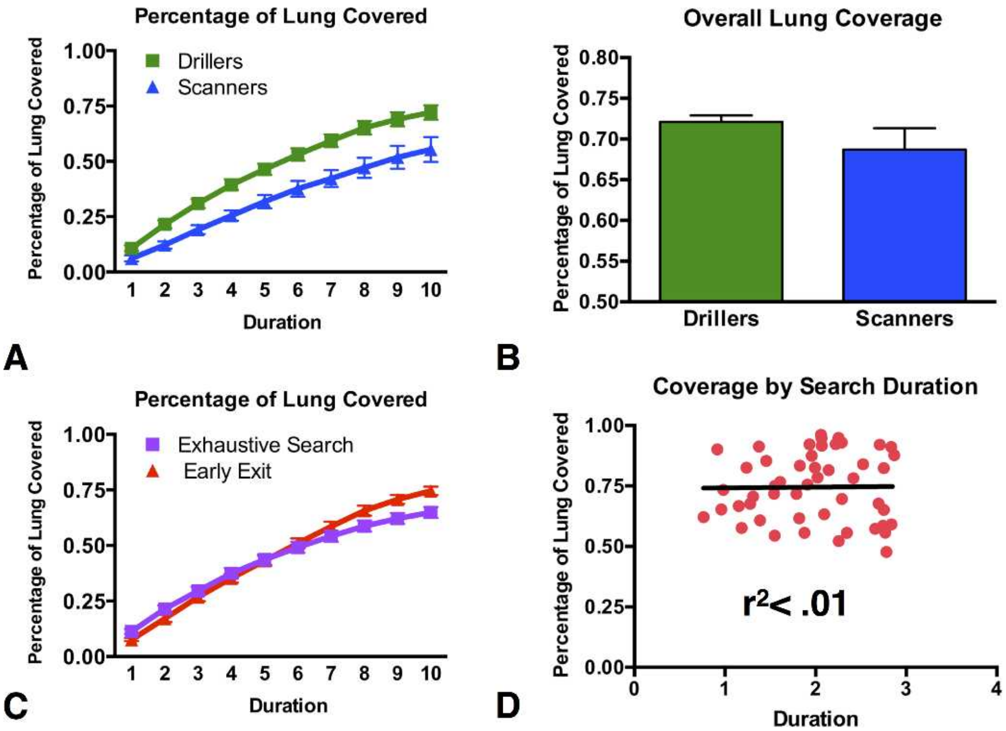

Figure 5. Coverage metrics. (A) Coverage for the two search strategies over time. Duration was divided into 10 equally sized epochs on each trial. (B) Total lung coverage for the two groups. (C) Lung coverage for trials where search continued for the full trial duration (3 min) or left early. (D) Correlation between total lung coverage and time spent searching for those trials where radiologists elected to finish early.

Coverage

What percentage of the lung tissue do radiologists actually search while performing a lung nodule detection task? The answer to this question depends greatly on how one estimates the "useful field of view" (UFOV). We recorded the *x*, *y*, and *z* position of the gaze 500 times per second. Each coordinate in this large matrix represents a single point in space. In order to determine coverage, we need to estimate the size of the region of space around this single point that we can think of as attended. This undoubtedly depends on the task at hand. For instance, the UFOV for simple stimuli (i.e., detecting a red bar amongst green bars) will be much larger than the UFOV for complex stimuli (i.e., detecting nodules; Young & Hulleman, 2013). Kundel et al. (1987) found that low contrast targets were detected roughly 80% of the time as long as the target was within a 5° diameter of the eye position. This 5° diameter estimate has been used in a number of follow-up studies involving chest radiographs (e.g., Kundel et al., 1989; Nodine, Mello-Thoms, Kundel, & Weinstein, 2002).

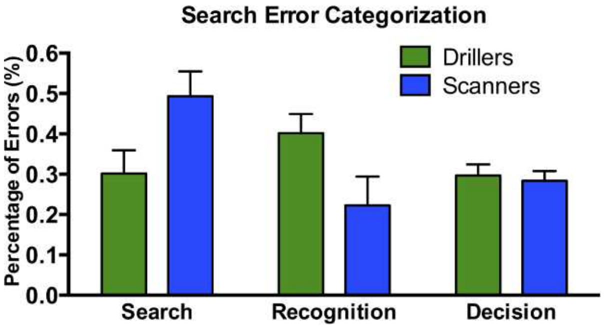

Figure 6. Search error categorization as a function of search strategy.

Given that these estimates were initially made to approximate nodule detection in chest radiograph, we used a 5° circle to estimate the total coverage of the lung. Computing total lung coverage was a three-step procedure. First, we used a simple image processing algorithm to extract the lung tissue from our chest CT scans and converted the images to a black (not lung tissue) and white mask (lung tissue). Second, we used our matrix of all *x*, *y*, and *z* eye positions over the course of the trial to place 5° black circles centered at each of these points. This amounts to "painting" the lung with a circular brush in all places where the eyes have visited. Theoretically then, a lung with 100% coverage would be completely black at this point. The final step of the procedure was to divide the number of white pixels in the newly created coverage lung image to the number of white pixels in the original lung image. To further understand how the process of searching the lung differed between the scanner and drillers, we analyzed lung coverage by group as a function of time. We divided each trial into ten equally sized epochs, then plotted cumulative coverage as function of time (Figure 5A). At each epoch, drillers covered more of the lung than scanners. The final total coverage differed as function of search strategy: drillers, 72%, covered more of the lung than scanners, 55%: $t(23) = 2.23$, $p = 0.029$. Overall we found that when using a 5° estimate of the useful field of view, an average of 69% (range: 44%–92%) of the lung tissue was covered by our radiologists.

Since we imposed a 3-min time limit on each trial, we cannot claim that radiologists would only search through two thirds of the lung in normal practice. However, while there is no time limit in the reading room, radiologists are generally under time pressure imposed by case load. The 3-min limit was imposed in an effort to simulate a typical clinical search time based on conversations with and observation of radiologists engaged in chest CT reading. In our study, 42% of the trials (49/118) were voluntarily terminated prior to our time limit. If we compare lung coverage for the "early exit" trials and those trials where the radiologists searched until the time limit ("exhaustive search"), we see that early exit trials produce reliably higher coverage than the exhaustive search trials, in the final epoch, 75% versus 65%: $t(116) = 3.14$, $p < 0.001$. This may suggest that, without a time limit, radiologists who used the full time in the current experiment would have produced somewhat greater coverage, but not much. Experienced readers of chest CT may be somewhat faster readers of chest CTs (unsurprisingly). Radiologists who reported that they read chest CT scans regularly ended search in less than 3 min on a higher proportion than the radiologists who were not regular chest CT readers, 49% versus 26%: $\chi 2$ (1, $N = 118$) = 6.37, $p < 0.05$. In the early exit trials, there was no correlation between time spent searching and coverage ($r^2 < 0.01$: Figure 5D).

These data represent one of the first quantitative estimates of coverage of a volumetric medical image. Previously, Ellis et al. asked four radiologists to search through 16 slice subsections of the chest CT scans (2006). They found that less than 50% of the lung was scrutinized by foveal vision when using a 2 DVA estimate of FOV. Using much more liberal estimate of FOV, our data similarly suggests that a substantial proportion of the lung is not closely scrutinized. This suggests that we are either underestimating the ability of radiologist to extract information from peripheral vision or that radiologists do not completely search the lung during lung cancer screening. If the lung is not completely searched during this task, it is likely that this is a cause of some false negative errors. Alternatively, this level of coverage may reflect the expertise of the radiologists. If certain areas within the lung hardly ever contain lung nodules, it would be logical to avoid spending additional time in these low-yield areas, especially when considering the constant time-pressure in the radiology reading room.

Despite all the benefits of the evolution from chest radiographs to chest CT scans, maintaining a representation of what areas have been searched is much more difficult in volumetric space than in the simple 2-D images of chest radiographs. Given the evidence that holding information in working memory decreases the efficiency of visual search in the laboratory (Oh & Kim, 2004; Woodman & Luck, 2004), we believe that these additional mnemonic costs with associated volumetric images are an important factor that should be considered when evaluating how to improve search performance in these complex displays. Furthermore, as noted by Solman et al. (2012), the observer is much more likely to get disoriented and waste valuable time in search tasks that ask the searcher to navigate through volumetric space. From this perspective, the low level of coverage in the current dataset suggests that providing the radiologist with some representation of what areas of the lung have not yet been searched could improve lung nodule detection.

Characterizing false negative errors

In order to better understand the cause of the observed false negative errors, we used the eye-tracking data to categorize each false negative in terms of how long the lesion in question was fixated. This method was pioneered by Kundel et al. (1978), who used it to analyze the different types of errors that occurred during search for lung nodules in 2-D chest radiographs. It has since been extended to a variety of medical images, including bone fractures (Hu, Kundel, Nodine, Krupinski, & Toto, 1994), and mammograms (Krupinski, 1996; Krupinski & Nishikawa, 1997). Kundel et al. proposed a three-way categorization of errors based on the fixation data. If a target never fell within the UFOV of the eyes, the resulting error was categorized as a search error. If a target attracted foveal vision for a long period of time (> 1 s), then, presumably, it was identified as a possible lesion worthy of scrutiny but was deemed innocuous or irrelevant. This is a decision error. Finally, targets that fell within the UFOV for less than a second were labeled as recognition errors. We used these definitions to categorize the errors made in our nodule detection task. As pointed out by Kundel et al. (1978), the distribution of these types of errors is entirely contingent upon the definition of useful field of vision. They adopted a circular region around the fixation point that subtended 5 DVA. This estimate of UFOV has been used in many of the papers that have followed Kundel's classic study (e.g., Krupinski, 1996; Krupinski & Nishikawa, 1997; Mello-Thoms et al., 2005; Nodine et al., 2002) and we adopted this estimate as well.

To assess the role that search strategy had on search error type, we analyzed the amount of time spent within 2.5 DVA of all false-negative nodules (a 5 DVA diameter UFOV). Scanner and drillers showed a significantly different distribution of error types, $\chi 2$ (2, $N = 529$) $= 20.77$, $p < 0.001$. This result is driven by a higher proportion of search errors (45% vs. 32%) in the scanner population and a higher rate of recognition errors in the driller population (23% vs. 39%). Furthermore, in the original Kundel et al. (1978) study, the distribution of errors was 30% search, 25% recognition, and 45% decision errors. We find a significantly different pattern in our data, $\chi 2$ (2, $N = 2$) $= 56.22$, $p < 0.001$. Decision errors are less common in our data, perhaps reflecting an improvement, brought about by the advent of CT imaging, in the ability to identify nodules once they are located. However, given the many differences between these studies, more work needs to be done to directly compare the distribution of errors in chest radiograph and chest CT cases.

Conclusions

Advances in computer technology have revolutionized diagnostic radiology, dramatically increasing the scope and variety of medical images. While we know a great deal about how search is accomplished in 2-D images, such as chest radiographs, we know much less about search in the sorts of volumetric images that are becoming standard in radiology reading rooms all over the world. Although adding a depth dimension to search undoubtedly enables searchers to detect smaller abnormalities, converting to this type of image also fundamentally changes the process by which the observer searches for targets. It is therefore important to use methods, such as the ones derived in this paper, to better understand how search is accomplished in these complex images. This is one of the first studies to analyze how radiologists search through a volume of images. By tracking the eye movements of these expert observers, we were able to uncover a series of novel results:

1. In a lung nodule detection task on chest CT scans, radiologists tended to adopt one of two search strategies, here termed drilling and scanning. These strategies were consistent within individuals across the multiple trials.
2. Drillers covered more of the lung than scanners.
3. Drillers correctly localized more lung nodules than scanners.
4. Driller and scanners differ in their distribution of error types. Most nodules, missed by scanners, were never looked at (search errors), whereas a plurality of nodules, missed by drillers were looked at briefly but not classified as possible nodules (recognition errors).

As eye tracking becomes easier and less intrusive, it will be valuable to acquire eye-movement data in true clinical settings. This information would allow a more definitive assessment of our hypothesis that the drilling strategy works better than the scanning strategy. Data of this sort could also be used to inform the training of the next generation of radiologists.

Keywords: visual search, eye movements, medical image perception

Acknowledgments

A portion of this research was funded by a grant from Toshiba Medical Systems, BWH Agreement No. A203079. Additional support was provided for TD by the NIH: 1F32EB011959-01 and JW: R01 EY017001. Also thanks to the gracious radiologist observers from Brigham and Women's Hospital and the American Board of Radiology who volunteered their time.

Commercial relationships: none.
Corresponding author: Trafton Drew.
Email: traftondrew@gmail.com.
Address: Department of Surgery, Brigham & Women's Hospital, Cambridge, MA, USA.

References

Aberle, D. R., Adams, A. M., Berg, C. D., Black, W. C., Clapp, J. D., Fagerstrom, R. M., et al. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, *365*(5), 395–409.

Andriole, K. P., Wolfe, J. M., Khorasani, R., Treves, S. T., Getty, D. J., Jacobson, F. L., et al. (2011). 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.

Armato, S. G., III, McNitt-Gray, M. F., Reeves, A. P., Meyer, C. R., McLennan, G., Aberle, D. R., et al. (2007). The lung image database consortium (LIDC): An evaluation of radiologist variability in the identification of lung nodules on CT scans. *Academic Radiology, 14*(11), 1409.

Berbaum, K. S., Franken, E. A., Dorfman, D. D., Miller, E. M., Caldwell, R. T., Kuehn, D. M., et al. (1998). Role of faulty visual search in the satisfaction of search effect in chest radiography. *Academic Radiology*, *5*(1), 9–19.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Cooper, L., Gale, A., Darker, I., Toms, A., & Saada, J. (2009). *Radiology image perception and observer performance: How does expertise and clinical information alter interpretation? Stroke detection explored through eye-tracking*. Paper presented at Proceedings of SPIE: Medical Imaging, February 7, 2009, Lake Buena Vista, FL, USA.

Cooper, L., Gale, A., Saada, J., Gedela, S., Scott, H., & Toms, A. (2010). *The assessment of stroke multidimensional CT and MR imaging using eye movement analysis: Does modality preference enhance observer performance?* Paper presented at Proceedings of SPIE: Medical Imaging, February 13, 2010, San Diego, CA, USA.

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the psychophysics toolbox. *Behavior Research Methods, 34*(4), 613–617.

Ellis, S. M., Hu, X., Dempere-Marco, L., Yang, G. Z., Wells, A. U., & Hansell, D. M. (2006). Thin-section CT of the lungs: Eye-tracking analysis of the visual approach to reading tiled and stacked display formats. *European Journal of Radiology*, *59*(2), 257–264.

Hu, C. H., Kundel, H. L., Nodine, C. F., Krupinski, E. A., & Toto, L. (1994). Searching for bone fractures: A comparison with pulmonary nodule search. *Academic Radiology, 1*(1), 25–32.

Krupinski, E. A. (1996). Visual scanning patterns of radiologists searching mammograms. *Academic Radiology*, *3*(2), 137–144.

Krupinski, E. A., & Nishikawa, R. M. (1997). Comparison of eye position versus computer identified microcalcification clusters on mammograms. *Medical Physics*, *24*(1), 17–23.

Kundel, H. L., Nodine, C. F., & Carmody, D. (1977). Search, recognition, and decision-making in lung nodule detection. *Investigative Radiology, 12*(5), 431–431.

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern-recognition and decision-making in pulmonary nodule detection. *Investigative Radiology, 13*(3), 175–181.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: Gaze-tracking study. *Radiology*, *242*(2), 396–402.

Kundel, H. L., Nodine, C. F., & Krupinski, E. A. (1989). Searching for lung nodules—Visual dwell indicates locations of false-positive and false-negative decisions. *Investigative Radiology, 24*(6), 472–478.

Kundel, H. L., Nodine, C. F., Thickman, D., & Toto, L. (1987). Searching for lung nodules—A comparison of human-performance with random and systematic scanning models. *Investigative Radiology, 22*(5), 417–422.

Kundel, H. L., Nodine, C. F., & Toto, L. (1991). Searching for lung nodules—The guidance of visual scanning. *Investigative Radiology, 26*(9), 777–781.

Manning, D., Barker-Mill, S., Donovan, T., & Crawford, T. (2006). Time-dependent observer errors in pulmonary nodule detection. *British Journal of Radiology*, *79*(940), 342–346.

Manning, D., Ethell, S., Donovan, T., & Crawford, T. (2006). How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography*, *12*(2), 134–142.

Mello-Thoms, C., Dunn, S., Nodine, C. F., Kundel, H. L., & Weinstein, S. P. (2002). The perception of breast cancer: What differentiates missed from reported cancers in mammography? *Academic Radiology*, *9*(9), 1004–1012.

Mello-Thoms, C., Dunn, S. M., Nodine, C. F., & Kundel, H. L. (2001). An analysis of perceptual errors in reading mammograms using quasi-local spatial frequency spectra. *Journal of Digital Imaging*, *14*(3), 117–123.

Mello-Thoms, C., Hardesty, L., Sumkin, J., Ganott, M., Hakim, C., Britton, C., et al. (2005). Effects of lesion conspicuity on visual search in mammogram reading. *Academic Radiology*, *12*(7), 830–840.

Nodine, C. F., Mello-Thoms, C., Kundel, H. L., & Weinstein, S. P. (2002). Time course of perception and decision making during mammographic interpretation. *American Journal of Roentgenology*, *179*(4), 917–923.

Oh, S. H., & Kim, M.-S. (2004). The role of spatial working memory in visual search efficiency. *Psychonomic Bulletin & Review*, *11*(2), 275–281.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Phillips, P., Manning, D., Crawford, T., Burling, D., Tam, C.-L., & Taylor, A. (2008). *Searching in axial and 3D CT visualisations*. Paper presented at

Proceedings of SPIE: Medical Imaging, February 16, 2008, San Diego, CA, USA.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology (Hove)*, *62*(8), 1457–1506.

Seltzer, S. E., Judy, P. F., Adams, D. F., Jacobson, F. L., Stark, P., Kikinis, R., et al. (1995). Spiral CT of the chest: Comparison of cine and film-based viewing. *Radiology, 197*(1), 73–78.

Smith, A. D., Gilchrist, I. D., Cater, K., Ikram, N., Nott, K., & Hood, B. M. (2008). Reorientation in the real world: The development of landmark use and integration in a natural environment. *Cognition, 107*(3), 1102–1111.

Smith, A. D., Hood, B. M., & Gilchrist, I. D. (2010). Probabilistic cuing in large-scale environmental search. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *36*(3), 605–618.

Solman, G. J., Cheyne, J. A., & Smilek, D. (2012). Changing perspective: Zooming in and out during visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 348–364.

Solman, G. J., Wu, N., Cheyne, J. A., & Smilek, D. (2013). In manually-assisted search, perception supervises rather than directs action. *Experimental Psychology*, *60*(4), 1–12.

Woodman, G. F., & Luck, S. J. (2004). Visual search is slowed when visuospatial working memory is occupied. *Psychonomic Bulletin & Review*, *11*(2), 269–274.

Young, A. H., & Hulleman, J. (2013). Eye movements reveal how task difficulty moulds visual search. *J Exp Psychol Hum Percept Perform*, *39*(1), 168–190.