Suspicious minds: effect of using a lesion likelihood score on reader behaviour with interactive mammographic CAD

DOI:
10.1117/12.2556472

Document Version
Accepted author manuscript

Suspicious minds: effect of using a lesion likelihood score on reader behaviour with interactive mammographic CAD

Ethan Du-Crow[a], Susan M Astley[a], Johan Hulleman[b]
[a]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Stopford Building, Oxford Road, Manchester, UK
[b]School of Biological Sciences, Division of Neuroscience and Experimental Psychology, University of Manchester, Manchester, UK

ABSTRACT

Computer-Aided Detection (CAD) systems are used to help readers in interpreting screening mammograms. Traditional use of CAD in mammography involves an expert reader searching the image initially unaided, and then once again with the aid of CAD prompts that automatically indicate suspicious regions. An alternative approach is interactive CAD, where prompts are only displayed when readers query a suspicious region for which a prompt is available. These prompts are typically displayed with a given confidence of malignancy for that region. Two non-expert observer studies of interactive CAD were conducted to investigate its effect on the visual search of synthetic images containing microcalcification clusters. Experiment 1 (n=44) used no-CAD and interactive CAD conditions, whereas Experiment 2 (n=43) used interactive CAD in both conditions but in one there was an additional 'image score' denoting the likelihood that an image contained a cluster. In both experiments, the addition of interactive CAD (Experiment 1) and an image score (Experiment 2) did not change sensitivity or specificity compared to no-CAD and interactive CAD-alone, respectively. In Experiment 1, the higher the confidence value for a given prompt, the more likely a participant was to act on it. This effect was stronger for true prompts than false prompts. In Experiment 2, participants spent longer viewing images with higher image scores. When no prompt was available, they were more likely to make false positive errors on those images. However, decisions made on available prompts were influenced primarily by confidence values of the prompts rather than overall image score.

Keywords: Breast Cancer, Computer-Aided Detection, CAD, interactive CAD, Visual Search, Eye-tracking, mammography

1. INTRODUCTION

Mammographic CAD systems automatically detect and prompt abnormal regions. Traditional CAD, where the reader reviews the image first unaided, and then again with the help of CAD, has been shown to improve sensitivity over single-reading alone, but at the cost of an unacceptably increased recall rate[1]. Beyond this traditional approach, a wide variety of prompting techniques have been proposed and investigated. Jorritsma et al. (2015) proposed ways of improving trust between readers of medical images and CAD systems[2]. This includes providing confidence ratings, which could improve trust in the system if CAD prompts that the reader would dismiss anyway have a low confidence rating.

In addition, a local rationale could be provided that provides an explanation of why that region has been marked. An example of this is ImageChecker's PeerView mode, which enhances queried regions, outlining masses and individual calcifications[3]. 'Analogue CAD', where the prompts themselves indicate the probability of being a target, was proposed by Cunningham et al. (2016). It was shown to increase sensitivity for non-expert observers finding clusters of dots with a specific colour, compared to binary CAD (where prompts mark potential targets with no extra information)[4]. In that study, the analogue marks outlined targets and distractors and the colour of the outline corresponded to the probability of being a target. This is similar ImageChecker's EmphaSize feature which uses prompt size to denote the likelihood of malignancy[3].

Interactive-CAD incorporates some of these approaches. Prompts are only displayed on the image when the reader queries a region and this has the potential to improve sensitivity without increasing false positives. Prompts are accompanied with confidence ratings, denoting the probability that the prompted region marks an abnormality. This method of prompting

has been demonstrated to lead to a significant improvement in the partial AUC compared to unaided search and traditional CAD[5]. Compared to a traditional CAD approach, interactive-CAD can afford more available false positive prompts since they are not all necessarily going to be seen. Therefore it can operate at a higher sensitivity. One study showed that radiologists' AUC was significantly improved from 83% unaided to 86% aided[6], and another reported a 7.2% increase in radiologists' AUC compared to unaided search[8]. Importantly, recall rate in these studies did not change with the introduction of CAD.

There has yet to be a study into the effect of interactive CAD on visual search behaviour and decision making. We therefore conducted two studies with non-expert observers to investigate the interaction between both prompt confidence and image score on the one hand and visual search strategy on the other. By studying interactive-CAD in a lab setting, we aimed to investigate the mechanism behind the reported improvement in observer performance with this approach, and thus how it could be further improved.

2. MATERIALS AND METHODS

Two visual search tasks were created where non-expert observers searched for microcalcification clusters in $1/f^{1.5}$ noise distributions. In Experiment 1, non-expert observers searched images in two conditions: unaided (referred to as no-CAD); and using interactive CAD (referred to as CAD). In Experiment 2, interactive CAD was used in both conditions, but in one condition an overall image score was provided below the image denoting the likelihood that the image contained a target; these are referred to as the CAD and CAD+Score conditions.

The experimental setup and images used in these studies were the same as a previous study[8]. Synthetic images were created in MATLAB using open-source code[9], with a spectral roll-off factor (image roughness) of 1.5. These backgrounds resemble the glandular component of mammograms, although they lack linear structures. Targets were malignant microcalcification clusters that were extracted from magnified images of slices of mastectomies[10]. Clusters were inserted into the images by multiplying the cluster pixels by the background pixels, placed randomly at a point within a 5×5 grid on the image. The grid was defined in such a way that the entirety of the cluster was within the image if it was placed at the edge of the grid. Images were displayed at 800×800 pixels on a 21-inch ViewSonic VX2268WM LCD monitor with a resolution of 1680×1050 pixels. Viewing distance was 73cm. Eye movements were tracked with an EyeLink 1000 desktop eye tracker, and observers used a chin rest to restrain head movements. For classification of fixations, gaze must be contained within 1.5 degrees visual angle from the fixation start point for a minimum of 100ms. The experiment code was written using PyGaze (v0.6.0)[11], with a PyGame back-end (v1.9.2).

A total of 100 background images were generated (set A), and then rotated by 180 degrees to form a second set of 100 (set B). From set A, 40 images were randomly selected for cluster insertion, along with the corresponding rotated images from set B. Initially, 80 unique cluster targets were randomly split into two sets of 40 and were inserted into the images of set A and B. To avoid predictability of target position, the position of the clusters on the grid were assigned randomly but distributed evenly so that for both cluster target sets each of the 25 possible locations was used once, and 15 were used twice. An example experimental image is shown in Figure 1. Using data from a previous study with these images[8], the target images in sets A and B were swapped until the distribution of target detectabilities and the median target detection rate were matched between the two image sets.

All participants had corrected or corrected-to-normal vision (20/20 or higher) as confirmed by the Freiburg Vision Test (FrACT)[12]. Participants were initially presented with a training set of images with on-screen feedback to enable them to understand target appearance and experimental set-up. The training set consisted of images of isolated clusters (not inserted in the synthetic backgrounds), followed by 10 images with targets of varying difficulty. The training also introduced how the prompts worked. The order of presentation of the 200 experimental images was randomised for each participant. Presentation was blocked by experimental condition. Half of the participants started with the condition 1 (no-CAD/CAD) and the other half started with condition 2 (CAD/CAD+Score). For both condition 1 and 2, half of the participants saw images drawn from set A and half saw images drawn from set B. Participants were randomly allocated into these groups.

Target prevalence was 40%, and out of the 40 cluster targets in each image set, 4 were selected to be used as unprompted targets. The prompts placed on background structures (non-targets) were in the same positions in sets A and B (i.e. rotated 180 degrees for set B). Participants were not informed of the exact target prevalence, but knew it was between 1% and 50%. Participants placed a marker on areas that they believed to be a cluster, and were also able to remove markers if they

changed their mind. No time limit was enforced, but participants were advised not to spend longer than around one minute on each image.

In the conditions with CAD, participants were able to 'query' suspicious regions by clicking on them, which displayed a prompt (if available on that region) and a given confidence score (between 1 and 100) for the prompt. For the prompted condition, the prompts marked 90% of the clusters, with an average of 2 false prompts per image. The prompts consisted of a coloured circle, where the colour varied from yellow (low confidence) to red (high confidence). Observers were not informed of the exact sensitivity of the prompts, but were told that the prompts were not perfect and may mark regions that are not true clusters.

Prompt confidence for true prompts was assigned using data from a previous study[8]. For each target, prompt confidence was the associated difficulty in detecting it according to the proportion of participants that successfully located the target in the unaided condition of the previous experiment. For false prompts, the same distribution of target detectability was used to randomly assign confidence values. For Experiment 2, image score assignment for target present images was done as follows: 28/40 assigned a score of 10, 3/40 a score of 9, 2/40 a score of 8, and 1 target per score from 7 to 1. Target absent images were equally split between the ten score categories with 6 images in each. The score distribution was the same in each image set.

Statistical tests used a bootstrap technique[13], which treated observers and images as random effects. From this, the t-statistic and p-value were calculated, as well as the 95% confidence intervals, and statistically significant differences were identified. This approach provides a conservative estimate of an effect.

Forty-two participants (median age 22, age range 18-58, 34 female) and 43 participants (median age 21, age range 18-47, 35 female) were recruited for Experiment 1 and 2, respectively, and informed consent was obtained. Sixteen participants in each experiment were undergraduate psychology students and received course credit for taking part. The rest received £10 in exchange for their time. The study was approved by the University of Manchester Research Ethics Committee (2018-4586-6410).

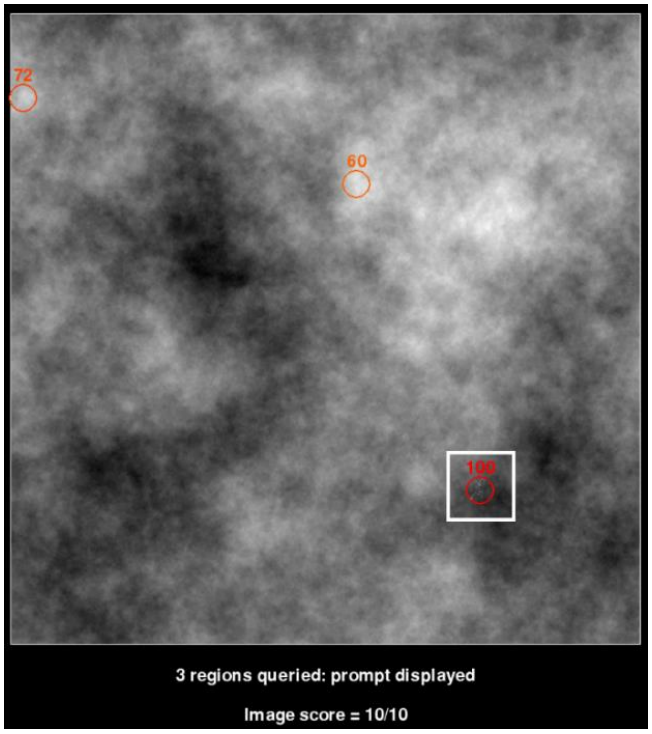

Figure 1. Example of an image with a cluster target (outlined by the white square for illustrative purposes), a true prompt on the target with a prompt confidence score of 100, and two false prompts with confidence values of 60 and 72, respectively. In addition to the displayed scores, the prompt colour also reflects the confidence score, from yellow (low) to red (high). The image score was displayed below the image as shown, but only available in the CAD+Score condition of Experiment 2.

3. RESULTS

Results are given for the no-CAD and CAD conditions in Experiment 1, and the CAD and CAD+Score conditions in Experiment 2. Image coverage was calculated using a circle of radius 2.5 degrees visual angle around the centre of fixations (the typical value associated with useful visual field in mammography visual search[14]) and points within that region were considered covered. Sensitivity was defined as the proportion of targets correctly located. False positive rates per image indicate the mean number of incorrect observer-placed marks per image.

As shown in Table 1, the mean trial time was not significantly different in the CAD condition compared to no-CAD in Experiment 1, although, numerically, the CAD condition took 1.35s longer. There was also no difference in the percentage image coverage. Overall sensitivity was numerically worse with CAD and went from 67.3% in no-CAD to 65.7% in CAD, but this difference of 1.55% was not significant. The mean number of false positive responses per image was 0.26 for no-CAD and 0.22 for CAD, a non-significant difference of 0.04.

Table 1: Results for Experiment 1, given for the No-CAD and CAD conditions. The 95% CIs, *t*-statistic, and *p*-value were calculated using a bootstrap technique over readers and images. Statistically significant results are highlighted in bold.

	No-CAD	CAD	Difference [95% CIs]	***t***-value	***p***-value
Trial time (s)	12.09	13.43	1.35 [0.09−2.57]	1.78	0.083
Image coverage (%)	65.07	65.97	0.90 [-1.41−3.12]	0.65	0.517
Sensitivity (%)	67.3	65.7	1.55 [-4.1−7.3]	0.45	0.658
FPs/image	0.26	0.22	0.04 [-0.02−0.11]	1.02	0.313

Table 2 shows the results for Experiment 2, where there was an increase in mean trial time with the introduction of the image score of 1.87s. The three other variables only differed numerically. The percentage image coverage was 2.52% larger in the CAD+Score condition, but this difference failed to break the .05 threshold. Overall sensitivity was 64.1% for CAD and 67.1% for CAD+Score, a non-significant difference of 3.0%. The false positive rate was 0.40 per image for the CAD+Score condition and 0.33 for CAD, a numerical difference of 0.06.

Table 2: Results for Experiment 2, given for the CAD and CAD+Score conditions. The 95% CIs, *t*-statistic, and *p*-value were calculated using a bootstrap technique over readers and images. Statistically significant results are highlighted in bold.

	CAD	CAD+Score	Difference [95% CIs]	***t***-value	***p***-value
Trial time (s)	13.86	15.73	1.87 [0.4−3.3]	2.13	**0.039**
Image coverage (%)	65.5	68.0	2.52 [0.25−4.77]	1.83	0.075
Sensitivity (%)	64.1	67.1	3.0 [-1.4−7.5]	1.07	0.289
FPs/image	0.33	0.40	0.06 [0.01−0.11]	1.79	0.081

As shown in Table 3, the number of regions queried by participants ranged from 1.18 to 1.44 per image between the three CAD conditions of the two experiments. Participants were more likely to place a marker in reaction to true prompts than false prompts, with an average of 87.3% for true prompts vs 68.0% for false prompts across the three conditions. In Figure 2 it can be seen how the confidence value affected how participants acted on prompts. The higher the prompt confidence, the more likely a participant was to believe that it marked a target and subsequently place a marker on that region. True prompts were more likely to be acted upon than false prompts for all confidence values, although this was less distinguishable for confidence values below 60 in the CAD condition of Experiment 2.

Table 3: Observer behaviour by prompt type for Experiments 1 and 2: true prompts (where the prompt marks a target), false prompts (where the prompt does not mark a target), and no prompts (where the participant queries but no prompt is available for that location). For each prompt type, the median number of queries for that type is given, along with the median percentage of queries of that type where a marker is placed by the participant.

		Median number of regions queried per image (range)	True prompts		False prompts		No prompts	
			Queried out of 36 (median)	Acted on (%)	Queried out of 200 (median)	Acted on (%)	Queried (median)	Acted on (%)
Experiment 1	CAD	1.44 (0.29−5.31)	15	92.8	10	71.3	107.5	1.2
Experiment 2	CAD	1.18 (0.03−5.09)	11	83.3	11	68.9	96	3.6
	CAD+Score	1.39 (0.03−5.20)	10	85.7	11	63.9	106	4.1

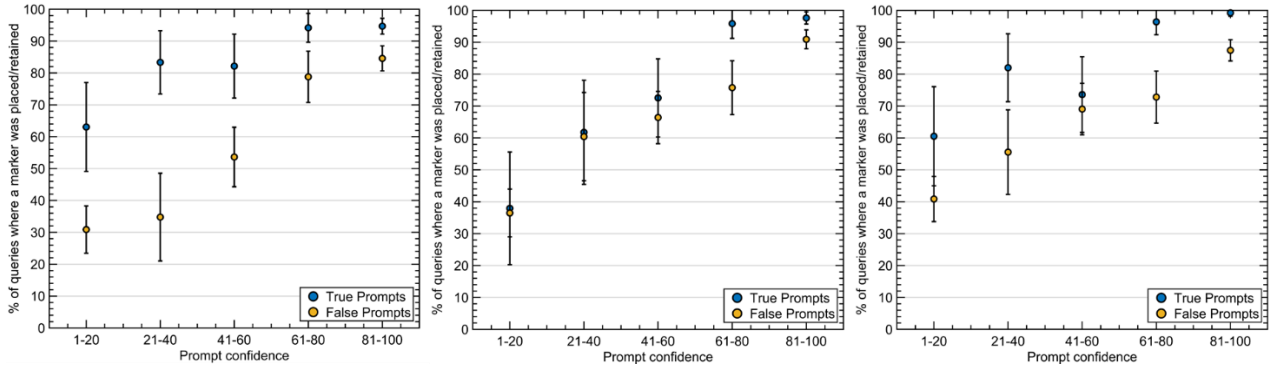

Figure 2: Mean percentage of queries where a participant subsequently clicked on that location or retained a marker they had already placed there. Results are for Experiment 1 (left: CAD) and Experiment 2 (middle: CAD & right: CAD+Score). The error bars are the standard errors across participants.

As shown in Figure 3 (left), in Experiment 2, the higher the image score in the CAD+Score condition, the longer participants spent on those images compared to the same images in the CAD condition without a score, and vice versa. This correlation was significant ($r(8)=0.96$, $p<0.001$). Above a score of 5 there was a >10% increase in viewing time, and below 4 there was a reduction of >5%. In Figure 3 (right), it can be seen that an image score ≥5 led to a greater number of false positive errors compared to those same images in the CAD condition, with a >34% increase above a score of 7. An image score of 1, 2 or 4 led to a decrease in the number of false positives, with a >13% reduction. This correlation between the change in false positive error rate and image score was again significant ($r(8)=0.92$, $p<0.001$). For the 28 images with a score of 10, the sensitivity was numerically larger by 4.6% compared to those same images in the CAD condition, but this was not significant ($p=0.47$). For the remaining score values for target-present images, there were too few images for each score (3 or below) to reliably calculate sensitivity change.

In Figure 4, the interaction between prompt confidence and image score is shown. For a given prompt, the likelihood the participant will act on it by placing a marker was mostly invariant with overall image score, and appears to be primarily influenced by the confidence value of that prompt. So, when a prompt was available for the area queried, it almost completely overruled the general tendency of participants to rely on the image score (as shown in Figure 3, right). However, please note that most of the time the image score was all participants had in Experiment 2, since most areas queried did not contain a prompt.

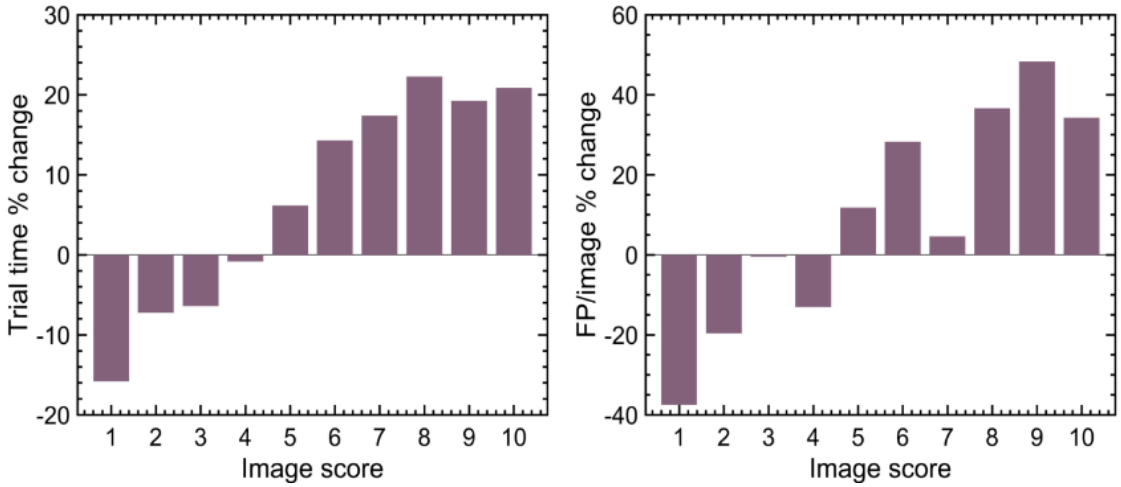

Figure 3: Results from Experiment 2. The percentage change in trial time (left) and false positive responses per image (right) when going from CAD to CAD+Score as a function of image score.

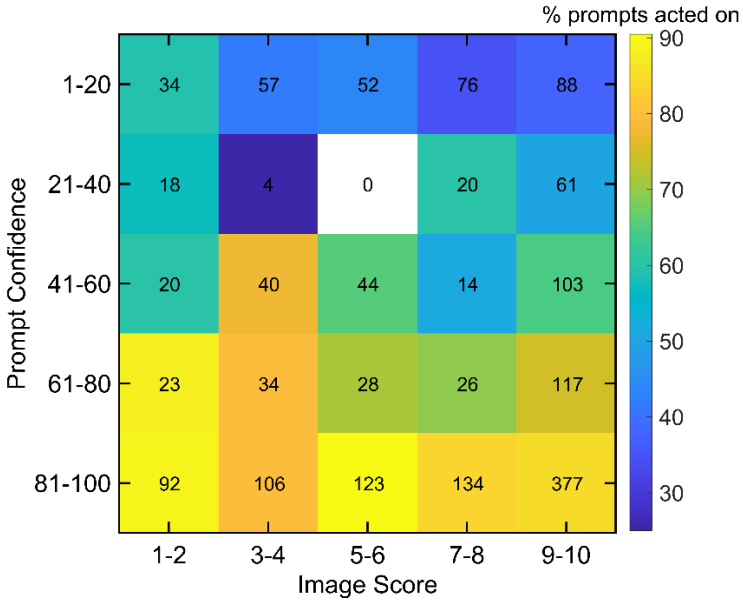

Figure 4: Results from the CAD+Score condition in Experiment 2. The percentage of true and false prompts acted on (marked by a participant) as a function of prompt confidence and image score. Both prompt confidence (1-100) and image score (1-10) have been divided into five bands. The % acted on ranges from 25.0% to 90.5%. The number in each square is the number of data points available for that image score/prompt confidence combination.

4. DISCUSSION

This study investigated how interactive-CAD affects image search and observer performance compared to unaided search, and whether providing an overall image score has an additional effect. In Experiment 1, CAD had no positive effect on the sensitivity. Moreover, participants using CAD also achieved similar specificity to no-CAD. Overall search of images did not differ in terms of coverage, but participants tended to spend longer on images in the CAD condition (Experiment 1), and spent significantly longer in the CAD+Score compared to CAD alone (Experiment 2). In both experiments, participants were more likely to believe prompts with higher confidence values, for both true and false prompts, with true prompts being more likely to be acted on than false prompts for all confidence values for Experiment 1 and confidence values above 60 for Experiment 2. In Experiment 2 participants spent longer on images with a high overall image score,

and conversely spent less time on low score images. When no prompt was available on a query, participants made more false positive errors on images with a higher image score, and fewer false positive errors on those images with a low image score. When a prompt was available, the prompt confidence determined whether any particular prompted region was considered to be a target, rather than the overall image score.

The lack of improvement in sensitivity in Experiment 1 is in contrast with clinical studies[6,7]. The interactive CAD model used in our study was comparable to those in the clinical studies, with an equivalent sensitivity, false positive rate, appearance of prompts, and target prevalence. However, there are some key differences between the studies. Firstly, our study used non-expert readers with synthetic images, rather than experts reading mammograms. This may have meant that our observers were less motivated in finding the targets compared to radiologists searching for cancers. But since here we are interested in within-participant differences between the conditions this should not have an effect on our results, as long as motivation was consistent between conditions. One could argue that the lack of improvement with CAD is simply due to the fact that if a participant did not see the target, then they could not query it, and therefore did not benefit from the information CAD would have provided on that region. While this contradicts clinical results, this argument may hold some truth in our study, since the synthetic images lack the clear anatomical structure present in mammograms, and there were no locations where targets were more likely to be found. Therefore, our participants may have struggled to make informed choices on where they should be looking and should subsequently query. While the prompts were placed on the images such that they were often in areas that may raise suspicion, the majority of queries made were on areas without available prompts.

When compared to our previous study with a more traditional CAD approach[8], the CAD 'signal' in this study was likely to be stronger than previously, since the ratio of true to false positive prompts was 1:1.36 in this study, compared to 1:1.96 in the previous study. This means participants were probably more likely to believe a prompt when they found one, and in fact (on average) 68.0% of the time false prompts were acted on in the three CAD conditions across Experiments 1 and 2, whereas only 14% were acted on with the traditional CAD of the previous study. Despite this, the specificity in Experiment 1 did not change with the addition of CAD, unlike in the previous experiment where it was significantly reduced. The most likely explanation is that although participants had a higher trust in the prompts, the number of false prompts that were seen by the participants in the current experiments was far fewer: a median of between 10 and 11 were fixated in this study versus 48 in the previous experiment. In addition to this, in the three CAD conditions of the current study participants queried a median of between 96 to 107.5 regions that did not have a prompt available, and on average on only 3.0% of these queries was a marker placed. Therefore, participants may have reduced false positive responses by querying regions they thought might be a target, and used no prompt as a reassurance it was in fact not a target.

The addition of an overall image score in Experiment 2 had a clear impact on observer behaviour. The changes in trial time for different image scores was comparable to that reported in a clinical study with interactive-CAD, where providing a score was shown to reduce reading time by >5% for cases with a score of 5 or lower and increase reading time by >5% for image scores of 9 or 10 [6]. This 'prioritising' of images deemed to be more important did not confer any tangible benefit in Experiment 2 though. For image scores above 7, there was a large increase in the false positive rate (>34%), but there was no apparent benefit for sensitivity. For instance, for images with a score of 10 there was no significant difference between the score and no-score conditions. It may have been expected that prompt confidence and image score would act in concert. I.e. an image score of ≤2 and a prompt confidence ≤20 would have the lowest proportion of participants acting on those regions, and an image score ≥9 and prompt confidence ≥81 would have the highest. However, we actually observed that, when available, prompt confidence was the most important factor, and the proportion of prompts acted on was not correlated with image score for a given prompt confidence band.

In future, a further study will need to establish whether the results observed in this experiment hold for expert readers when analysing mammograms, with an eye-tracking setup capable of recording gaze over a full radiology workstation.

5. CONCLUSIONS

Our study of non-expert observers found that interactive CAD did not improve sensitivity compared to no-CAD, and the addition of an image score did not improve sensitivity compared to CAD-alone, despite both of these additions leading to longer trial times. Specificity was also unchanged with the introduction of interactive CAD. Participants in both experiments were more likely to believe a prompt when it had a higher lesion likelihood score, and were more likely to

make a false positive error for higher image scores. Prompt confidence was the key factor influencing whether a participant would mark a prompted region, much more so than overall image score.

ACKNOWLEDGEMENTS

This work was supported by the Medical Research Council; grant reference number MR/N013751/11. Dr Astley is supported by the NIHR Manchester Biomedical Research Centre. We thank Dr Lucy Warren for providing the simulated calcification clusters. The clusters were developed as part of the OPTIMAM research programme funded by Cancer Research UK, and the Engineering and Physical Sciences Research Council Cancer Imaging Programme in Surrey, in association with the Medical Research Council and Department of Health (England).

REFERENCES

[1] Noble, M., Bruening, W., Uhl, S. and Schoelles, K. (2008). Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. *Archives of Gynecology and Obstetrics*, 279(6), pp.881-890. [doi: 10.1007/s00404-008-0841-y].

[2] Jorritsma, W., Cnossen, F. and van Ooijen, P. (2015), 'Improving the radiologist–CAD interaction: designing for appropriate trust', *Clinical Radiology* 70(2), pp.115–122. [doi: 10.1016/j.crad.2014.09.017].

[3] Roehrig, J. (2005), 'The manufacturer's perspective', *The British Journal of Radiology*, 78(1), pp.S41-S45. [doi: 10.1259/bjr/25058162].

[4] Cunningham, C., Drew, T. and Wolfe, J. (2016), 'Analog Computer-Aided Detection (CAD) information can be more effective than binary marks', *Attention, Perception, & Psychophysics* 79(2), pp.679–690. [doi: 10.3758/s13414-016-1250-0].

[5] Hupse, R., Samulski, M., Lobbes, M., Mann, R., Mus, R., den Heeten, G., Beijerinck, D., Pijnappel, R., Boetes, C. and Karssemeijer, N. (2013), 'Computer-aided Detection of Masses at Mammography: Interactive Decision Support versus Prompts', *Radiology* 266(1), pp.123–129. [doi: 10.1148/radiol.12120218].

[6] Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.-J., Schilling, K., Heywang-Köbrunner, S. H., Sechopoulos, I. and Mann, R. M. (2019), 'Detection of breast cancer with mammography: Effect of an artificial intelligence support system', *Radiology* 290(2), pp.305–314. [doi: 10.1148/radiol.2018181371].

[7] Watanabe, A., Lim, V., Vu, H., Chim, R., Weise, E., Liu, J., Bradley, W. and Comstock, C. (2019), 'Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography', *Journal of Digital Imaging*, 32(4), pp.625-637. [doi: 10.1007/s10278-019-00192-5].

[8] Du-Crow, E., Astley, S. and Hulleman, J. (2019), 'Is there a safety-net effect with computer-aided detection?', *Journal of Medical Imaging*, 7(02), p.1. [doi: 10.1117/1.JMI.7.2.022405].

[9] Methven, T. and Qi, L. (2012), 'Texturelab Edinburgh – Resources – Scripts'. Available at: http://www.macs.hw.ac.uk/texturelab/resources/scripts/ [Accessed 25 Jan. 2018].

[10] Warren, L., Mackenzie, A., Cooke, J., et al. (2012), 'Effect of image quality on calcification detection in digital mammography,' *Medical Physics* 39(6), pp.3202–3213. [doi: 10.1118/1.4718571].

[11] Dalmaijer, E., Mathôt, S. and Van der Stigchel, S. (2014), 'PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eye tracking experiments', *Behavior Research Methods* 46, pp.913–921. [doi: 10.3758/s13428-013-0422-2].

[12] Bach, M. (1996), 'The Freiburg visual acuity test—automatic measurement of visual acuity', *Optometry and Vision Science* 73(1), pp.49–53.

[13] Efron, B., and Tibshirani, R. J. [An introduction to the bootstrap], New York, N.Y.; London: Chapman & Hall (1993).

[14] Kundel, H. and Nodine, C., "Modeling visual search during mammogram viewing," Proc. SPIE 5372, 110-115 (2004). [doi: 10.1117/12.538063].