Reader behavior in a detection task using single- and multi-slice image datasets

Asli Kumcu[a], Ljiljana Platiša[a], Milan Platiša[b], Ewout Vansteenkiste[a], Karel Deblaere[c], Aldo Badano[d], and Wilfried Philips[a]

[a]TELIN-IPI-IBBT, Ghent University, Ghent, Belgium;
[b]TASS N.V., Leuven, Belgium;
[c]Dept of Neuroradiology, Ghent University Hospital, Ghent, Belgium;
[d]CDRH-FDA, Silver Spring, MD, USA

ABSTRACT

We assess human reader behavior such as reading times and browsing trends in a signal detection experiment with synthetic single-slice (ss) and multi-slice (ms) image datasets of varying task complexity, defined in this study as the ratio of the background lump size to the signal width. Three dataset types were generated by inserting one 3D Gaussian target of fixed size into the center of 3D volumes of correlated Gaussian noise with three different kernel sizes. Corresponding signal intensities were determined separately for the three background types using the staircase method targeting an AUC of 0.7 for ss datasets. Non-expert human readers were presented with ss (central slice of the volume) and ms datasets (slice-by-slice viewing in a stack-browsing mode). Readers were aware of the target's approximate location within the slice or volume. Readers could scroll freely through the ms datasets at arbitrary speed and direction with no time limit. Experiments were conducted in a controlled viewing environment on a 5MP digital mammography display. AUCs were 0.68–0.73 for ss; 0.82–0.98 for ms datasets. Reading time (ms, ss), the number of repetitions through the stack (ms), and the average number of slices per repetition (ms) were assessed. Browsing speeds were in the range of 1–7 slices per second. Results show that readers spent the shortest time and fewest repetitions reading TP cases, with FP and FN cases requiring the most attention. The reported trends concur with earlier chest x-ray and mammography studies which report that readers fixate longer on regions subsequently rated incorrectly.

Keywords: Observer study, multi-slice, reader behavior, browsing patterns, visual search

1. INTRODUCTION

One of the active areas of investigation in medical image perception research is the analysis of visual search patterns of radiologists,[1] in order to understand how they perceive images and potential findings, and to potentially explain why they make errors in detecting and classifying lesions. Visual search patterns and image interpretation times in planar (2D) modalities, such as chest x-ray and mammography, have been studied with and without eye-tracking technology in order to understand the differences between correctly and incorrectly diagnosed findings, with promising findings.[2,3] Earlier studies on 2D modalities report that wrong decisions are accompanied by longer reading times compared to correct decisions. Reading times are often compared to the decision outcome, which we define as the decision made by the reader about signal presence or absence compared with the ground truth: true positive (TP), true negative (TN), false positive (FP), or false negative (FN). Nodine et al.[2] found that for interpretation of mammography cases as measured by eye-tracking dwell time, TP decisions were made quicker than FP decisions, and detection performance decreased as decision time increased. Manning et al.[3] reported that in pulmonary nodule detection, FN decisions took significantly longer in terms of dwell time than TN decisions, and TPs were decided quicker than FPs (except for novices, who took longer on TPs). Saunders and Samei[4] studied detection time in mammography without eye-tracking, and found that radiologists decided quicker for TPs than for FPs, and TNs took less time than FNs.

Send correspondence to Asli Kumcu, E-mail: asli.kumcu@telin.ugent.be

These and other studies indicate that there appears to be some correlation between reading/dwell time and the probability of a correct detection/classification outcome (TP, TN). In other words, cancers which attract long dwell times but are subsequently not reported are not due to an inability to detect the signal but rather from errors in recognition and decision[3] or interpretation.[5] Likewise, observers may interpret FPs as real lesions[5] but take longer to classify than TPs.[3]

The relationship between detection performance and visual search in volumetric (3D) modalities such as Computed Tomography (CT), Breast CT, Breast Tomosynthesis, Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI) has been less frequently investigated. Since these 3D modalities by nature generate a set of multiple 2D slices, the radiologists will typically scroll through, or browse, the dataset stack in a slice-by-slice fashion. Visual search in these modalities can thus be characterized not only by the search path on a single slice, but in addition the search patterns through the stack: which slices have been viewed, in which order, and for how long. Wang et al.[6] reported on a study where the main goal was to compare the detection performance of slice-by-slice stack browsing, maximum intensity projection (MIP), and stereoscopic display modes for chest CT datasets; when they looked at the browsing patterns, they found that in slice-by-slice stack browsing mode, 25% of the missed lesions (FNs) "received extra attention." Potential applications of measuring reading behavior in the clinical setting include improving training strategies for novice radiologists and incorporating feedback loops in the diagnostic process.[1,2,4]

The main question in this study is whether there is a relationship between reading time, browsing patterns through 3D (multi-slice) image stacks, and detection performance in multi-slice datasets, as has already been established for 2D (single-slice) image reading. A second question is whether the image characteristics influence reading times and browsing patterns.

In this work, we report the results of human reader behavior trends in a signal detection task on synthetic single-slice (ss) and multi-slice (ms) datasets. We examine the relationship between behavior aspects such as reading duration and multi-slice browsing patterns with decision outcome (TP,TN,FP,FN) and varying image data characteristics (the ratio of the background correlation length (frequency) and the signal width). This study uses synthetically generated backgrounds and signals as described in Sec. 2.1 in order to better control the image data characteristics. Limitations in interpretation due to this setup are discussed in Sec. 4.

The companion paper to this work by Platiša et al.[7] discusses the detection performance of the human observers in detail and the significance of the findings with respect to varying the image data characteristics.

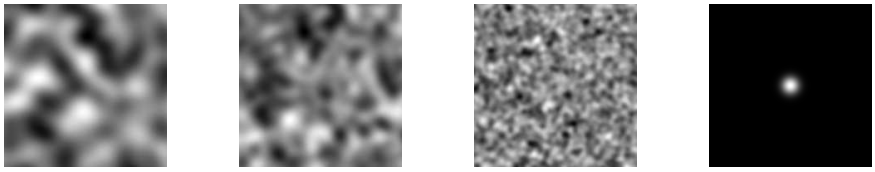
2. METHODS

2.1 Image Data

The stimuli were synthetic datasets generated by inserting a 3D Gaussian target of fixed size into the center of 3D volumes of correlated Gaussian noise. A region of 196 x 196 pixels x 63 slices was extracted from a volume of 256 x 256 x 256, and averaged over every 3 slices in the z-direction to simulate slice thickness, resulting in a volume of 196 x 196 pixels x 21 slices. Background and signal volumes were processed in the same manner. Three background types were generated, shown in Fig. 1, by selecting three different kernel sizes for the 3D Gaussian noise filters: $\sigma_{b1} = 11$ (B11), $\sigma_{b2} = 7$ (B07), $\sigma_{b3} = 3$ (B03). The width of the signal volume was kept constant: $\sigma_{s1} = 5$ (S05).

We introduce the term *task complexity* as the ratio between the background kernel width and the width of the signal; task complexity increases as the ratio of the background/signal widths decreases. Thus, the low complexity task is B11-S05, where the background lump size is considerably larger than the signal; the medium complexity task B07-S05, where the background lump size is somewhat larger than the signal; and the high complexity task B03-S05, where the background lump size is smaller than the signal.

Abnormal datasets were generated by adding the signal volume to the background volume, while *normal* datasets consisted of only the background volume. Multi-slice (ms) datasets consisted of the entire volume (all 21 slices), while single-slice (ss) datasets were generated by extracting the middle slice (slice 11) of the ms dataset. Note that the peak of the 3D Gaussian signal occurs in the middle slice (slice 11) of the volume and the center of the slice (pixel (98,98) in the slice). Corresponding peak signal intensities a_s for



(a) B11 (b) B07 (c) B03 (d) Signal

Figure 1: Example of the stimuli used in this experiment: (a)–(c) the three background types in decreasing order of kernel size and increasing order of *task complexity*, from left to right ($\sigma_{b1} = 11$, $\sigma_{b2} = 7$, $\sigma_{b3} = 3$); (d) is the central slice of the 3D target (the signal, $\sigma_{s1} = 5$).

the abnormal datasets were determined for each the three background types using the staircase method[8] targeting an Area Under the ROC Curve (AUC) of 0.7 for ss datasets. A slightly higher signal amplitude $a_{1.1s}$ (where $a_{1.1s} \approx 1.1a_s$) was also considered. There were six reading setups in total for both ms and ss datasets: $B11_{as1=60}$, $B11_{as2=65}$, $B07_{as3=120}$, $B07_{as4=130}$, $B03_{as5=135}$, and $B03_{as6=145}$.

2.2 Study design

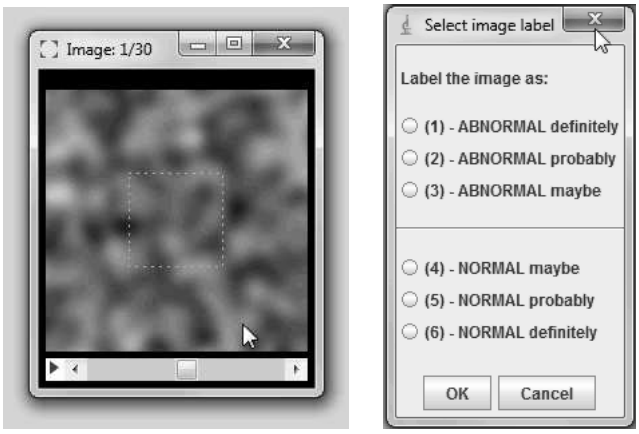
The study was designed for Multi-Reader Multi-Case (MRMC) Receiver Operating Characteristic (ROC) analysis, with Area Under the ROC Curve (AUC) as the figure of merit. Twenty-two trained human readers were recruited. Two of them were very experienced with the tasks (LP and AK), two were moderately experienced (they also participated in the pilot studies) and the rest were trained for the task. Two additional observers participated in the pilot readings only, including one expert neuroradiologist (KD). Of the 22 readers, 5 are female and 17 are male. They are between 24 and 36 years of age with a mean of 30.5. 2 readers are students, 14 are PhD student researchers, 4 are post-doctoral researchers, and 2 work as engineers in industry. Most participants are researchers in digital image processing, some have experience with medical image processing, and some have previous experience reading images or evaluating image quality.

The six reading setups explained in Sec. 2.1 (combination of three background widths and one signal width at two amplitudes) were used as stimuli, with two treatments per reading setup: ss and ms reading scenarios, each containing an equal number of normal (signal-absent) and abnormal (signal-present) datasets. The number of cases for each background type was chosen based on sample size estimations from a pilot study (not reported here). For each background type, readers were randomly assigned to the a_s or $a_{1.1s}$ signal level.

All readers participated in four reading sessions which took place on different days. The first session consisted of an explanation of the study followed by 50 training datasets per treatment per background type (a total of 300 datasets were rated). During training, readers received feedback after rating each dataset, indicating whether they were right or wrong; they were allowed to re-examine the dataset again after feedback before continuing to the next dataset. These ratings were not used for performance analysis.

The three following reading sessions consisted of one session per background type, in which both ms and ss datasets were rated. Each session started with a short reminder training segment consisting of 30 (B11) or 40 (B07, B03) datasets for each of the ss or ms testing trials. The session then continued with the actual rating trials, consisting of 64 (B11), 84 (B07), or 94 (B03) datasets per treatment. Only the results from the rating trials were used for the performance analysis. All testing and rating datasets were independent. Additional details about the numbers of datasets are given in the companion paper.[7]

An ImageJ plugin developed for this study was used to visualize datasets and collect reader ratings and behavior. Cases were rated using a 6 point confidence scale (definitely abnormal, probably abnormal, maybe abnormal, maybe normal, probably normal, definitely normal) as shown in Fig. 2b. Readers were aware of the approximate location of the target within the slice (ss and ms) and within the stack (ms). They could freely scroll through the ms datasets at arbitrary speed and direction with no time limit. A standard mouse with scroll wheel was used as the navigation interface; reader were allowed to use either the scroll wheel



(a) Viewing interface (b) Rating scale

Figure 2: Example of the user interface. (a) Image visualization and navigation window with region of interest (ROI) indicators – the dotted square for the ROI in the xy place; the dot in the upper-left corner of the window for the ROI in the z-direction. The signal is centered in the center of the ROI. Navigation through the stack is made either with the scroll wheel of the mouse or by clicking-and-dragging the button on the scroll-bar at the bottom of the image. (b) The 6-point confidence scale used for rating cases.

or the scroll-bar click-and-drag functionality to navigate through the stack. No image manipulation (zoom, pan, window level) was allowed. An example of the user interface is given in Fig. 2.

Experiments were conducted in a psychophysical test room[9] to ensure consistent experimental conditions: low ambient light and surface reflectance. The chair position was fixed such that readers viewed images at 50 cm from the display; they were allowed to lean back and forth. All sessions were conducted on a 5 MP medical-grade monochrome LCD display (Barco Coronis digital mammography display), which was warmed up for at least 1 hour before every experiment to ensure temperature stability.

2.3 Data Collection and Analysis

The ImageJ interface was used to collect the following data for each case read: reader ID, case ID, ground truth (whether a signal was actually present or not), confidence rating given by the reader, timestamp when reading started, and timestamp when rating completed. For ms cases, additional data was collected on each slice that was visited, together with a timestamp.

After the reader finished all four reading sessions, they answered a profile questionnaire including the following questions:

- Gender, Age, Profession
- Length and type of experience in image processing
- Length and type of experience with medical images
- Experience in reading images
- What time of the day do you prefer to work / when are you most efficient?
- Which image setup was most difficult to read? (a) B11 (b) B07 (c) B03
- For which image setup did you feel the least confident? (a) B11 (b) B07 (c) B03
- Was the training in the study useful? (a) No, not at all (b) Yes, somewhat useful (c) Yes, very useful

- How sufficient was the training? (a) Insufficient (b) Somewhat Sufficient (c) Almost Sufficient (d) Sufficient

Detection accuracy was measured using AUC as the figure of merit. DBM MRMC software version 2.3 build 3[10–13] was used to estimate the ROC curves from the confidence ratings and conduct the MRMC analysis with readers and cases as random effects. The trapezoidal/Wilcoxon method was used to estimate the per-reader ROC curves and corresponding AUCs. Values are reported as AUC_{mean} together with the 95% confidence interval (CI) range.

Associations between AUCs and reader profiles (for example age, experience level) were examined using the Kruskall-Wallis non-parametric one-way Analysis of Variance (ANOVA) test with significance level $\alpha = 0.05$.

Confidence ratings were processed in two additional ways as a complement to the AUCs. In order to quantify the improvement in performance for a ms dataset compared to its ss counterpart, the difference in the confidence ratings per case were compared using a plot inspired by the heat map charts of Freedman and Osicka.[14] The heat maps of Freedman and Osicka indicate the confidence ratings given for a set of cases; the differences in confidence ratings between two readings are also plotted in the same way. In our work, we compare the change in confidence ratings between the ms and ss datasets for the same case. The following methodology is used to compute the difference in confidence ratings between ms and ss cases, $\Delta Conf_{ms-ss}$:

- For *normal* cases: $\Delta Conf_{ms-ss} = \quad Conf_{ms} - Conf_{ss}$
  For example, if a *normal* case is rated "definitely normal" in ms mode ($Conf_{ms} = 6$) and "maybe normal" in ss mode ($Conf_{ss} = 4$), then $\Delta Conf_{ms-ss} = 2$, and a *green* square of size 2 is plotted. If the reader gives a higher confidence rating to the ss dataset, for example by rating the *normal* case as "maybe normal" in ms mode ($Conf_{ms} = 4$), and "definitely normal" for ss mode ($Conf_{ss} = 6$), then $\Delta Conf_{ms-ss} = -2$, and a *red* square is plotted.

- For *abnormal* cases: $\Delta Conf_{ms-ss} = -(Conf_{ms} - Conf_{ss})$
  For example, if an *abnormal* case is rated as "definitely abnormal" in ms mode ($Conf_{ms} = 1$), and the ss case is rated as "maybe abnormal" ($Conf_{ss} = 3$), then $\Delta Conf_{ms-ss} = 2$, and a *green* square of size 2 is plotted. If however the reader gives a higher confidence rating to the ss dataset, for example by rating the *abnormal* case as "maybe abnormal" in ms mode ($Conf_{ms} = 3$), and "definitely abnormal" for ss mode ($Conf_{ss} = 1$), then $\Delta Conf_{ms-ss} = -2$, and a *red* square is plotted.

Separately, confidence ratings were converted to a binary outcome by setting a hard threshold in the middle of the rating scale): cases rated between 1 and 3 were considered *abnormal* and those rated between 4 and 6 were labeled *normal*. Combining these with the reader's decision (*right* or *wrong* decision) gave four *decision outcome* categories: true positive (TP), true negative (TN), false positive (FP), or false negative (FN). Cases are later grouped by decision outcome to compare various behavioral measures.

The following behavioral measures were extracted from the browsing logs: reading duration for ms and ss (amount of time each slice was viewed, including rating time), the number of repetitions for ms (number of times the reader went back and forth through the stack while passing through the middle slice; abbreviated as *Reps*), the average repetition spread for ms (average number of slices per repetition), and the overall browsing speed.

Reading duration and the number of repetitions were analyzed using the method of *survival curves*.[3,4] Survival analysis is often used in healthcare models to assess the probability of surviving an *event* over a certain time when the incidence rate is not constant over time;[15] one example is the probability that an ex-smoker remains a quitter. In this work, a case is considered to *survive* until the time t at which it is rated; an *event* occurs when the reader gives the case a rating. The survival probability then corresponds to the length of time needed to rate a case. Survival curves for each decision outcome were computed using the product-limit method.[15] Reading times and repetitions were grouped by decision outcome and pooled over all readers and cases. The three task complexities were analyzed separately.

Table 1: Reader performance and feedback

Reading setup	Reader Performance		Reader Feedback			# Readers
	AUC_{mean} [95% CI]		Most difficult setup		Least confident setup	
	ss	ms	ss	ms	ss & ms	
$B11_{as1=60}$	0.71 [0.62,0.80]	0.98 [0.95,1.00]	3	2	3	11
$B11_{as2=65}$	0.73 [0.64,0.82]	0.97 [0.94,1.00]				12
$B07_{as3=120}$	0.71 [0.63,0.79]	0.96 [0.93,0.99]	0	0	2	11
$B07_{as4=130}$	0.68 [0.60,0.76]	0.98 [0.96,1.00]				12
$B03_{as5=135}$	0.68 [0.61,0.75]	0.82 [0.74,0.90]	16	17	15	12
$B03_{as6=145}$	0.70 [0.62,0.78]	0.85 [0.78,0.92]				10

The *median* and the $[25\%, 75\%]$ interquartile range were computed for reading times, number of repetitions, slices per repetition, and browsing speed. Values were pooled over all readers/cases and grouped by decision outcome, for each task complexity separately.

SAS Enterprise Guide 4[16] was used for all statistical analysis, except for the DBM MRMC ROC analysis.

3. RESULTS

3.1 Reader performance

The detection performance results for the six reading setups are given in Table 1. AUCs for ss datasets are in the range of 0.68–0.73 (as targeted in the pilot study). AUCs for ms datasets are 0.82,0.85 for the two B03 setups (high-complexity task) and 0.96–0.98 for the four B07 and B11 setups (medium- and low-complexity tasks, respectively).

Since the gain in performance for the ms treatment is highest for the medium- and low-complexity tasks, these reading setups have greater benefit from the availability of additional slices in the ms reading mode. This effect can also be seen in Fig. 3, which shows the differences in the confidence ratings between ms and ss datasets for all cases. Confidence ratings for ms datasets are in general higher than the ss datasets for the medium- and low-complexity tasks (light-colored squares). Conversely, in the high-complexity task there are more datasets where readers rate the ss dataset the same or higher confidence rating than the ms dataset (dashes and dark-colored squares).

Also shown in Table 1 is the feedback the readers gave on the difficulty level and the overall level of confidence they felt for the three levels of task complexity. Most readers indicate that the high-complexity task is the most difficult one, and they feel least confident for this setup. Reader feedback parallels their performance and confidence ratings: readers feel that the high-complexity task is the most difficult setup and show the smallest performance improvement in ms mode compared to ss mode.

We also examine a possible correlation between reader profiles and reader performance. No significant relationship between AUC and the following factors has been found: image processing experience, medical imaging experience, previous experience evaluating images, profession, or age ($\alpha = 0.05$). Interestingly, the female gender is positively associated with higher AUC for the high-complexity task (Kruskall-Wallis non-parametric one-way ANOVA, $\chi^2_{B03-as5} = 6.23$, $\chi^2_{B03-as6} = 4.36$, $df = 1$, $0.01 < p < 0.05$).

Platisa et al.[7] contains additional analysis of the differences in reader performance from this study and discusses the relationship between task performance, ss versus ms reading mode, and task complexity in more detail.

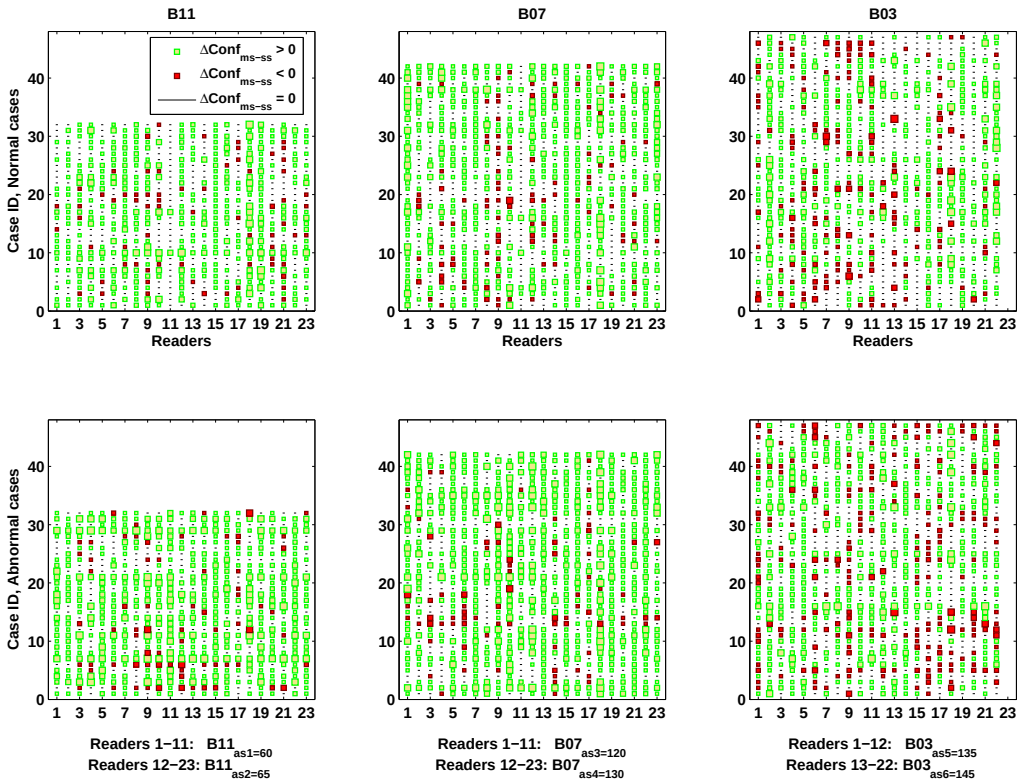
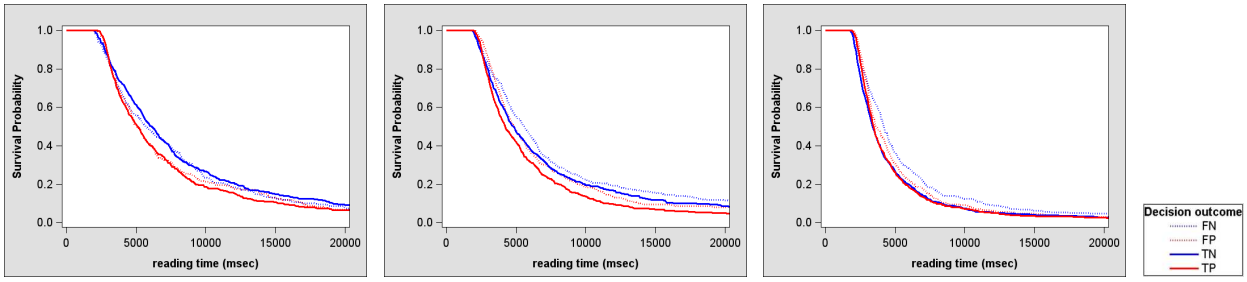


Figure 3: Change in confidence ratings $\Delta Conf_{ms-ss}$ between the ms and the ss datasets of the same case. Charts from left to right are in increasing order of task complexity (B11, B07, and B03); charts on the top row are for normal cases, charts on the bottom are for abnormal cases. For each case, a green (light-colored) square indicates that the reader gives a higher confidence rating on the ms dataset, whereas a red (dark-colored) square indicates that the ss dataset receives a higher confidence rating; the size of the square is proportional to the difference in confidence ratings between the ms and ss datasets. A dash indicates no change in the confidence rating. A 1–6 scale is used for the confidence ratings (see Fig. 2b), where 1 = *definitely abnormal* and 6 = *definitely normal*. The difference in confidence ratings for *normal* cases was computed as: $\Delta Conf_{ms-ss} = Conf_{ms} - Conf_{ss}$, while the difference for *abnormal* cases was computed by taking the negative value of the difference: $\Delta Conf_{ms-ss} = -(Conf_{ms} - Conf_{ss})$. See Sec. 2.3 for further explanation and examples.



(a) B11-ss (b) B07-ss (c) B03-ss

Figure 4: Survival plots for the time it took readers to rate cases, ss reading mode. Values are pooled over all readers and cases, and grouped by decision outcome. Maximum reading time shown is 20 seconds.

3.2 Reading duration

Median reading times for ss cases are shown in Table 2 and the survival curves in Fig. 4. Reading times for ms cases are in Table 3 with the corresponding survival curves in Fig. 5. Cases are pooled over all readers and cases and grouped by decision outcome (TP,TN,FP,FN) for each task complexity setup separately. Reading times of the two signal intensities per task complexity are pooled, as there is no significant difference in performance between the two setups for all background types (see Table 1 for corresponding AUCs and 95% CIs).

Reading times for ss cases are shorter than ms cases, as would be expected. Among ss cases, FN cases are consistently scored the slowest (5.7, 5.6, 4.2 seconds for B11, B07, B03). In addition, a slight trend of increased reading times for incorrectly scored cases (FN,FP) can be seen especially for B03 and B07. For B11, although correctly and incorrectly scored cases cannot be distinguished, cases rated as *abnormal* are scored quicker than those rated *normal* (about 5 seconds for TP and FPs, and about 6 seconds for TN and FNs). Overall the high-complexity datasets (B03) are rated the quickest (3.5 seconds for TPs, up to 4.2 seconds for FNs) with the smallest (25%,75%) interquartile range. Readers are slightly slower on the medium- and low-complexity tasks (4.2, 5.1 seconds for TPs; 5.6, 5.7 seconds for FNs, respectively B07,B11), and have a larger interquartile range, indicating more variance in the reading times. The survival curves given in Fig. 4 exhibit the same trends as indicated by the curve drop-off rate as well as the differences between the decision outcome curves. Quicker curve decay indicates shorter reading times and smaller variance in reading times.

Among ms cases, differences between correctly and incorrectly scores cases are more apparent, with less difference between the three levels of task complexity. TPs are consistently read the quickest (8.5, 7.9, 8.5 seconds for B11, B07, B03), followed by TNs (12.7, 11.4, 10.9 seconds). FPs are consistently read the slowest (14.5, 14.4, 12.0 seconds). FNs typically take more time than TPs and fall in between the reading times of TNs and FPs. Overall the B03 datasets are rated approximately 2 seconds quicker than the B11 and B07 datasets on TN, FP, and FNs cases. The variance in reading time for incorrectly scored cases is also slightly smaller for B03 (but there are also more observation points in this sample).

3.3 Browsing trends

The results of various browsing measures for ms cases are given in Table 3 and include, in addition to the total reading time:

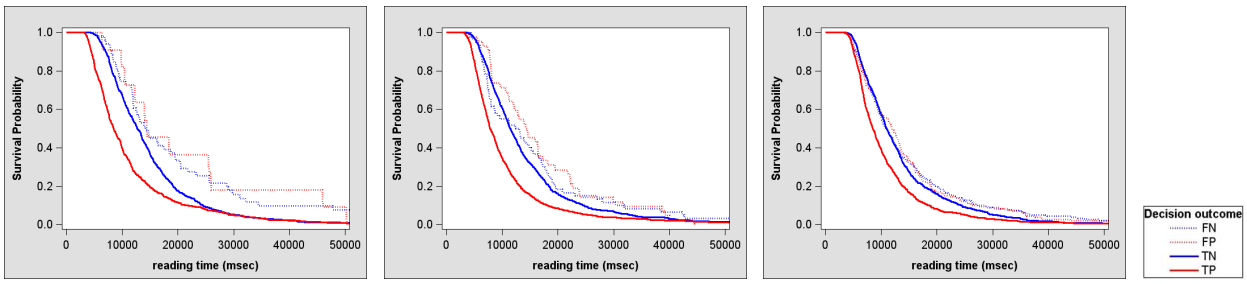
- # Reps: the median number of repetitions through the stack (which indicates how many times readers reversed scrolling direction after going through the middle slice, where the peak of the 3D signal is located)
- # Slices/Rep: the median of the average number of slices per repetition (which indicates how far readers browsed in the stack before reversing the scrolling direction)

Table 2: Reading times for ss cases (Values are pooled over all readers and cases and are given as *median* [25%, 75%] interquartile range.)

Reading Setup	Decision Outcome	Reading time (sec)	# cases
B11	TP	5.1 [3.4, 8.3]	481
	TN	6.0 [3.7, 10.3]	543
	FP	5.0 [3.6, 8.9]	193
	FN	5.7 [3.4, 9.7]	255
B07	TP	4.2 [3.0, 6.9]	641
	TN	4.8 [3.1, 8.3]	618
	FP	4.8 [3.5, 8.2]	348
	FN	5.6 [3.5, 9.2]	325
B03	TP	3.5 [2.7, 5.1]	694
	TN	3.4 [2.5, 5.2]	689
	FP	3.7 [2.9, 5.6]	345
	FN	4.2 [2.9, 6.5]	340

Table 3: Reading times and browsing patterns for ms cases (Values are pooled over all readers and cases and are given as *median* [25%, 75%] interquartile range. See Sec. 2.3 for an explanation of the behavior measures.)

Reading Setup	Decision Outcome	Reading time (sec)	# Reps	# Slices/Rep	Speed (slices/sec)	# cases
B11	TP	8.5 [5.9, 13.0]	3 [1,5]	2.3 [2.0, 4.0]	1.9 [1.5, 3.0]	684
	TN	12.7 [8.9, 17.6]	3 [2,6]	4.7 [3.2, 7.0]	2.0 [1.6, 3.8]	725
	FP	14.5 [10.4, 26.0]	7 [2,13]	4.9 [3.3, 8.5]	2.3 [1.7, 5.1]	11
	FN	14.2 [9.6, 25.6]	6 [3,9]	3.4 [2.2, 7.1]	1.9 [1.4, 2.6]	52
B07	TP	7.9 [5.7, 11.9]	3 [2,4]	2.3 [2.0, 4.0]	1.9 [1.6, 2.7]	906
	TN	11.4 [8.0, 16.9]	3 [2,5]	4.8 [3.1, 6.5]	2.0 [1.6, 3.4]	923
	FP	14.4 [8.3, 22.1]	5 [2,8]	4.7 [2.7, 8.0]	1.7 [1.3, 3.5]	42
	FN	12.7 [7.3, 18.5]	4 [2,8]	5.5 [3.0, 8.1]	2.7 [1.5, 7.0]	60
B03	TP	8.5 [6.3, 12.8]	3 [2,5]	2.3 [2.0, 4.0]	1.8 [1.5, 2.4]	762
	TN	10.9 [7.5, 15.7]	3 [2,5]	3.6 [2.3, 5.5]	1.9 [1.4, 2.6]	855
	FP	12.0 [7.1, 17.2]	4 [2,7]	3.7 [2.3, 5.5]	1.9 [1.4, 4.1]	179
	FN	10.8 [7.7, 17.5]	4 [2,6]	3.9 [2.3, 7.0]	2.2 [1.5, 4.8]	272



(a) B11-ms (b) B07-ms (c) B03-ms

Figure 5: Survival plots for the time it took readers to rate cases, ms reading mode. Values are pooled over all readers and cases, and grouped by decision outcome. Maximum reading time shown is 50 seconds.

- Speed (slices/sec): the median browsing speed, in terms of slices per second (which indicates how fast readers navigated through the stack)
- # cases: the number of cases classified at that decision outcome

Cases are pooled over all readers and cases and grouped by decision outcome (TP,TN,FP,FN) for each task complexity setup separately. Values for the two signal intensities per task complexity are also pooled, as there is no significant difference in performance between the two setups for all background types (see Table 1 for corresponding AUCs and 95% CIs). Survival curves for the number of repetitions are given in Fig. 6, and represent the probability that a case "survives" a certain number of repetitions before being rated (see Sec. 2.3 for an explanation of survival curves).

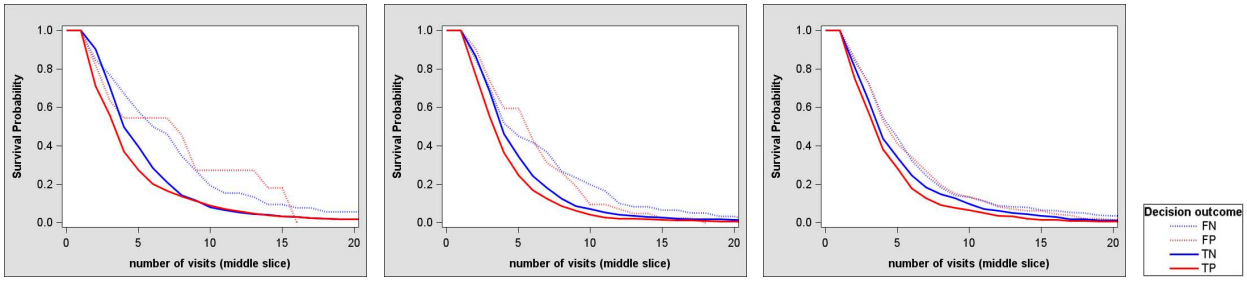
TP and TN cases require fewer repetitions through the stack (about 3 reps), than FNs and FPs, which require between 1 and 4 additional repetitions. The interquartile range is also larger by several repetitions, for FNs and FPs. The differentiation between decision outcomes can also be appreciated in the survival curves in Fig. 6: curve decay is fastest for TPs, a bit slower for TNs, and slowest for FNs and FPs. When comparing task complexities, no significant differences for the different decision outcomes are seen, except for incorrectly scored cases at the low-complexity task (B11), which require a larger number of repetitions and have a larger interquartile range than the medium- and high-complexity tasks (B07 and B03).

Similarly, the number of slices per repetition are fewest for TPs (2.3 slices/rep for all three task complexities). TNs, FPs, and FNs all require more slices per repetition (3.4–4.9 slices/rep) and have a larger interquartile range than TPs. Among reading setups, in general the high-complexity task requires the fewest slices per repetition and has the smallest interquartile range for all decision outcomes, compared to the medium- and low-complexity tasks.

Lastly, readers' browsing speeds hover around 2 slices per second, with an interquartile range of 1.3–7.0 slices per second over all readings. Based on the measurements of two medical-grade 5 MP monochrome LCD displays made by Liang and Badano,[17] the maximum transition time between any two gray levels was around 70 ms (which would correspond to a maximum browsing speed of about 14 slices/sec). Therefore, the browsing speeds of the readers in this study are slow enough (<14 slices/sec) to ensure a minimal if at all significant effect due to the slow temporal response of the LCD.

4. DISCUSSION

The difference between ss and ms detection performance and confidence ratings show an inverse relationship with task complexity: the ss-ms AUC difference is smaller for high-complexity datasets (B03). Since ss AUC is around 0.7 for all task complexity setups, the smaller improvement in ms performance on the high-complexity task indicates that readers find less benefit from having additional slices available in ms mode. The change in confidence ratings, shown in Fig. 3 also indicates that the ms reading mode allows readers to give more confident ratings on the medium- and low-complexity tasks (B07 and B11). Reader feedback



(a) B11-ms (b) B07-ms (c) B03-ms

Figure 6: Survival plots for the number of repetitions through the stack (number of times the reader browsed back-and-forth while passing through the middle slice of the stack; the peak of the 3D Gaussian signal is centered in the middle slice), ms reading mode. Values are pooled over all readers and cases, and grouped by decision outcome.

is also associated with performance and task complexity: readers feel that the high-complexity task is the most difficult one, and the task for which they have the least confidence.

Analysis of visual search patterns can be a useful tool for understanding how people perceive images, and to potentially explain why they make errors in detecting and classifying lesions; this is of special interest in the field of medical imaging.[1] In this study, reader behavior is found to be associated with decision outcome (TP,TN,FP,FN). TP cases consistently require the least attention (shortest reading times, fewest repetitions through the stack, and slices per repetitions). Cases that are rated correctly (TP,TN) in general associated with shorter reading times, fewer number of repetitions, and fewer slices per repetition, than incorrectly rated cases (FP,FN). In addition, cases that are classified as *abnormal* (signal present) are also associated with shorter readings times and fewer repetitions, compared to cases classified as *normal* (signal absent) . These trends concur with earlier studies for ss modalities (chest x-ray, mammography) which report that wrong decisions are accompanied by longer dwell/reading times compared to correct decisions;[2–5] as well as a ms study showing that more attention is given to missed detections.[6]

Among ss cases, readers give FN cases the most attention, whereas reading times for FP cases are more similar to TPs. It is possible that in the ss mode, readers interpret FPs as real lesions and therefore the short FP reading times reflect this. Another possibility is that the longer reading times for FN cases indicate that the reader is attracted to something in the image (e.g. a signal) but subsequently decide to rate it *normal* anyway; this mechanism has been reported in medical imaging as errors of decision[3] or interpretation,[5] rather than an inability to detect the signal.

Conversely, when reading ms cases, although FN cases receive more attention than TP cases, more attention is consistently given to FP cases in terms of all three browsing parameters. One explanation is that some cases contain 3D structures that appear like a signal, as can be seen in the confidence plots in 3: for several cases, all readers consistently give lower confidence ratings to the ms version of the case. However, this behavior is more exaggerated for the medium- and low-complexity tasks. This may be due to the overall high performance levels for these two tasks: readers on average give higher confidence ratings to the ms cases for medium- and low-complexity tasks and have AUCs of around 0.97. Therefore during reading they are probably in general very confident as to whether a signal is present or not. Then, when they do encounter cases with false signals, this probably creates much more doubt than if they are not as confident in the task. These effects are not seen as strongly in the high-complexity task: FNs are given a similar amount of attention as TNs, probably indicating errors of detection rather than interpretation; FPs are also given more attention than TPs but cannot be well distinguished from TNs and FNs. It is possible that the lack of additional benefit from multiple slices in the high-complexity tasks is reflected in browsing patterns: readers give less attention to incorrectly marked cases compared to the medium- and low-complexity tasks, either because the task is more difficult or they feel less confident, and as a result they "give-up" earlier.

Trends in reading time and browsing patterns that are correlated with decision outcome may have a

potential application in improving reader performance.[1,2,4] For example, cases that receive excessive attention could be flagged to the user indicating that they have spent too long examining the case (or too many repetitions through a particular slice), and are probably likely to give an incorrect decision. Given feedback, the reader could then adjust their strategy in order to improve detection and interpretation, for example by asking a colleague for a second opinion, examining the data from another perspective (e.g. coronal rather than axial slice), or asking for additional follow-up if the case remains inconclusive.

Finally, we would like to point out several limitations of the study. First, the reading times measured in this study only indicate how long the case went without being rated – it is possible that readers did not spend 100% of that time gazing at the image, navigating through the dataset, and deciding how to rate it (since eye-tracking is not used in this study). Therefore reading times should be interpreted with caution. Nevertheless, for ms cases we have shown that additional behavior measures, such as the number of repetitions through the stack and the number of slices per repetition, may also be useful indicators of browsing behavior and are in fact are associated with the trends in reading times. Second, the images used in this study are artificially generated correlated noise backgrounds with Gaussian signals. The application of the results to images and signals with different statistical properties, and especially clinical images, should be made very carefully. Likewise, the signal location in this study is fixed in both the slice and the volume in order to control testing conditions. Readings times and browsing trends will most likely be very different (and potentially longer) when the signal location is unknown. The characterization of browsing trends will also require different measures when the signal location is unknown: readers may exhibit different browsing behavior when scrolling through the dataset attempting to detect signals versus when they interrogate a potential finding (in this study, only the latter behavior is considered since the reader is aware of the approximately location of the signal in the volume).

Our current work-in-progress includes the analysis of reading behavior by reader experience level as well as confidence rating. Preliminary results shows that readers with medical imaging experience have faster ratings times, higher AUCs, and slower browsing speeds.

In the future, the study of browsing patterns by task complexity should include ms reading setups with similar performance levels; in this study there is a wide range of performances in ms mode, from 0.84 AUC to 0.98 AUC, making meaningful behavioral comparisons between task complexities in ms mode difficult. In addition, future study designs should incorporate more clinically relevant aspects, including the use of radiologists as readers, the use of clinical data, and a randomized signal-location task.

5. CONCLUSION

This study examines the relationship between the behavior exhibited by readers while reading and rating cases, in terms of ss and ms reading times and ms browsing trends, with performance (AUC) and decision outcome (TP,TN,FP,FN). We also examine the relationship between reading behavior and task complexity as defined by the ratio between the lump width of 3D correlated noise backgrounds and the kernel width of 3D Gaussian signals.

Overall, the results from this study suggest that reader behavior is linked to detection performance, particularly during stack-browsing through volumetric (multi-slice) datasets. In general, TP cases are read faster or with fewer repetitions than TN cases, and incorrectly marked cases are given more attention than correctly marked cases. The increased reading times, number of repetitions, and slices per repetition accompanying FN decisions may suggest that while readers suspected the presence of a lesion, further browsing through the dataset did not help them detect the signal confidently. The trends found in this study in general concur with earlier studies for ss modalities (chest x-ray, mammography) which report that wrong decisions are accompanied by longer dwell and reading times compared to correct decisions;[2–5] as well as a ms study showing that more attention is given to missed detections.[6] These results encourage further investigation into the mechanisms that underlie differences in reading behavior and browsing patterns during stack-browsing of multi-slice datasets in a signal detection task.

REFERENCES

[1] Krupinski, E. A. and Berbaum, K. S., "The Medical Image Perception Society update on key issues for image perception research," *Radiology* **253**(1), 230–233 (2009).

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

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

[4] Saunders, R. S. and Samei, E., "Improving mammographic decision accuracy by incorporating observer ratings with interpretation time," *British Journal of Radiology* **79**(Special Issue 2), S117–S122 (2006).

[5] Mello-Thoms, C., "The problem of image interpretation in mammography: effects of lesion conspicuity on the visual search strategy of radiologists," *British Journal of Radiology* **79**(Special Issue 2), S111–S116 (2006).

[6] Wang, X., Durick, J., Lu, A., Herbert, D., Golla, S., Foley, K., Piracha, C., Shinde, D., Shindel, B., Fuhrman, C., Britton, C., Strollo, D., Shang, S., Lacomis, J., and Good, W., "Characterization of radiologists search strategies for lung nodule detection: Slice-based versus volumetric displays," *Journal of Digital Imaging* **21**, 39–49 (2008).

[7] Platisa, L., Kumcu, A., Platisa, M., Vansteenkiste, E., Deblaere, K., Badano, A., and Philips, W., "Volumetric detection tasks with varying complexity: human observer performance," *Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment* **8318**(1), 83180S, SPIE (2012).

[8] Garcia-Perez, M. A., "Yes-no staircases with fixed step sizes: Psychometric properties and optimal setup," *Optometry & Vision Science* **78**(1), 56–64 (2001).

[9] Marchessoux, C. and Kimpe, T., "15.3: Specificities of a psycho-physical test room dedicated for medical display applications," *SID Symposium Digest of Technical Papers* **38**(1), 971–974, SID (2007).

[10] Dorfman, D., Berbaum, K., and Metz, C., "Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method," *Investigative Radiology* **27**, 723–731 (1992).

[11] Hillis, S., "Monte carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification," *Academic Radiology* **12**, 1534–1541 (2005).

[12] Hillis, S., "A comparison of denominator degrees of freedom for multiple observer ROC analysis," *Statistics in Medicine* **26**, 596–619 (2007).

[13] Hillis, S. L., Berbaum, K. S., and Metz, C. E., "Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis," *Academic Radiology* **15**(5), 647–661 (2008).

[14] Freedman, M. and Osicka, T., [*The Handbook of Medical Image Perception and Techniques*], ch. 21: Perceptual effects of CAD in reading chest radiographs, 290–303, Cambridge University Press, New York (2010).

[15] Rosner, B., [*Fundamentals of Biostatistics*], Cengage Learning, 7th ed. (2011).

[16] SAS Institute Inc., Cary, NC, "SAS Enterprise Guide 4." software.

[17] Liang, H. and Badano, A., "Temporal response of medical liquid crystal displays," *Medical Physics* **34**(2), 639–646 (2007).