Low Prevalence Search for Cancers in Mammograms: Evidence using Laboratory Experiments and Computer Aided Detection

Melina A. Kunar[1], Derrick G. Watson[1], Sian Taylor-Phillips[2] & Julia Wolska[1]

1) Department of Psychology, The University of Warwick, Coventry, CV4 7AL, UK
2) Warwick Medical School, The University of Warwick, Coventry, CV4 7AL, UK

Email: m.a.kunar@warwick.ac.uk

Tel: +44 (0)2476 522133

Running Title: CAD in LP Mammogram Search

Word Count: 14150

Abstract

People miss a large proportion of targets when they only appear rarely. This Low Prevalence (LP) Effect could lead to serious consequences if it occurred in the real world task of searching for cancers in mammograms. Using a novel mammogram search task, we asked participants to search for a pre-specified cancer (Experiments 1-2) or a range of masses (Experiments 3-5) under high or low prevalence conditions. Experiment 1 showed that an LP Effect occurred using these stimuli. Experiment 2 tested an over-reliance hypothesis and showed that the use of Computer Aided Detection (CAD) led to fewer missed cancers with a valid CAD prompt yet, a large proportion of cancers were missed when CAD was incorrect. Experiment 3 - 5 showed that false alarms also increased when searching for a range of masses and that CAD reduced miss errors when it correctly cued the target but increased miss errors and false alarms when it did not. Furthermore, when a mass fell outside the CAD prompt it was more likely to be misidentified. No LP Effect was observed with the addition of CAD when people were asked to search for a range of targets. Theories and implications for mammogram search are discussed.

Public Significance Statement

This study suggests that search for low prevalence cancers in mammograms can be aided by Computer Aided Detection (CAD) but there is also a cost. CAD highlights areas in mammograms that are likely to contain a mass. When CAD successfully prompts a mass, observers are better at detecting it. However, observers also show an over reliance on CAD, and are more likely to miss cancers if the cue is incorrect.

Introduction

People often perform visual search tasks in everyday life, some of which have important consequences for society. For example, baggage screeners search for dangerous items when screening luggage at an airport and radiologists search for indicators of cancer when examining mammograms. In these two examples the target (e.g. weapon or cancer) typically appears rarely. This rarity is important because previous work has found that miss errors dramatically increase when the prevalence of a target is low (Wolfe, Horowitz & Kenner, 2005).

Researchers study how well people perform visual search tasks in the laboratory by asking participants to search for a pre-specified target among distracting items. Typically, in these search tasks scientists measure reaction times (RTs) and error rates (e.g. Watson & Kunar, 2010, 2012). When the target appears frequently (e.g. 50% of the time), miss errors, where the target is physically present but the participant responds that it is absent, tend to be low (around 5%, Wolfe, 1998). In contrast, Wolfe, et al. (2005) investigated search for targets that only appeared rarely. In their study participants searched a grey-scale display containing semi-transparent, overlapping objects. Participants were asked to search for a tool that could appear 50% of the time in a High Prevalence (HP) condition and 1% of the time in a Low Prevalence (LP) condition. Consistent with previous work, their results showed that miss errors in the HP condition were low (7%). However, when the target appeared only rarely in LP conditions, miss errors dramatically increased (up to 30%). This increase in miss errors with a decrease in target prevalence is known as the 'Low Prevalence' (LP) Effect and has been found to be robust across a variety of stimulus types, including search through perceptually simple displays (e.g., search for horizontal line among vertical lines, Rich et al., 2008, or search for a T among L distractors, Rich et al., 2008, Kunar, Rich & Wolfe, 2010, Russell & Kunar, 2012) and complex

displays (e.g., searching for a weapon in baggage screening images Van Wert, Horowitz & Wolfe, 2009, Mitroff & Biggs, 2014).

There have been several theories proposed to explain the LP Effect. One theory is that it occurred due to a speed-accuracy trade-off (Fleck & Mitroff, 2007). There is some evidence for this. RTs for absent trials in LP conditions tend to be faster than those in HP conditions (e.g., Wolfe et al., 2007, Rich et al., 2008, Russell & Kunar, 2012). Thus participants missed more targets because they responded too quickly to search the display properly. However, the results could not be entirely explained by a speed-accuracy trade off as the LP Effect was still observed under conditions in which participants were explicitly told to slow down during the task if they were responding too fast (Wolfe et al., 2007) or were required to wait a minimum amount of time before responding (Rich et al., 2008). Fleck and Mitroff (2007) also suggested that participants might make more response-execution motor errors under LP conditions. To investigate this they added a self-correction key so that participants could change their response if they were aware they had pressed the wrong key. This self-correction option reduced the LP Effect showing that part of the effect could be due to motor errors. However, it did not explain the whole effect as subsequent work showed that, although self-correction led to fewer miss errors, the LP Effect still occurred even after motor errors were accounted for (e.g. Van Wert, et al., 2009, Kunar, et al., 2010, Russell & Kunar, 2012).

Other theories have suggested that the LP Effect occurs because people were quitting their search too soon, before they had a chance to find the target. On the whole this would be a good strategy because responding target absent, without performing an exhaustive search, would lead to a correct outcome on the majority of trials in LP search. However, early quitting of search would, of course, lead to a high rate of miss errors when the target was actually present.

Evidence for this theory comes from eye movement data. Under LP conditions the majority of miss errors were shown to occur when participants quit their search before fixating the target (Rich et al., 2008, see also Peltier & Becker 2016, who found a decrease in quitting threshold and an increase in identification errors using eye movements under LP conditions). Furthermore, Signal Detection Theory (SDT, Green and Swets, 1967, Macmillan & Creelman, 2005) has also been used to understand the mechanisms behind LP search. It has been found that under LP search peoples' sensitivity in finding the target (as measured by d') did not change with target prevalence (Wolfe et al., 2007). Instead having the target only appear rarely changed people's criteria (as measured by c) so that people were less willing to respond 'target present' in LP conditions (Wolfe et al., 2007, Van Wert et al. 2009, Wolfe & Van Wert, 2010, Russell & Kunar, 2012). Wolfe and Van Wert (2010) proposed a Multiple Decision Model of visual search where the LP Effect could be explained by a change in two factors: (i) a change in decision criteria involving perceptual decisions of each item inspected, and (ii) a change in the quitting threshold leading to reduced search times on target absent trials. The Multiple Decision Model also predicted that there would be fewer false alarms at LP because participants' would be less willing to respond that a target was present under low prevalence conditions.

There are a number of important real world tasks in which observers need to search for a low prevalence target. Relevant to the work presented in the current study is the example of radiologists searching a mammogram for a cancer. In the UK alone over 2 million women are screened for breast cancer each year (NHSBSP, 2015). Radiologists examine the mammograms produced from this screening for evidence of cancerous indicators and a recent large scale study (over 1 million women screened) has found no decrease in cancer detection rate with radiologist time on task (Taylor-Phillips et al., 2016). However, the percentage of cancers

occurring in these mammograms are rare (estimated at 0.86% in the UK, Rayat, 2016). There have only been a few studies examining whether the LP Effect occurs in medical image reading and these have produced mixed results. Reed et al. (2011) found that when examining posteroanterior chest images for pulmonary lesions eye movements changed across prevalence rates whereby there was a greater number of fixations and a longer scrutiny time under higher prevalence conditions. Furthermore an LP Effect has been reported in medical searches where cytologists searched for an abnormality in a cervical screening test (Evans et al., 2011). Evans, Birdwell and Wolfe (2013) also investigated whether an LP Effect occurred in real-world mammography by inserting 50 mammograms containing a cancer, over the course of 6 months, into the normal workflow of a breast cancer screening service. The results showed that observers missed more cancers at LP than they did at HP. However, other work has suggested changing the prevalence rate does not affect search in a clinical setting (e.g., Gur et al., 2003, Hancock 2013, Taylor-Phillips et al., 2016). For example, Gur et al. (2003) found that varying the prevalence rate of a range of abnormalities (e.g., nodules, rib fractures, alveolar disease) had minimal effect on search performance (as measured by area under the receiver operating curve, ROC).

Note that there are very few studies directly investigating how changing prevalence rates affect search in medical images. This is because it is often difficult to detect miss errors in clinical settings, thereby making it difficult to measure the LP Effect in situations like real-time mammography. By definition screening centres will not know if a cancer has been missed until either the woman becomes symptomatic or the developed cancer is picked up at the next screening session (which can be up to 3 years later). One way around this is to embed ‘truth cases’ into a clinical setting, as in the work conducted by Evans et al., (2013). However, there are important and strict ethical issues surrounding real-world manipulations which limit what

can be studied. First, for ethical reasons, the number of truth cases that can be embedded into a clinical setting is very low so as not to cause disruption to the reader's normal success rate of abnormality detection. Second, data collection is a lengthy and limited process (e.g., as mentioned above due to the nature of the task, Evans et al., 2013, took 6 months to collect 50 data points of target-present cases). Third, it is often very difficult to recruit radiologists for these studies due to the lack of their availability. For example, some breast screening studies only used 2 or 3 readers (either radiologists or trained non radiologist film readers) for each observational study (e.g. Freer et al., 2001). Fourth, there is a limit to which manipulations can be studied in the clinical setting as any intervention is ethically bound to make sure that it does not interfere with a reader's typical reading technique.

The above points mean that data collected in a clinical setting may result in low experimental power because of issues related to collecting sufficient data for reliable analyses. In an ideal world the best way to perform this research would be to have an appropriate number of radiologists search for enough pre-known truth cases under LP conditions to gather sufficient data. However, given the lack of availability of radiologists and the fact that LP experiments are very time-consuming (due to the increased number of target absent trials needed in LP search) this is often not a feasible option. It is clear that researchers need to find other ways to address this question. For example, to investigate this issue we can either: (i) insert experimental paradigms into a clinical setting, or (ii) simulate clinical tasks in a laboratory setting. Both techniques have their uses. Evans et al. (2013) investigated the former technique. In the current study we examine the latter.

We created a laboratory based mammogram search task that was used to investigate the LP Effect. We trained participants to search for a spiculated cancerous mass (Experiments 1 - 2)

or a range of cancerous and benign masses (Experiments 3 – 5) embedded in real mammogram images. The use of mammogram images in LP search provides a number of practical and theoretical benefits. From a practical perspective it is important to study, replicate and extend work showing an LP Effect when searching mammograms, given that missing a cancer can have serious medical consequences. From a theoretical perspective, our work examined the effect of prevalence rates on false alarms. Wolfe and Van Wert's (2010) Multiple Decision Model (2010) proposed that under LP conditions false alarms should decrease. This aspect of the model has been difficult to test previously due to the very low number of false alarms produced overall when the target item was clearly defined and unambiguous (e.g. Wolfe et al., 2005, Rich et al., 2008, Kunar et al., 2010, Russell & Kunar, 2012)[1]. We investigated this aspect of the model in Experiments 3 – 5 in which we used a range of targets, thereby making the target identity from trial to trial more ambiguous. With less clearly defined targets we expected to find an increase in the number of false alarms from which we tested the predictions made by the Multiple Decision Model. False alarms are important to investigate in mammogram search as in clinical settings they could lead to invasive and unnecessary medical procedures (including needle biopsies of breast tissue). These procedures could increase the patient's experience of worry associated with breast cancer for up to a year afterwards (Aro, 2000) and also incur extra and unnecessary financial costs associated with increased medical tests.

Our work also examined the use of Computer Aided Detection (CAD) in mammogram search. CAD systems use algorithms that are designed to detect cancers by highlighting suspicious areas to readers (e.g., Bennett et al., 2006). Several clinical studies have been implemented to

[1] Wolfe and Van Wert (2010) investigated whether false alarms changed with prevalence rate. However, here they used conditions where the prevalence of the target was 98% rather than 2%. Very few studies have investigated false alarm data using low target prevalence rates (e.g. 2%).

examine the effectiveness of CAD. However, the clinical ability of CAD to improve cancer detection in mammography reading remains controversial (e.g., Fenton et al., 2007; Philpotts, 2009). Gur et al. (2004) compared two CAD systems along with an in-house scheme and found that all systems could be improved upon and mass detection differed depending on which CAD scheme was used. Bennett et al. (2006) completed a literature review that compared the accuracy of single reading with CAD to double reading procedures (in which two readers independently read each mammogram) across eight studies. Their results found that given the very large differences in the studies (e.g., methodologies, prevalence rate, number of readers, experimental findings etc.) the benefits or costs of CAD were inconclusive. Furthermore, although a multicentre randomised control trial (CADET II) found comparable cancer detection rates for single reading with CAD compared to double reading procedures (Gilbert et al., 2008) it was difficult to calculate miss errors in this study. Clinical studies have also indicated that CAD leads to differing levels of cancerous detection depending on prompt validity. For example, Zheng et al. (2004) found that CAD cues might reduce cancer detection in non-cued areas (see also Samulski et al., 2010)[2]. We suggest that this could be because observers become over-reliant on CAD and so fail to detect a visible cancer if the CAD cue fails to prompt it. We call this the over-reliance hypothesis. However, the study by Zheng et al. (2004) again only had a limited number of radiologists and only used conditions in which the prevalence rate of the cancer was high.

In this paper we present five experiments. Experiment 1 investigated whether the LP Effect occurred using mammogram laboratory stimuli, in which participants were asked to search for a pre-defined cancer. In this experiment we also included a letter visual search task (search for

[2] Russell & Kunar (2012) and Drew et al. (2012) reported similar findings suggesting that non-cued targets were missed more often than cued ones. However, both these studies used search tasks where people searched for a target, T among distractor, Ls rather than search for a cancer in a mammogram.

a T among Ls) to act as a baseline condition as we know that an LP Effect occurs with these stimuli (e.g., Rich et al., 2008, Kunar et al., 2010, Russell & Kunar, 2012). The letter search also enabled a direct comparison of the LP Effects between letter visual search and mammogram displays. If the two search tasks showed similar underlying mechanisms this could mean that interventions found to combat the prevalence effect using simpler laboratory stimuli may also be effective in a clinical setting (e.g. Wolfe et al., 2007). Experiment 2 investigated the use of CAD under LP conditions. In this experiment, we directly tested the over-reliance hypothesis determining whether participants' judgements were affected by the presence of CAD. We predict that if participants came to over-rely on CAD, when a target was not highlighted by CAD they would be less likely to find it.

In Experiment 3 participants searched for a range of masses (some cancerous, some benign) in LP search (rather than one specific cancer as in Experiment 1). From a practical perspective, searching for a range of multiple masses better reflects the clinical task where radiologists search for a range of suspicious targets, some of which are cancerous, some of which are benign. However, searching for multiple targets simultaneously has been shown to lead to dual-target costs where response times are slower and accuracy falls (e.g. Godwin, Menneer, Donnelly, & Cave, 2010, Menneer, Barrett, Phillips, Donnelly & Cave, 2007; Menneer, Cave, & Donnelly, 2009; Menneer, Donnelly, Godwin, & Cave, 2010). Mestry et al. (2016) suggested that the reason for this decline in accuracy is that when searching for two possible targets, one of the targets becomes classified as the non-preferred target and is 'shed' (i.e., participants give up searching for the non-preferred target in favour of the preferred one). This shedding is thought to occur as the non-preferred targets were less well represented in Visual Working Memory (VWM) given that VWM has a limited processing capacity (Mestry et al., 2016, see also Cowan, 2001, Garavan, 1998, & McElree & Dosher, 1989, Menneer et al., 2007 &

Oberauer, 2002). On the basis of this theory we predict that requiring participants to hold target templates for a range of masses would exceed the capacity of visual working memory. In turn this would lead to an increased proportion of errors in comparison to when participants searched for one mass. Please note that, Mestry et al. (2016) only reported overall accuracy and did not separate their data into miss errors or false alarms. We propose that both miss errors and false alarms would be affected by the multiple-target cost. If searching for multiple targets was impaired due to weaker representations of each target in VWM, then miss errors would increase due to a failure to match the perceptual input of the current target to the poorer VWM representations. Furthermore, we predict that false alarms would increase as an ambiguous non-mass item that weakly resembled one of the target templates would be more likely to be accepted as a target compared to when the VWM representation of the target type was strong.

Experiments 4 and 5 implemented CAD using these stimuli where participants were asked to search for a range of masses. Based on the work presented by Mestry et al. (2016) we predict that when searching for multiple targets both miss errors and false alarms would be higher than those observed in Experiment 2 (where there was only one target, and therefore a stronger VWM representation). Furthermore, we again predict an over-reliance hypothesis where participants' judgements were swayed by the presence of the CAD cue.

Experiment 1: LP Effect in Mammogram Search

Experiment 1 investigated whether an LP Effect occurred using our laboratory mammogram task.

Method

Participants:

Twenty-four participants (M = 25.3 years, SD = 4.6, 15 female) took part in the experiment. All had normal or corrected-to-normal vision. Ethical approval for all studies was granted by the Humanities and Social Sciences Research Ethics Committee at the University of Warwick. A power analysis using the effect sizes reported in Russell and Kunar (2012, who investigated the effect of cues on LP letter visual search) showed that the minimum number of participants needed to achieve a power of 0.8 for each experiment would be 7. Therefore, we would expect that testing 24 participants per experiment should provide ample power to detect significant effects, if present.

Stimuli and Procedure:

The experiment was programmed using Blitz3D and presented on a PC. For the mammogram condition, images were taken from the selection of 'normal' mammograms (those not containing a cancer) of the Digital Database for Screening Mammography (DDSM) database (Heath et al., 2001, 1998). All images were selected at random. In total, 1100 images were selected – 1000 for the LP condition (2% target prevalence) and 100 for the HP condition (50% target prevalence). Nine-hundred and eighty of these images were selected to act as target absent trials for the LP condition and 40 of these images were selected to act as target absent trials for the HP condition. This led to the LP and HP conditions having different mammogram images. Images were presented in the centre of the display and were approximately 10.7 degrees by 18.6 degrees at a viewing distance of 57 cm in size (although the individual size of

each image varied because they were real mammograms)[3]. For target present trials an image of a cancerous mass was selected from one of the cancer cases on the DDSM and transposed onto the remaining mammogram images using imaging editing software. The cancers could appear on any area of the breast tissue (mimicking conditions in a clinical setting), as long as it was clearly distinguishable once fixated (see Figure 1 for examples). All mammogram displays were created offline.

For the Letter Search condition, the stimuli were rotated Ts and Ls. Each stimulus had a visual angle of 1.7 degrees x 1.7 degrees at a viewing distance of 57 cm and the vertical lines of the Ls were slightly offset from their horizontal line (see Figure 1). All stimuli were white and presented on a grey background. On each trial, there were always 12 stimuli presented (on 'target present' trials – 11 distractors and 1 target; on 'target absent' trials – 12 distractors) and they were presented randomly within an invisible 6 x 6 matrix, subtending an area of 25.2 degrees by 25.2 degrees. The target, if present, was a 'T' and was presented on either 2% (low prevalence) or 50% (high prevalence) of trials. The distractor items were offset L shapes. All stimuli were presented randomly in one of four orientations (0 degrees, 90 degrees, 180 degrees or 270 degrees) with equal probability.

Figure 1 about here

---

[3] Please note that some of the images from the DDSM contained dates and/or artefacts on the background of the image similar to images seen by radiologists in clinical mammography. However, as the dates/artefacts only appeared on the background of the image they did not affect the actual search task.

There were four blocks of trials: two high prevalence conditions – one with the mammogram stimuli and one with the letter search stimuli – and two low prevalence conditions – one with the mammogram stimuli and one with the letter search stimuli. Participants completed all four conditions. In the mammogram conditions, a blank screen appeared for 500ms and was followed by a central fixation point for 500ms. Following this one of the mammogram images was presented and remained on the screen until response (order of the images were randomised across participants). For the visual search condition, a blank screen appeared for 500ms and was followed by a central fixation point for 500ms. Following this the letter search display was presented and remained on the screen until response. In all conditions participants indicated whether the target was 'present' or 'absent' by pressing either the 'm' or the 'z' key respectively. Participants were instructed to respond as quickly but as accurately as possible and were informed of the prevalence rates of the target prior to the condition starting. If no response was made within 10 seconds, the trial 'timed-out' and the next trial started automatically[4]. Following a response or 'time-out', the blank screen was again displayed before the next fixation point and trial.

In line with Fleck and Mitroff (2007), and to correct for motor errors, participants had the option of correcting their response. If a participant recognized that they had made an error, they were able to correct it on the following trial, by pressing the 'Escape' key. This would automatically log in the data file that the participant had noticed their mistake so that motor errors could be calculated. They would then proceed with the current trial as normal, responding with an 'm' or 'z' key if the target was present or absent, respectively. No feedback was given after any response or correction was made.

[4] Note that this aspect of the task differed to that in a clinical setting in which observers do not have a time limit to view each mammogram.

For each of the high prevalence conditions there were 80 trials (40 present and 40 absent). For each of the low prevalence conditions, in which the target was present 2% of the time, there were 1000 trials (20 present and 980 absent). To familiarise themselves with the stimuli, participants were shown examples of the mammogram displays and cancers prior to the experiment. They were also given a short practice block before each experimental block. During this practice block the experimenter ensured that participants were able to recognise the cancer in the Mammogram condition. If any of the participants had difficulties identifying the cancer they would be shown more examples and could undergo the practice condition again until both the participant and experimenter were confident that they were able to identify the cancer. However, all the participants responded correctly in the first practice session and none were asked to repeat it. RTs and error rates were recorded. In the Low Prevalence blocks breaks occurred automatically every 200 trials, after which participants continued with the experiment when they were ready. The presentation order of the blocks was randomised across participants.

Results and Discussion

Consistent with the findings of Fleck and Mitroff (2007) miss errors were reduced after self-correction in all conditions (see Table 1, all ts > 3.4, ps < 0.01). However, as we are primarily interested in cognitive rather than motor response errors throughout the paper we focus our analysis on the self-corrected data. Error rates and mean correct reaction times for all conditions are presented in Tables 2-5.

Table 1 - 5 about here

Miss Errors

Examining the miss errors a 2 x 2 within-participants ANOVA with factors of Search Type (Mammogram vs Letters) and Prevalence (High vs Low) showed there to be a main effect of Search Type, $F(1, 23) = 63.02$, $p < 0.01$, $\eta_p^2 = 0.733$, where more targets were missed in the Letter Visual Search condition compared to the Mammogram search. There was also a main effect of Prevalence, $F(1, 23) = 28.03$, $p < 0.01$, $\eta_p^2 = 0.549$, with more targets missed in LP compared to HP search. However, the Search Type x Prevalence Interaction was not significant, $F(1, 23) = 2.59$, $p = 0.12$.

False Alarms

For the false alarms a 2 x 2 within-participants ANOVA with factors of Search Type (Mammogram vs Letters) and Prevalence (High vs Low) showed there was no main effect of Search Type, $F < 1$, nor of Prevalence, $F < 1$. Neither was the Search Type x Prevalence interaction significant, $F(1, 23) = 1.46$, $p = 0.24$. Overall, the false alarm rate was low, consistent with previous work where people searched for a clearly defined target (e.g. Wolfe et al., 2005, Kunar et al., 2010, Russell & Kunar, 2012).

RTs

In all experiments, Present and Absent RTs were analysed separately to reflect RTs for hits and correct rejections, respectively. Examining the RTs for target present trials a 2 x 2 within-participants ANOVA with factors of Search Type (Mammogram vs Letters) and Prevalence (High vs Low) revealed a main effect of Search Type, $F(1, 23) = 269.47$, $p < 0.01$, $\eta_p^2 = 0.921$. RTs were slower in the Letter Visual Search condition compared to the Mammogram search. There was also a main effect of Prevalence, $F(1, 23) = 32.93$, $p < 0.01$, $\eta_p^2 = 0.589$, where RTs were slower in the LP condition compared to HP. However, the Search Type x Prevalence interaction was not significant, $F < 1$.

For target absent trials a 2 x 2 within-participants ANOVA with factors of Search Type (Mammogram vs Letters) and Prevalence (High vs Low) revealed a main effect of Search Type, $F(1, 23) = 156.47$, $p < 0.01$, $\eta_p^2 = 0.872$, in which RTs were slower in the letter visual search condition compared to the mammogram search. There was also a main effect of Prevalence, $F(1, 23) = 11.38$, $p < 0.01$, $\eta_p^2 = 0.331$, in which RTs were faster in the LP condition compared to HP. The Search Type x Prevalence interaction was borderline significant, $F(1, 23) = 3.44$, $p = 0.08$, $\eta_p^2 = 0.130$, in which the difference in RTs across prevalence rate was greater for the Letter Visual Search stimuli than it was for the Mammogram displays.

Summary of Experiment 1

Miss errors were reduced overall when participants were given the option to self-correct. However, even when self-corrected people missed more targets when the prevalence was low compared with when it was high. This LP Effect occurred for both Letter Visual Search and Mammogram stimuli with no reliable difference between the LP Effect across displays, denoted by the lack of a significant Search Type x Prevalence interaction. These data extend the results of other studies showing the LP effect in different display types (e.g. Van Wert et al., 2009, Mitroff & Biggs, 2014, Wolfe et al., 2005, Evans et al., 2013) and importantly show that an LP Effect occurs when searching for cancers in these mammogram images. One reason for the high miss errors could be that participants were faster at responding to target absent trials at LP than at HP. These data replicate and extend previous findings in the literature suggesting that under LP search people were quitting their search too soon. In contrast to the miss error data, people were making very few false alarms.

The results also showed that participants were faster at responding and missed fewer targets in Mammogram search than in Letter Visual Search. Please note that we do not believe the ease of which participants responded to cancers (in terms of reduced RTs and error rates) reflect those in a clinical setting where search for a cancer would be more difficult. It may be that participants learned to efficiently identify the specific cancer used in the mammogram search in this experiment and so became better at searching for it in comparison to search for a T among an L which is known to be an inefficient search task (Treisman & Gelade, 1980, Kunar & Humphreys, 2006, Wolfe et al., 1989). However, the same would not occur in a clinical setting where readers have to search for a range of potential masses rather than one specific cancer exemplar. We investigated this further in Experiments 3 – 5. Nevertheless, despite being able to respond better to the cancer compared to a letter target in this experiment, overall, participants still showed an LP Effect, even when searching mammogram images.

Experiment 2: CAD in Mammogram Search

Experiment 1 showed that an LP Effect occurred using mammogram stimuli. Experiment 2 investigated how cancer detection might be affected by the use of CAD prompts. The practical benefits of CAD prompts in mammography are inconclusive (e.g. Bennett et al., 2006). Although many studies report the benefits of CAD, very few focus on the potential negative consequences (e.g., the possibility of observers becoming over-reliant on the CAD prompt and missing targets that are not prompted). This over-reliance hypothesis was tested in Experiment 2.

Method

Participants:

Twenty-four participants (M = 21.7 years, SD = 3.1, 11 female) took part in the experiment. All had normal or corrected-to-normal vision.

Stimuli and Procedure:

The stimuli were the same as in the mammogram displays from Experiment 1, except that on some trials the outline of a salient red box (visual angle of box size 9 degrees x 8.5 degrees, line thickness 0.3 degrees, at a viewing distance of 57 cm), which acted as the CAD prompt would appear on the image. There were three conditions where the red box would appear: (i) Present Correct CAD trials – with the target located within the bounds of the box; (ii) Present Incorrect CAD trials – with the target present but outside the box presented at a random location within the breast tissue; or (iii) Absent Incorrect CAD trials – where the target was absent and the box appeared in a random position within the breast tissue.

There were two experimental conditions: high prevalence and low prevalence. The HP condition was used as a baseline to analyse the specific effects of using a red box to cue the possible target location. In this condition, there were 200 trials with 100 (50%) being target-present trials. Of these target-present trials, 60 included the red box containing the target (Present Correct CAD), 20 had the red box with the target outside the box[5] (Present Incorrect CAD) and a further 20 had no box at all (Present No CAD). The remaining 100 trials (target-absent) included 24 with the red box randomly positioned in the display (Absent Incorrect

[5] In these trials the target and box could appear at any location as long as the target fell outside the CAD prompt and that both the target and the CAD prompt appeared on the breast image.

CAD) and 76 trials where no box was present (Absent No CAD)[6]. Example images can be seen in Figure 2. The HP condition demonstrated how effective the red box cue was at capturing attention when the target was frequent.

Figure 2 about here

The LP condition comprised a total of 5000 trials with 100 (2%) target-present trials – 60 included the red box containing the target (Present Correct CAD), 20 had the red box with the target outside the box (Present Incorrect CAD) and a further 20 had no box at all (Present No CAD). These trial numbers were the same as those in the HP condition. The remaining 4900 trials were target absent trials of which 1170 of them contained a red box (Absent Incorrect CAD) and 3730 of them had no box present (Absent No CAD). As the prevalence of the target appearing in the LP and HP trials needed to be kept at 2% and 50% respectively, the absolute numbers of the different trial types could not be equated, however, the percentage of the different trial types used in the LP condition were the same as in the HP condition (with 24% Absent Incorrect CAD and 76% Absent NO CAD trials at both high and low prevalence). The CAD cue, when present, was presented at the same time as the mammogram image. Please note, that target present trials were more likely to contain a CAD cue than target absent trials as in the field the CAD algorithms used would be more likely to display a prompt when a cancer appears than when it is absent. Participants were aware that the target, if present, was

[6] It is difficult to identify the proportion of real world cases that show a CAD prompt as it varies depending on the exact algorithm implemented by the CAD technology. Studies have reported different ranges of CAD sensitivity that vary from 57% (Soo et al., 2005) to 85% (Obenauer et al., 2006) depending on the type of cancer and suspiciousness of the lesion. However, it is generally accepted that CAD is able to prompt the majority of cancers without creating too many false prompts. Therefore, we have chosen these proportions of trials to reflect this outcome. Furthermore, by using these proportions the data can be directly compared to other LP experiments in the field using explicit visual cues (see Russell & Kunar, 2012, who used these validity rates).

likely to be cued by the CAD prompt, however, they were also told that on some trials there would be no CAD prompts on present trials, or the target could appear outside the CAD cue.

The total number of LP trials was split equally into 5 sessions of 1000 trials. Given the volume of mammograms needed to create target absent trials in the LP conditions these displays were repeated across sessions[7], however target present trials were never repeated across the experiment. Participants completed the HP condition during one of these sessions, the order of which – including the position of the HP condition within one of these sessions – was randomised across participants. There was at least 30 minutes between each session and they could be spread over a number of days. Participants were given a short practice block before each experimental session. They were asked to respond as quickly but as accurately as possible and also told that the red box was there to assist them and would most likely highlight the target, if it was present.

Results and Discussion

Error rates and mean correct RTs for all conditions are presented in Tables 2 -5.

Miss Errors

Examining the miss errors a 2 x 3 within-participants ANOVA with factors of Prevalence (High vs Low) and CAD (Correct, Incorrect and No CAD) showed there to be a main effect of Prevalence, $F(1, 23) = 23.13$, $p < 0.01$, $\eta_p^2 = 0.501$, where more targets were missed when there was a LP compared to a HP. There was also a main effect of CAD, $F(2, 46) = 21.89$, $p < 0.01$, $\eta_p^2 = 0.488$. Planned t-tests revealed that fewer targets were missed in the Correct CAD

[7] Repeating the target absent trials across the experiment should have little effect on the results as previous work has shown no evidence of learning displays when the target is not there (Kunar & Wolfe, 2011)

condition compared to the Incorrect or No CAD conditions ($t(23) = 4.32$, $p < 0.01$, $d = 0.733$ and $t(23) = 4.06$, $p < 0.01$, $d = 1.036$ respectively). More targets were missed in the No CAD condition compared to the Incorrect CAD condition ($t(23) = 4.99$, $p < 0.01$, $d = 0.419$). The Prevalence x CAD Interaction was also significant, $F(2, 46) = 26.15$, $p < 0.01$, $\eta_p^2 = 0.532$. Planned t-tests revealed that the LP Effect was largest in the No CAD condition compared to the Incorrect CAD and Correct CAD conditions (($t(23) = 4.86$, $p < 0.01$, $d = 0.640$ and $t(23) = 5.54$, $p < 0.01$, $d = 1.109$ respectively). The LP Effect was also larger in the Incorrect CAD condition than in the Correct CAD condition ($t(23) = 3.84$, $p < 0.01$, $d = 0.589$).

False Alarms

Examining the false alarm rates a 2 x 2 within-participants ANOVA with factors of Prevalence (High vs Low) and CAD (Incorrect and No CAD) showed there to be no main effect of Prevalence, $F < 1$. Neither was there a main effect of CAD, $F(1, 23) = 1.99$, $p = 0.17$. The Prevalence x CAD interaction was not significant, $F < 1$. Similar to Experiment 1, the false alarm rate was low, consistent with previous work where people searched for a clearly defined target (e.g. Wolfe et al., 2005, Kunar et al., 2010, Russell & Kunar, 2012).

RTs

A 2 x 3 within-participants ANOVA on mean correct RTs for target present trials with factors of Prevalence (High vs Low) and CAD (Correct, Incorrect and No CAD) showed there to be a main effect of Prevalence, $F(1, 23) = 36.49$, $p < 0.01$, $\eta_p^2 = 0.613$. RTs were faster in HP trials compared to LP. There was also a main effect of CAD, $F(2, 46) = 24.33$, $p < 0.01$, $\eta_p^2 = 0.514$, in which RTs were faster in the Correct CAD condition compared to the Incorrect CAD and No CAD conditions ($t(23) = 4.81$, $p < 0.01$, $d = 0.354$ and $t(23) = 6.46$, $p < 0.01$, $d = 0.766$, respectively). RTs were also faster in the Incorrect CAD compared to the No CAD condition,

t(23) = 3.30, $p < 0.01$, d =0.416. The Prevalence x CAD Interaction was not significant, F(2, 46) = 2.37, p = 0.11.

For target absent trials a 2 x 2 within-participants ANOVA with factors of Prevalence (High vs Low) and CAD (Incorrect and No CAD) showed there to be a main effect of Prevalence, $F(1, 23) = 6.39$, $p < 0.05$, $\eta_p^2 = 0.217$, where RTs were faster in the LP condition compared to HP. The main effect of CAD was significant, $F(1, 23) = 50.01$, $p < 0.01$, $\eta_p^2 = 0.685$, in which RTs were fastest in the No CAD condition compared to the Incorrect CAD. There was also a significant Prevalence x CAD interaction, $F(1, 23) = 9.23$, $p < 0.01$, $\eta_p^2 = 0.286$, in which the difference in RTs between the HP and LP conditions was biggest in the Incorrect CAD condition than in the No CAD condition.

Summary of Experiment 2

Similar to Experiment 1, false alarm rates were low and showed no difference across prevalence rate or CAD presentation. However, the same could not be said about miss errors where people missed more cancers at low compared to high prevalence. Importantly, there was also an effect of CAD on target detection. When the CAD prompt was used to highlight the target miss error rates were reduced. However, when there was no CAD prompt or the CAD prompt highlighted a non-target area miss errors increased. This increase was particularly pronounced under LP conditions where approximately 25% of targets were missed when there was no CAD prompt. The data showed that the use of a CAD cue was effective under LP conditions, but only when it correctly cued the target location. This is in accordance with the over-reliance hypothesis, where participants began to over-rely on the CAD prompt to correctly identify the target, leading to impaired search when the CAD cue was inaccurate.

Examining the RTs, participants responded faster under LP conditions compared to HP conditions when the target was absent. This was similar to the results of Experiment 1 and consistent with the theory that participants were quitting their search too soon. There was also an effect of CAD on RTs. Unsurprisingly, if the target was cued by the CAD then RTs were faster. In contrast, an incorrect CAD cue on target absent trials slowed RTs, especially when the prevalence of a target was high.

Overall, the findings have implications for CAD use in mammogram search. Correct CAD cues on target present trials resulted in a reduced number of miss errors and clearly helped observers detect the target. However, should the cancer fall outside the CAD cue or if there was no CAD presented on target present trials miss errors were much higher. In fact, miss errors in the No CAD condition here were significantly higher than miss errors in Experiment 1 which did not use CAD at all, $t(46) = 2.67$, $p = 0.01$, $d = 0.769$. Under these conditions, introducing a CAD aid to search affected search negatively in comparison to when the search aid was not implemented at all.

Experiment 3: Search for a Range of Masses

Experiments 1 and 2 showed an LP effect occurred in mammogram displays where more miss errors were found in LP conditions compared to HP. However, false alarms were negligible across experiments. This differs from clinical settings where the proportion of false alarms is higher (9.3% in the USA, NIH, 2015). One of the reasons for the low false alarm rates could be that in Experiments 1 and 2 participants were searching for a specific cancer. This may have made the task easier as people knew on each trial what target to look for. Experiments 3 - 5 investigated what happens to miss rates and false alarm rates if we add target uncertainty from

trial to trial by asking participants to look for one of a range of potential masses, some of them cancerous and some of them benign. In Experiment 3 participants searched for the mass without CAD to see how well observers detected cancers without the use of an automated prompt, whereas in Experiments 4 and 5 we investigated the effect of CAD on cancer detection when the identity of the mass on each trial was uncertain. In screening practice it is important to recall women for further tests if they have cancerous masses so that they can be treated, but not if they have benign masses. Benign masses are clinically unimportant, so recalling those women is undesirable. In the clinical task one of the key skills is distinguishing between benign and malignant masses. According to Mestry et al., (2016) adding multiple target exemplars will lead to a weakening in the representation of target representations in Visual Working Memory. We propose that this will result in a large proportion of miss errors and false alarms observed both at HP and at LP for cancerous and benign masses. Furthermore, Experiment 3 also tested the hypothesis proposed by the Multiple Decision Model of Visual Search that under LP conditions false alarm rates should decrease (Wolfe & Van Wert, 2010).

Method

Participants:

Twenty-four participants (M = 21.4 years, SD = 2.5, 15 female) took part in the experiment. All had normal or corrected-to-normal vision.

Stimuli and Procedure:

The stimuli were similar to those used in the mammogram displays from Experiment 1, except that participants were asked to look for any one of four different cancerous masses or any one of four different benign masses on target present trials. To create these displays, 80 'normal' mammograms (those not containing a cancer) were randomly selected from the DDSM (40 for HP trials and 40 for LP trials). These images were then digitally edited to include either a cancerous mass or a benign mass. The masses were also taken from the DDSM and included four cancerous masses (selected at random from the 'Cancer' cases) and four benign masses (selected at random from the 'Benign' cases). Each mass was then transposed onto ten of the 'normal' mammogram images to create 40 cancerous mass mammograms (20 to be used in the HP condition and 20 to be used in the LP condition) and 40 benign mass mammograms (20 to be used in the HP condition and 20 to be used in the LP condition). Similar to Experiment 1, the masses could appear on any area of the breast tissue, as long as it was clearly distinguishable once fixated. Example images can be seen in Figure 3. All mammogram displays were created offline. Target absent trials were created in a similar manner to Experiment 1 by randomly selecting from the DDSM 40 'normal' mammograms for the HP condition and 960 'normal' mammograms for the LP condition.

Figure 3 about here

For the High Prevalence conditions, where a mass was present 50% of the time, there were 80 trials. Half of these trials were target present and contained either a cancerous or benign mass (20 trials with a cancerous mass and 20 trials with a benign mass). The other half of the displays were target absent displays. For the Low Prevalence conditions there were 1000 trials (20 trials

with a cancerous mass and 20 trials with a benign mass and 960 absent). This meant there was a mass present 4% of the time of which 2% of trials contained a cancerous mass and 2% contained a benign mass. To familiarise themselves with the stimuli, participants were given a training session where they were shown examples of the mammogram displays, benign masses and cancers prior to the experiment. During this training session, participants were shown example displays of mammogram images and asked to locate and identify each mass. Once the experimenter was confident the participants could identify the mass, participants each took part in a training test before the experiment proper. The training test examined participants' ability to recognise and classify a mass as either benign or cancerous and included 24 examples of mammogram displays (12 containing a cancer and 12 containing a benign mass). Participants were asked to press the letter 'b' if they thought the mass was benign and a 'c' if they though the mass was cancerous. Participants only continued onto the experiment once they had successfully completed the training test by being able to correctly identify all masses.

The procedure was similar to the procedure used in the mammogram condition in Experiment 1. However, participants were asked to indicate whether a mass (either cancer or benign) was present or absent by pressing either the 'm' or the 'z' key respectively. If they pressed the present key they were then given a follow up question asking them to identify the mass as either cancerous or benign by pressing either the 'c' or the 'b' key respectively. The presentation order of blocks was randomised across participants.

Results and Discussion

Error rates and mean correct reaction times for all conditions are presented in Tables 2 - 5.

Miss Errors

Examining the miss errors a 2 x 2 within-participants ANOVA with factors of Prevalence (High vs Low) and Mass (Benign vs Cancer) revealed a main effect of Prevalence, $F(1, 23) = 28.22$, $p < 0.01$, $\eta_p^2 = 0.551$, in which more masses were missed at LP than at HP. There was a main effect of Mass, $F(1, 23) = 22.16$, $p < 0.01$, $\eta_p^2 = 0.491$, in which more cancers were missed compared to benign masses. This might be because the benign masses we used were less spiculated than their cancerous counterparts. As the benign masses had a much smoother texture this might lead them to be segmented more easily from the background leading them to a greater detection rate (e.g. Julesz, 1981). The Prevalence x Mass interaction was also significant, $F(1, 23) = 8.46$, $p < 0.01$, $\eta_p^2 = 0.269$, in which the LP Effect was greater for cancerous masses than for benign masses.

False Alarms

Examining the false alarm rate, a two-tailed paired t-test showed that there was no significant difference in false alarms between HP and LP conditions, $t < 1$. However, false alarm rates in both the HP and LP conditions were greater here than those witnessed in Experiment 1 ($t(46) = 3.49$, $p < 0.01$, $d = 0.696$ and $t(23) = 3.04$, $p < 0.01$, $d = 0.880$ for HP and LP conditions, respectively), where participants had only one target to search for.

RTs

A 2 x 2 within-participants ANOVA on mean correct RTs for target present trials with factors of Prevalence (High vs Low) and Mass (Cancer vs Benign) revealed a marginal main effect of Prevalence, $F(1, 23) = 3.17$, $p = 0.09$, $\eta_p^2 = 0.121$, in which there was a trend for RTs to be faster in HP conditions compared to LP. There was also a main effect of Mass, $F(1, 23) = 10.78$, $p < 0.01$, $\eta_p^2 = 0.319$, in which RTs were faster for benign masses compared to cancers. The

Prevalence x CAD Interaction was not significant, $F(1, 23) = 1.31$, $p = 0.26$. Examining the absent trials, there was no difference in RTs across HP and LP conditions, $t < 1$[8].

Accuracy of Mass Identification

When asked to identify each detected mass participants performed accurately 78.5% of the time. Breaking this down into accuracy for each mass type the data showed that participants accurately identified 82% of all cancers and 74% of all benign masses. There was no difference in identification accuracy rates across mass type, $t(23) = 1.7$, $p = 0.09$). As correct mass identification did not vary with prevalence (78.6% vs 77.7% for HP and LP, respectively, $t < 1$) we do not analyse this factor further here or in subsequent experiments.

Summary of Experiment 3

The results showed an overall LP Effect, where participants missed more masses at LP compared to HP. This could have occurred as participants were responding too quickly in LP trials. However, although there was a numerical difference in RTs for target absent trials between HP and LP, unlike previous experiments this was not significant here. Of interest is the result that having people search for a range of potential masses increased the proportion of false alarms observed. This agrees with what we would expect to find in a clinical setting (e.g., National Institute of Health, 2015). On the other hand, our data showed little support for the Multiple Decision Model (Wolfe & Van Wert, 2010) as there was no reliable difference in false alarm rates across prevalence. However, our data can be explained by the theory presented by Mestry et al. (2016) which predicted that VWM representations would be weaker with multiple target exemplars (see also Menneer et al., 2007). In this case having a weaker representation

[8] As, by definition, there was no mass in the target absent trials we can only analyse the data across Prevalence rates as the factor of Mass does not exist in these data.

would mean that there would be a greater chance of an ambiguous item being accepted as a target compared to when the representation of the target was strong. This would occur both at HP and at LP, which explains why there was little difference in false alarms across prevalence rates.

Experiment 4: CAD on High and Low Prevalence Search for a Range of Masses

Method

Experiment 3 showed that an LP Effect occurred under conditions where participants were searching for a range of possible masses. Using these stimuli, false alarms also increased. Experiment 4 examined the effect of CAD on miss errors and false alarms when searching for a range of targets. Here we again tested the over-reliance hypothesis, which predicted an increase in both miss errors and false alarms when the CAD cue was incorrect.

Participants:

Twenty-four participants (M = 20.8 years, SD = 1.6, 15 female) took part in the experiment. All had normal or corrected-to-normal vision.

Stimuli and Procedure:

The stimuli and procedure were similar to those used in the mammogram displays from Experiment 3, except that some critical trials contained the red box CAD cue. In the HP condition (50% mass prevalence) there were 400 trials in total. These contained 200 trials where a mass was present. Of these target-present trials, 60 included the red box containing a

cancer and 60 included the red box containing a benign mass (Cancer Present - Correct CAD and Benign Present - Correct CAD, respectively), 20 had the red box with a cancer outside the box and 20 had the red box with a benign mass outside the box (Cancer Present - Incorrect CAD and Benign Present - Incorrect CAD, respectively), 20 contained a cancer and had no box at all and 20 contained a benign mass and had no box at all (Cancer Present - No CAD and Benign Present - No CAD, respectively). The remaining 200 trials were target absent trials and included 50 with the red box randomly positioned in the display (Absent Incorrect CAD) and 150 trials where no box was present (Absent No CAD).

In the LP condition there were 5000 trials in total (4% mass prevalence of which 2% were cancerous). These contained 200 trials where a mass was present. Of these target-present trials, similar to the HP condition, 60 included the red box containing a cancer and 60 included the red box containing a benign mass (Cancer Present - Correct CAD and Benign Present - Correct CAD, respectively), 20 had the red box with a cancer outside the box and 20 had the red box with a benign mass outside the box (Cancer Present - Incorrect CAD and Benign Present - Incorrect CAD, respectively), 20 contained a cancer and had no box at all and 20 contained a benign mass and had no box at all (Cancer Present - No CAD and Benign Present - No CAD, respectively). The remaining 4800 trials were target absent trials and included 1200 with the red box randomly positioned in the display (Absent Incorrect CAD) and 3600 trials where no box was present (Absent No CAD). The presentation order of blocks was randomised across participants.

Results and Discussion

Error rates and mean correct reaction times for all conditions are presented in Tables 2 - 5.

Miss Errors

Examining the miss errors a 2 x 2 x 3 within-participants ANOVA with factors of Prevalence (High vs Low), Mass (Benign vs Cancer) and CAD (Correct, Incorrect or No CAD) revealed a main effect of Mass, $F(1, 23) = 17.15$, $p < 0.01$, $\eta_p^2 = 0.427$, in which more cancers were missed compared to benign masses. There was a main effect of CAD, $F(2, 46) = 11.99$, $p < 0.01$, $\eta_p^2 = 0.343$, in which more masses were missed in the Incorrect CAD condition than in the No CAD or Correct CAD conditions ($t(23) = 2.64$, $p < 0.01$, $d = 0.654$ and $t(23) = 3.90$, $p < 0.01$, $d = 0.791$, respectively). Furthermore, more masses were missed in the No CAD condition than in the Correct CAD condition, $t(23) = 4.95$, $p < 0.01$, $d = 0.385$. There was no main effect of Prevalence, $F < 1$. The Mass x CAD interaction was significant, $F(2, 46) = 8.64$, $p < 0.01$, $\eta_p^2 = 0.273$, in which more cancerous masses than benign masses were missed in the Incorrect CAD and No CAD conditions ($t(23) = 2.21$, $p < 0.05$, $d = 0.162$ and $t(23) = 4.12$, $p < 0.01$, $d = 0.817$) but not in the Correct CAD condition, $t < 1$. None of the other interactions were significant (all $Fs < 1.2$, $ps > 0.3$).

False Alarms

For the false alarms a 2 x 2 within-participants ANOVA with factors of Prevalence (High vs Low) and CAD (Incorrect or No CAD) showed there to be a main effect of CAD, $F(1, 23) = 11.02$, $p < 0.01$, $\eta_p^2 = 0.324$, where there were more false alarms in the Incorrect CAD condition compared to when there was no CAD. Contrary to what was predicted by the Multiple Decision Model, there was no main effect of Prevalence, $F(1, 23) = 1.22$, $p = 0.28$. Neither was the Prevalence x CAD interaction significant, $F(1, 23) = 2.70$, $p = 0.11$.

RTs

For present trials, a 2 x 2 x 3 within-participants ANOVA on mean correct RTs, with factors of Prevalence (High vs Low), Mass (Benign vs Cancer) and CAD (Correct, Incorrect or No CAD) showed there was a main effect of Mass, $F(1, 23) = 13.49$, $p < 0.01$, $\eta_p^2 = 0.370$, in which RTs were faster for benign masses compared to cancers. There was a main effect of CAD, however, there was no main effect of Prevalence, $F < 1$. There was a significant Prevalence x CAD interaction, $F(2, 46) = 3.20$, $p = 0.05$, $\eta_p^2 = 0.122$ and the Mass x Prevalence interaction was marginally significant, $F(1, 23) = 3.90$, $p = 0.06$, $\eta_p^2 = 0.145$. None of the other interactions were significant (all $Fs < 1$).

For absent trials, a 2 x 2 within-participants ANOVA on mean correct RTs, with factors of Prevalence (High vs Low) and CAD (Correct, Incorrect or No CAD) showed that there was no main effect of Prevalence or CAD (all $Fs < 1$). Neither was the Prevalence x CAD interaction significant, $F(1, 23) = 1.26$, $p = 0.27$.

Accuracy of Mass Identification

For correct target present trials, participants accurately identified the mass as either cancerous or benign 73.3% of the time. Table 6 shows the percentage of masses correctly identified across the different CAD conditions. A 2 x 3 ANOVA showed that there was no main effect of Mass ($F < 1$). However there was a main effect of CAD, $F(2, 46) = 5.83$, $p < 0.01$, $\eta_p^2 = 0.202$. Participants correctly identified more masses in the Correct CAD and No CAD conditions than in the Incorrect CAD condition ($t(23) = 2.31$, $p < 0.05$, $d = 0.494$ and $t(23) = 2.64$, $p < 0.05$, $d = 0.585$, respectively). There was no difference in mass identification between No CAD and Correct CAD trials, $t < 1$. There was also a significant Mass x CAD interaction $F(2, 46) = 3.42$, $p < 0.05$, $\eta_p^2 = 0.130$, which reflects the fact that for benign masses participants were better at identifying the mass in the Correct CAD condition compared to the Incorrect and No CAD

conditions ($t(23) = 2.03$, $p = 0.05$, $d = 0.375$ and $t(23) = 3.30$, $p < 0.01$, $d = 0.198$, respectively). However, for cancers observers were worse at identifying masses in the Incorrect CAD condition compared to the No CAD and Correct CAD conditions ($t(23) = 3.35$, $p < 0.01$, $d = 0.499$ and $t(23) = 2.49$, $p < 0.05$, $d = 0.434$, respectively) while there was no difference between Correct CAD and Incorrect CAD trials, $t < 1$. Clearly, CAD not only affects detection of the mass, but also its identification. We discuss this further in the General Discussion.

Table 6 about here

Summary of Experiment 4

Similar to Experiment 2, there was an overall effect of CAD on miss errors. When the target was highlighted by the CAD prompt participants missed fewer targets than when it was not highlighted by the CAD prompt or when no CAD prompt appeared. False alarms also showed an effect of CAD presence. Participants made more false alarms on target absent trials when a CAD cue was presented compared to when it was not. The effect of having an incorrect CAD present increased both miss errors and false alarms, consistent with the over-reliance hypothesis.

Surprisingly, there was no overall effect of target prevalence on miss errors or false alarms in this experiment. Examining Table 2, the lack of LP Effect seems to be driven due to higher miss errors than expected in the HP condition rather than a lowering of miss errors in the LP condition. There are two potential reasons why miss errors might be higher under HP conditions in this experiment. First it could be that adding *both* (i) target uncertainty in terms of having participants search for a range of masses (both cancerous and benign), which would

weaken target representations in VWM (Mestry et al., 2016) and (ii) CAD cues (some of which are misleading) meant that the search and decision process for what is a cancerous mass became difficult regardless of prevalence. Therefore, in this case participants may have chosen to over-rely on the CAD cue to a greater extent than in situations where the search task was easier. This over-reliance on the CAD cue increased errors at both high and low prevalence.

Second, it could be a result of participants completing many more LP trials than HP trials (e.g., the LP condition ran over five search sessions rather than the one search session used for HP trials). Although this is the typical design for LP conditions, with the increased difficulty of the task brought about from the target uncertainty and use of CAD it may be that people are using the strategies implemented from LP search (e.g. faster RTs, a change in response bias) in the HP condition as well, leading to increased miss errors. This is made possible as the HP and LP conditions were run as a within participant design meaning that LP strategies could be carried over to HP search. To investigate this Experiment 5 had participants complete the HP condition on its own. If an LP Effect was observed in these circumstances, where miss errors in this HP condition were fewer than those reported in the LP condition of Experiment 4, this would suggest that strategies, obtained over lengthy exposure to LP search affected HP response. However, if there was still no LP effect when the HP condition was run in isolation this would suggest that the task of searching for a range of targets led to an over-reliance on CAD cues regardless of prevalence rate.

Experiment 5: CAD on High Prevalence Search for a Range of Masses

In Experiment 5 participants searched for a range of masses in a condition where the target prevalence was 50% (HP). As this condition was run in isolation of the LP condition, no LP strategies could be carried over to affect the results.

Method

Participants:

Twenty-four participants (M = 21.5 years, SD = 2.2, 13 female) took part in the experiment. All had normal or corrected-to-normal vision.

Stimuli and Procedure:

The stimuli and procedure were the same as the HP condition in Experiment 4, except that this block was run in isolation without any LP trials.

Results and Discussion

Error rates and mean correct reaction times for all conditions are presented in Tables 2 - 5. To compare behavioural responses across prevalence rates a between experiment comparison was conducted using the LP condition of Experiment 4 and the HP condition tested here.

Miss Errors

Examining the miss errors a 2 x 2 x 3 mixed ANOVA with within experiment factors of Mass (Benign vs Cancer) and CAD (Correct, Incorrect or No CAD) and between experiment factors of Prevalence (HP condition from Experiment 5 and LP condition from Experiment 4) revealed a main effect of Mass, $F(1, 46) = 39.85$, $p < 0.01$, $\eta_p^2 = 0.464$. Similar to Experiment 4, miss

errors for cancerous targets were greater than those for benign masses. There was a main effect of CAD, $F(2, 92) = 18.55$, $p < 0.01$, $\eta_p^2 = 0.287$. Miss errors were highest in the Incorrect compared to Correct and No CAD conditions ($t(47) = 4.91$, $p < 0.01$, $d = 0.949$ and $t(47) = 2.89$, $p < 0.01$, $d = 0.484$, respectively). There were also more miss errors in the No CAD compared to Correct CAD condition, $t(47) = 7.47$, $p < 0.01$, $d = 0.962$. There was no main effect of Prevalence, $F < 1$. The Mass x CAD interaction was significant, $F(2, 92) = 26.39$, $p < 0.01$, $\eta_p^2 = 0.365$. More cancers were missed than benign masses in the Incorrect and No CAD conditions ($t(47) = 3.46$, $p < 0.01$, $d = 0.168$ and $t(47) = 6.89$, $p < 0.01$, $d = 0.938$, respectively). However, there was a trend for more benign masses to be missed than cancers in the Correct CAD condition, $t(47) = 1.72$, $p = 0.09$, $d = 0.186$. None of the other interactions were significant (all $Fs < 1.1$, $ps > 0.3$).

False Alarms

For the false alarms a 2 x 2 mixed ANOVA with within experiment factors of CAD (Incorrect or No CAD) and between experiment factors of Prevalence (High vs Low) showed there to be a main effect of CAD, $F(1, 46) = 13.98$, $p < 0.01$, $\eta_p^2 = 0.233$, in which there were more false alarms in the CAD condition compared to when there was no CAD. There was also a main effect of Prevalence, $F(1, 46) = 15.05$, $p < 0.01$, $\eta_p^2 = 0.247$. However, in contrast to what was predicted by the Multiple Decision Model there were more false alarms in the LP condition compared to the HP condition. The Prevalence x CAD interaction was also significant, $F(1, 46) = 4.49$, $p < 0.05$, $\eta_p^2 = 0.089$. There were more false alarms at LP than HP in the CAD condition compared to the no CAD condition.

RTs

For present trials, a 2 x 2 x 3 mixed ANOVA on mean correct RTs, with within experiment factors of Mass (Benign vs Cancer) and CAD (Correct, Incorrect or No CAD) and between experiment factors of Prevalence (High vs Low), showed there was a main effect of Mass, $F(1, 46) = 16.28$, $p < 0.01$, $\eta_p^2 = 0.261$, in which RTs were faster for benign masses compared to cancers. There was a main effect of CAD, $F(2, 92) = 31.10$, $p < 0.01$, $\eta_p^2 = 0.403$ and a marginal effect of Prevalence, $F(1, 46) = 3.52$, $p = 0.07$, $\eta_p^2 = 0.071$, in which there was a trend for RTs to be faster in the LP condition compared to the HP. The Mass x CAD x Prevalence interaction was significant, $F(2, 92) = 5.03$, $p < 0.01$, $\eta_p^2 = 0.098$. Breaking this down we see that there was a significant Mass x Prevalence interaction, $F(1, 46) = 10.82$, $p < 0.01$, $\eta_p^2 = 0.190$, and a significant Mass x CAD interaction, $F(2, 92) = 4.09$, $p < 0.05$, $\eta_p^2 = 0.082$. There was also a significant CAD x Prevalence interaction, $F(2, 92) = 4.06$, $p < 0.05$, $\eta_p^2 = 0.081$.

For absent trials, a 2 x 2 within-participants ANOVA on mean correct RTs, with within experiment factors of CAD (Correct, Incorrect or No CAD) and between experiment factors of Prevalence (High vs Low) showed that there was a main effect of Prevalence, $F(1, 46) = 8.37$, $p < 0.01$, $\eta_p^2 = 0.154$, in which RTs were faster for LP trials than HP. There was no main effect of CAD, $F(1, 46) = 2.06$, $p = 0.16$. Neither was the Prevalence x CAD interaction significant, $F(1, 46) = 1.76$, $p = 0.19$.

Accuracy of Mass Identification

For correct target present trials, participants accurately identified the mass as either cancerous or benign 78.1% of the time. Table 6 shows the percentage of masses correctly identified across the different CAD conditions. A 2 x 3 ANOVA showed that there was no main effect of Mass, $F < 1$. However, there was a main effect of CAD, $F(2, 94) = 7.34$, $p < 0.01$, $\eta_p^2 = 0.135$, in which participants were worse at identifying the mass in the Incorrect and No CAD conditions

compared to the Correct CAD condition ($t(47) = 3.06$, $p < 0.01$, $d = 0.473$ and $t(47) = 3.79$, $p < 0.01$, $d = 0.188$, respectively). Participants were also worse at identifying the mass in the Incorrect CAD versus the No CAD condition, $t(47) = 2.17$, $p < 0.05$, $d = 0.334$. The Mass x CAD interaction was marginally significant $F(2, 94) = 2.52$, $p = 0.09$, $\eta_p^2 = 0.051$, in which the difference between benign and cancerous identification accuracy was greater in the No CAD than in the Incorrect CAD condition, $t(47) = 2.09$, $p < 0.05$, $d = 0.19$, but not between the Correct CAD and the Incorrect CAD or No CAD conditions ($t < 1$ and $t(47) = 1.66$, $p = 0.1$, respectively). The data confirm those of Experiment 4 showing that CAD validity affects target identification and a person's decision making ability to classify a mass as cancerous or not.

Summary of Experiment 5

Experiment 5 was used to investigate whether the high miss errors produced in Experiment 4 were a result of participants adopting LP search strategies in HP search when the target was difficult to identify (due to there being multiple masses to search for and CAD presence). Here the HP condition was run in isolation so that the results could not be affected by transferable strategies from LP search. Despite this the results showed that miss errors were again higher than expected in HP search and there was no difference to those observed at LP. Instead the results suggest that, even when this HP task is completed in isolation, participants made a higher proportion of miss errors than those observed in tasks where there is only one target (Experiment 2) or no CAD use (Experiment 3). We consider this further in the General Discussion.

General Discussion

The current paper investigated whether the LP Effect occurred using a mammogram search task in which participants searched for a cancerous target (Experiments 1-2) or a range of cancerous and benign targets (Experiments 3-5). It also investigated the use of CAD prompting on cancer detection (Experiments 2, 4 and 5). Experiment 1 showed that an LP Effect occurred with the use of mammogram images. Participants missed more cancers when they occurred at low prevalence compared to when they appeared at high prevalence. Furthermore, there was no difference in terms of the prevalence effect between mammogram search and letter search. In both search tasks participants missed more targets at LP than they did at HP. Comparison of the data across search types are useful as, if the same underlying factor causes the prevalence effect to occur in both conditions, it is feasible that interventions previously found to be effective using simpler lab-based stimuli will also apply to mammogram search (e.g. Wolfe et al., 2007).

Experiment 2 again showed that an LP Effect occurred in mammogram search. Of note, the LP Effect even occurred when the cancer was saliently highlighted by the CAD prompt (see also, Rich et al., 2008, for evidence of LP effect in 'pop-out' visual search). Importantly, Experiment 2 also showed that target detection was affected by the presence of CAD, consistent with the over-reliance hypothesis. When the CAD cue correctly identified the target, miss errors were relatively low. This was the case for both HP and LP trials. However, miss errors were greatly increased in situations where CAD either incorrectly identified the target or when no CAD cue appeared. The presence of CAD affected target detection so that people were more likely to respond in accordance with what CAD was indicating rather than what was present in the search display. Clearly CAD is useful when it is correctly used however participant's over-reliance on the technology hinders performance in conditions when CAD prompts fail.

In Experiment 3 participants searched for a range of cancerous and benign masses. The results showed that an LP Effect occurred where people made more miss errors at LP than at HP. This occurred for both cancerous and benign masses, although the LP Effect was more pronounced with cancerous targets. False alarms were noticeably higher (at 12%) than those witnessed in Experiments 1 and 2. Experiment 4 examined the effect of CAD and the over-reliance hypothesis on mammogram displays when there were a range of masses to search for. Similar to Experiment 2, fewer miss errors occurred when the target was correctly prompted by the CAD cue compared to conditions in which the CAD prompt failed to highlight the target. However, in this experiment an LP Effect did not occur. This did not seem to occur because participants missed fewer targets at LP, but more because when searching for a range of masses and the addition of CAD participants missed a higher proportion of targets even at high prevalence.

Experiment 5 ruled out the possibility that participants were carrying over an LP strategy into HP search as a similar effect was observed when the HP task was run in isolation. It is not the first time that high error rates have been found in high prevalence search. Kunar and Watson (2011) found that with more complex search displays and increased uncertainty of target identity from trial to trial miss rates were observed to reach around 20-30% (see also Kunar & Watson, 2014, Kunar, Thomas & Watson, 2017). Note that having a range of targets on its own did not lead to the increased miss errors in HP mammogram search as an LP Effect was found in Experiment 3 in the current paper. Instead, the combination of target uncertainty from trial to trial and CAD presence led to the increase in miss errors. Furthermore, when searching for a range of potential masses, the presence of CAD affected false alarms, again consistent with the over-reliance hypothesis. More false alarms were made in the presence of a CAD cue in comparison to when one was not presented (Experiments 4 and 5). Again participants showed

an over-reliance on the prompt, choosing to trust CAD more than their own judgement, even in the absence of a target.

False Alarms in LP Search

Our study showed that when searching for a range of masses false alarm errors were high (Experiments 3 – 5). Furthermore, as our findings showed that false alarms did not differ reliably across prevalence rates (and were even sometimes higher at LP than they were at HP, see Experiment 5) they showed little support for the Multiple Decision Model, which predicts fewer false alarms in LP search (Wolfe & Van Wert, 2010). We propose the increase in false alarms at both HP and LP search occurred as having multiple target templates led to weaker target representations in VWM (Mestry et al., 2016). In this case the ability to match the perceptual input of the current target onto the representation held in VWM would be susceptible to errors and misidentifications, leading to a greater probability of false alarms. As observers were asked to search for multiple targets at both high and low prevalence they would have had poor VWM representations of targets in both prevalence conditions.

Of course, during early diagnosis one could argue that detection of potential cancers is more important than generating false alarms. Although this may be true it is also important to note that for the women involved, false alarms also have their costs. In a clinical setting, false alarms can lead to an increase in recall rate of women being screened and therefore costly, invasive and unnecessary medical procedures. It is important to avoid such unnecessary tests not just because of the financial cost, but also due to the costs to the women who may undergo needless invasive tests (including needle biopsies of breast tissue) and experience increased worry associated with breast cancer for up to a year afterwards (Aro, 2000). Therefore any screening system should be optimally designed to minimise false alarms alongside miss errors. Given

that CAD has the potential to greatly increase both of these types of errors with erroneous prompts it leads to the question whether the benefits of CAD actually outweigh the cost.

Visual Working Memory and CAD

Mestry et al. (2016) theorised that having multiple target templates would lead to a weakening of target representations in VWM. Our data concur showing that with a weakening of target representations participants missed more targets and made more false alarms as their ability to match the perceptual input of the current target onto the weaker representation held in VWM was impaired. Other work has suggested that in order to search for multiple targets (e.g. a range of masses) one has to store the representation of these items in visual working memory. However, it is widely believed that there is a limit of the number of items that can be stored. Cowan (2001) suggested that visual working memory can store up to three to four items (Cowan, 2001) although others have suggested that this limit is restricted even further and that only one item, can be held in the focus of attention at any given time (Oberauer, 2002, see also Garavan, 1998, & McElree & Dosher, 1989). In order to change the item in the focus of attention an alternate object needs to be retrieved from working memory (Oberauer, 2002) which results in a switch cost. This switch cost may explain why people relied more heavily on the CAD prompts in Experiment 4 and 5, leading to high error rates in both HP and LP search. Given that such switch costs necessary to search through visual working memory require time (taking up to 300 – 500 ms, Garavan, 1998, see also Oberauer, 2002) and there is a dual task cost of searching through multiple targets (e.g. Godwin et al., 2010, Menneer et al., 2007; Menneer et al., 2009; Menneer et al. 2010) people may instead chose a heuristic to more willingly trust CAD rather than search through the many possible target representations in Visual Working Memory. This led to good search when the CAD cue was correct, but inaccurate search when the CAD cue was wrong. Such a strategy may relieve the cognitive

load on the observer in a laboratory setting, in which the negative consequences of making an error are minimal. However, further work is needed to investigate whether the same strategy shift occurs in a clinical setting where the cost of making an error is much higher.

CAD Errors of Mass Identification

Not only did CAD affect detection of the target but it also affected the identification of the mass. When the mass fell outside the CAD prompt participants made more errors in identifying the mass compared to when there was no CAD prompt or when the CAD prompt was used correctly. These results are particularly disturbing given that the identification of cancers were impaired when they appeared outside the CAD area. Although it is clear why CAD affected target detection we are not sure why CAD affected target identification. One potential reason for this impairment could be that the over-reliance on the CAD prompt also led to biases in target identification. That is, not only did participants come to believe that if a cancer was present it was to be *located* within the CAD prompt, but that they also formed a heuristic to believe that CAD cues should accurately *identify* cancers and that any mass falling outside the CAD prompt was likely to be benign. Lee and See (2004) found that observers often showed an overestimation of trust in the performance ability of automatic aids. Furthermore, automatization can lead to errors of commission where users accept the decision aid as correct, even in the face of conflicting information (Parasuraman & Manzey, 2010). In our study, both of these factors might have led participants to falsely overestimate the ability of the CAD cue to discriminate between cancers and benign masses and incorrectly conclude that any mass falling outside the cue was likely to be benign. Future work is needed to investigate this issue further. However, for now our data show that participants' over-reliance on CAD to detect the target also generalised to their ability to categorise the mass, showing that CAD can lead to both searching and classification errors.

Implications for CAD in a Clinical Setting

The data have implications for the use of CAD in mammogram search in a clinical setting. CAD technologies are currently used in a number of countries to help readers detect potential masses and microcalcifications in mammograms (Gilbert et al., 2008). The data presented here suggest that if CAD presentation correctly coincides with a mass then its detection and identification is improved. This occurs regardless of prevalence rates or mass type and is in line with other research suggesting a benefit in target detection with the use of correct exogenous CAD cues (e.g., Drew et al., 2012, Russell & Kunar, 2012). More worryingly, however, our findings reveal negative consequences of CAD. These occurred in situations where a mass was present but the CAD cue incorrectly indicated another area or failed to appear at all. Both of these situations led to higher miss error rates than situations where CAD was not used as a search tool. Although Drew et al. (2012) and Russell and Kunar (2012) found similar findings using letter visual search stimuli, here we present the first laboratory experiment to show that search for masses in mammograms suffer with incorrect CAD use. Observer's over-reliance on CAD and the belief that it will prompt a target has the potential to lead to serious consequences if a failed CAD cue leads to more cancers being missed in patient diagnosis.

Of course, there may be several important differences between our experiments and mammogram reading in a clinical setting. For example, CAD can be implemented in different ways to those presented in these experiments (e.g., CAD may be used interactively and remain invisible to the reader unless it is activated) and cancers can take on many more forms than those presented in our experiments (varying from being very obvious to subtle in appearance). Furthermore, clinical readers have had extensive training (more so than is practical in a laboratory setting) meaning that with such expertise, trained readers may not be susceptible to

missing targets due to incorrect CAD prompts. Nonetheless, there is a hint that incorrect CAD prompts can lead to missed cancers in the clinical literature. Zheng et al. (2004) reported that masses were more likely to be missed by radiologists if they appeared in a non-cued area. However, their study was limited by the number of observers tested and, perhaps more importantly, had masses appear at a higher prevalence rate (45 mass-positive versus 65 mass-negative images). Our study therefore complements the clinical findings with the benefit of having sufficient data for rigorous analysis. This increase in statistical power along with behaviour witnessed in a clinical setting suggests that the presence of an incorrect CAD cue might well have serious consequences for cancer detection.

Furthermore, our work complements previous work in the human factors literature which has also shown benefits and costs of automation. Parasuraman and Manzey (2010) suggested that the presence of automation in a task can change people's operating performance. For example, automation may lead to complacency where operators fail to notice important events due to substandard monitoring if they become reliant on technology (Billings et al., 1976, Parasuraman, Molloy & Singh, 1993). This complacency effect is present in both experts as well as non-expert observers (Singh et al., 1998, Galster & Parasuraman, 2001) and occurs even when the reliability of automation is low (May et al., 1983). Wickens and Dixon (2007) compared human performance when there was automation to conditions that had no automation and found that when reliability was below a certain threshold (70%), having automation resulted in worse performance than having no automation at all. Furthermore, there is the potential to misuse and disuse automation (Parasuraman & Riley, 1997) if automatic cues are highly salient or if observers put their trust in automation above other sources of information (Parasuraman & Manzey, 2010, Lee & See, 2004). In conclusion, although the use of technology in medicine can help understanding (Phelps et al., 2016) and reduce medical related

errors (Bates et al., 2001), automatic decision aids also have limitations if operators show an over reliance and complacency that leads to a detriment in patient care.

Acknowledgements

The authors would like to thank Peter Carr, Ranjeet Bassi and Anna Heinen for their help with data collection. This work was supported by a grant awarded to Melina Kunar from the British Academy and the Leverhulme Trust (SG122252), the Experimental Psychology Society and Research Development Funds from the University of Warwick. Sian Taylor-Phillips is supported by the NIHR CLAHRC West Midlands initiative. This paper presents independent research and the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

Aro, A.R. (2000) .False-positive findings in mammography screening induces short-term distress — breast cancer-specific concern prevails longer. *European Journal of Cancer, 36,* 1089-1097

Bates, D. Cohen, M, Leape, L., Overhage, M., Shabot, M. & Sheridan, T. (2001) Reducing the Frequency of Errors in Medicine Using Information Technology. *Journal of the American Medical Informatics Association* (8)

Bennett RL, Blanks RG, Moss SM. (2006) Does the accuracy of single reading with CAD (computer-aided detection) compare with that of double reading?: A review of the literature. Clin Radiol.;61(12):1023-8.

Billings, C. E., Lauber, J. K., Funkhouser, H., Lyman, G., & Huff, E. M. (1976). Aviation Safety Reporting System (Technical Report TM-X-3445). Moffett Field, CA: National Aeronautics and Space Administration Ames Research Center.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114.

Drew, T., Cunningham, C, Wolfe, J. M. (2012). When and why might a Computer Aided Detection (CAD) system interfere with visual search? An eye-tracking study. *Academic Radiology, 19,* 1260-1267.

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If You Don't Find It Often, You Often Don't Find It: Why Some Cancers Are Missed in Breast Cancer Screening. *PloS one, 8(5),* e64366.

Evans K.K., Evered A., Tambouret R.H., Wilbur D.C. & Wolfe J.M. (2011) Prevalence of Abnormalities Influences Cytologists' Error Rates in Screening for Cervical Cancer. *Archives of Pathology & Laboratory Medicine, 135(12*), 1557-1560.

Fenton JJ , Taplin SH , Carney PA , et al (2007) . Influence of computer-aided detection on performance of screening mammography . *N Engl J Med; 356* (14): 1399 – 1409 .

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science , 18* (11), 943-947.

Galster, S., & Parasuraman, R. (2001). Evaluation of countermeasures for performance decrements due to automated-related complacency in IFR-rated general aviation pilots. In Proceedings of the International Symposium on Aviation Psychology (pp. 245–249). Columbus, OH: Association of
Aviation Psychology.

Garavan, H. (1998). Serial attention within working memory. *Memory & Cognition, 26,* 263–276.

Gilbert FJ, Astley SM, Gillan MG, Agbaje OF, Wallis MG, James J, Boggis CR, Duffy SW, (2008) the CADET II Group: Single reading with computer-aided detection for screening mammography. *N Engl J Med, 359:*1675-1684.

Godwin, H.J., Menneer, T., Donnelly, N. and Cave, K.R. (2010) Dual-target search for high and low prevalence x-ray threat targets. *Visual Cognition*, 18, (10), 1439-1463.

Green, D. M., & Swets, J. A. (1967). Signal detection theory and psychophysics. New York: John Wiley and Sons.

Gur D, Rockette HE, Armfield DR, Blachar A, Bogan JK, et al. (2003) Prevalence effect in a laboratory environment. *Radiology 228* 10–14.

Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, et al. (2008) The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology 249(1*): 47–53.

Hancock PA. (2013) In search of vigilance: the problem of iatrogenically created psychological phenomena. *Am Psychol. 68(2*):97-109.

Heath, M., Bowyer, K., Kopans, D., Moore, R. and Kegelmeyer, W.P. (2001) *Proceedings of the Fifth International Workshop on Digital Mammography*, M.J. Yaffe, ed., 212-218, Medical Physics Publishing, ISBN 1-930524-00-5.

Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, W.P., Moore, R., Chang, K. and MunishKumaran, S. (1998) *Digital Mammography*, 457-460, Kluwer Academic Publishers, Proceedings of the Fourth International Workshop on Digital Mammography.

Julesz, B. (1981). A theory of preattentive texture discrimination based on first order statistics of textons. *Biology and Cybernetics, 41,*131-138.

Kunar, M. A. & Humphreys, G. W. (2006). Object-based inhibitory priming in preview search: Evidence from the 'top-up' procedure. *Memory & Cognition, 34,* 459-474.

Kunar, M.A., Rich, A.N. & Wolfe, J.M. (2010). Spatial and temporal separation fails to counteract the effects of low prevalence in visual search. *Visual Cognition, 18,* 881-897.

Kunar, M. A., Thomas, S.V. & Watson, D.G. (2017). Time-based selection in complex displays: Visual Marking does not occur in Multi-Element Asynchronous Dynamic (MAD) search. *Visual Cognition.* doi: 10.1080/13506285.2017.1306006

Kunar, M.A. & Watson, D.G. (2011). Visual Search in a Multi-element Asynchronous Dynamic (MAD) World. *Journal of Experimental Psychology: Human Perception and Performance, 37(4),* 1017-1031.

Kunar, M. A. & Watson, D. G. (2014). When Are Abrupt Onsets Found Efficiently in Complex Visual Search?: Evidence from Multi-Element Asynchronous Dynamic Search. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 232-252.

Kunar, M. A. & Wolfe, J. M. (2011). Target Absent Trials in Configural Contextual Cueing. *Attention, Perception and Psychophysics. 73 (7),* 2077-2091.

Lee, J. D., & See, J. (2004). Trust in automation and technology: Designing for appropriate reliance. Human Factors, 46, 50–80.

Macklis RM, Meier T, Weinhous MS. (1998) Error rates in clinical radiotherapy. *J Clin Oncol.;16*:551–6.

Macmillan, N. A. & Creelman, C. D. (2005). Detection Theory: A User's Guide, 2nd Edition, Cambridge University Press.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. Psychological Bulletin, 98, 185-199.

May, P., Molloy, R., & Parasuraman, R. (1983, October). Effects of automation reliability and failure rate on monitoring performance in a multi-task environment. Paper presented at the annual meeting of the Human Factors Society, Santa Monica, CA.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General, 118,* 346–373.

Menneer, T., Barrett, D. J. K., Phillips, L., Donnelly, N., & Cave, K. R. (2007). Costs in searching for two targets: Dividing search across target types could improve airport security screening. *Applied Cognitive Psychology, 21,* 915-932.

Menneer, T., Cave, K. R., & Donnelly, N. (2009). The cost of search for multiple targets: the effects of practice and target similarity. *Journal of Experimental Psychology: Applied, 15,* 125-139.

Menneer, T, Donnelly, N, Godwin, H. J. and Cave, K. R. (2010) High or low target prevalence increases the dual-target cost in visual search. *Journal of Experimental Psychology: Applied*, 16, (2), 133-144.

Mestry, N, Menneer, T, Cave, K. R., Godwin, H and Donnelly, N. (2016) Dual-target cost in visual search for multiple unfamiliar faces. *Journal of Experimental Psychology: Human Perception and Performance*, 1-69.

Mitroff, S. R., & Biggs, A. T. (2014). The Ultra-Rare-Item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science, 25*(1), 284-289. DOI: 10.1177/0956797613504221

NHS Breast Screening Programme (2015) Breast Screening Program, England, Statistics for 2013-2014. Retreived 29th June 2016 from http://www.hscic.gov.uk/catalogue/PUB16803

National Institutes of Health. Breast Cancer Surveillance Consortium. Website. http://breastscreening.cancer.gov. Accessed February 10, 2015

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 411–421.

Obenauer S , Sohns C , Werner C , Grabbe E . (2006) Computer-aided detection in full-field digital mammography: detection in dependence of the BI-RADS categories. *Breast J* ;*12(1):16–19*

Parasuraman, R. & Manzey, D.D. (2010). Complacency and bias in human use of automation: an attentional integration. Human Factors, 52(3), 381–410.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." International Journal of Aviation Psychology, 3, 1–23.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39, 230–253.

Peltier, C., & Becker, M. W. (2016, May 5). Decision Processes in Visual Search as a Function of Target Prevalence. Journal of Experimental Psychology: Human Perception and Performance. Advance online publication. http://dx.doi.org/10.1037/xhp0000248

Phelps, E. E., Wellings, R., Griffiths, F., Hutchinson, C. & Kunar, M. A. (2016). Do medical images aid understanding and recall of medical information? An experimental study comparing the experience of viewing no image, a 2D medical image and a 3D medical image alongside a diagnosis. *Patient Education and Counseling,* doi: 10.1016/j.pec.2016.12.034

Philpotts LE (2009). Can Computer-aided Detection Be Detrimental to Mammographic Interpretation? *Radiology, 253*:17-22.

Rayat P. (2016) Breast Screening Programme England statistics for 2014-15. London, England. http://www.hscic.gov.uk/article/2021/Website-Search?productid=20270&q=breast+screening &sort=Relevance&size=10&page=1&area=both#top. Published February 24, 2016. Accessed April 21, 2016.

Rich, A. N., Kunar, M. A., Van Wert, M. J., Hidalgo-Sotelo, B., Horowitz, T. S., & Wolfe, J. M. (2008). Why do we miss rare targets? Exploring the boundaries of the low prevalence effect. *Journal of Vision , 8* (15), 1-17.

Russell, N. & Kunar, M. A. (2012). Color and Spatial Cueing in Low Prevalence Visual Search. *The Quarterly Journal of Experimental Psychology, 65,* 1327-1344.

Samulski, M., Hupse, R., Boetes, C. Mus, R., Heeten, G., Karssemeijer, N. (2010). Using computer-aided detection in mammography as a decision support. *Eur Radiol., 20,* 2323-2330.

Singh, I. L., Molloy, R., Mouloua, M., Deaton, J., & Parasuraman, R. (1998). Cognitive ergonomics of cockpit automation. In I. L. Singh & R. Parasuraman (Eds.), Human cognition: A multidisciplinary perspective (pp. 242–253). New Delhi, India: Sage.

Soo MS , Rosen EL , Xia JQ , Ghate S , Baker JA (2005) . Computer-aided detection of amorphous calcifications. *AJR Am J Roentgenol,* 184(3):887–892

Taylor-Phillips, Sian, Wallis, Matthew G., Jenkinson, David J., Adekanmbi, Victor, Parsons, Helen, Dunn, Janet A., Stallard, Nigel, Szczepura, Ala, Gates, Simon, Kearins, Olive, Duncan, Alison, Hudson, Sue, Clarke, Aileen. 2016. Effect of using the same vs different order for second readings of screening mammograms on rates of breast cancer detection : a randomized clinical trial. *JAMA: The Journal of the American Medical Association*, 315 (18), 1956-1965.

Treisman, A. M., & Gelade, G. (1980). A feature-integreation theory of attention. *Cognitive Psychology* (12), 97-136.

Van Wert, M. J., Horowitz, T. S., & Wolfe, J. M. (2009). Even in correctable search, some types of rare targets are frequently missed. *Attention, Perception & Psychophysics , 71* (3), 541-553.

Watson, D. G. & Kunar, M. A. (2010). Visual marking and change blindness: Moving occluders and transient masks neutralize shape changes to ignored objects. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1391-1405.

Watson, D. G. & Kunar, M. A. (2012). Visual Marking: Determining the capacity of time-based selection. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 350-366.

Wickens, C. D., & Dixon, S. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science, 8, 201–212.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science, 9* (1), 33-39.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15* (3), 419-433.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual search. *Nature , 435*, 439-440.

Wolfe, J. M., Horowitz, T. S., Ven Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology , 136* (4), 623-638.

Wolfe, J.M., and VanWert, M.J. (2010). Varying target prevalence reveals two, dissociable decision criteria in visual search. *Current Biology, 20*, 121-124.

Zheng B., Swensson, R.G., Golla, S., et al. (2004). Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments, *Acad. Radiol. 11*, 398–406.

Table 1: Percentage of Initial and Self-Corrected Miss Errors in Experiment 1. Standard Errors are reported in the Parentheses.

	Initial Errors	**Self-Corrected Errors**
Mammogram – HP	1.8 (0.4)	0.7 (0.4)
Mammogram – LP	18.5 (3.6)	10.2 (2.8)
Letter Visual Search – HP	19.3 (2.4)	15.4 (2.4)
Letters Visual Search - LP	39.8 (3.3)	32.1 (2.9)

Table 2: Percentage of Miss Errors in Experiments 1 - 5. Standard Errors are reported in the parentheses.

Condition	Miss Errors - HP	Miss Errors - LP
Experiment 1		
Mammogram	0.7 (0.4)	10.2 (2.8)
Letter Visual Search	15.4 (2.4)	32.1 (2.9)
Experiment 2		
Correct CAD	0.1 (0.1)	5.6 (1.8)
Incorrect CAD	2.5 (1.1)	14.8 (3.4)
No CAD	1.7 (1.1)	25.8 (5.1)
Experiment 3		
Benign	2.7 (0.8)	8.5 (1.6)
Cancer	8.5 (1.7)	20.0 (2.8)
Experiment 4		
Correct CAD - Benign	2.4 (0.8)	4.0 (1.5)
Incorrect CAD - Benign	22.3 (5.8)	18.8 (5.5)
No CAD - Benign	4.6 (1.3)	6.6 (1.7)
Correct CAD - Cancer	2.1 (0.7)	3.0 (1.0)
Incorrect CAD - Cancer	24.8 (5.6)	24.4 (5.1)
No CAD - Cancer	13.5 (3.3)	16.3 (3.1)
Experiment 5		
Correct CAD - Benign	3.3 (0.8)	n/a
Incorrect CAD - Benign	15.4 (4.2)	n/a
No CAD - Benign	4.4 (1.4)	n/a
Correct CAD - Cancer	2.6 (0.8)	n/a
Incorrect CAD - Cancer	17.7 (4.4)	n/a
No CAD - Cancer	14.6 (2.3)	n/a

Table 3: Percentage of False Alarms in Experiments 1 - 5. Standard Errors are reported in the parentheses.

Condition	False Alarms - HP	False Alarms - LP
Experiment 1		
Mammogram	0.6 (0.3)	0.2 (0.1)
Letter Visual Search	0.4 (0.2)	0.9 (0.7)
Experiment 2		
Incorrect CAD	2.8 (2.1)	2.8 (1.1)
No CAD	0.8 (0.3)	2.0 (0.9)
Experiment 3		
All Absent trials	10.8 (2.9)	13.1 (4.3)
Experiment 4		
Incorrect CAD	45.6 (8.1)	47.0 (8.0)
No CAD	27.4 (7.1)	32.2 (7.3)
Experiment 5		
Incorrect CAD	11.25 (3.1)	n/a
No CAD	6.9 (2.9)	n/a

Table 4: Mean RTs (in ms) for present trials across Experiments 1 - 5. Standard Errors are reported in the parentheses.

Condition	HP	LP
Experiment 1		
Mammogram	696.4 (27.3)	1082.9 (68.8)
Letter Visual Search	1706.2 (77.1)	2177.3 (113.0)
Experiment 2		
Correct CAD	669.3 (24.0)	900.0 (49.7)
Incorrect CAD	708.3 (26.8)	980.7 (58.4)
No CAD	762.8 (39.4)	1087.5 (60.6)
Experiment 3		
Benign	1652.7 (140.4)	2048.7 (138.1)
Cancer	2072.6 (202.6)	2318.4 (182.9)
Experiment 4		
Correct CAD - Benign	1128.2 (91.9)	1343.5 (140.7)
Incorrect CAD - Benign	1479.4 (129.1)	1555.4 (129.4)
No CAD - Benign	1193.6 (91.8)	1393.8 (141.3)
Correct CAD - Cancer	1251.4 (109.2)	1403.0 (119.4)
Incorrect CAD - Cancer	1622.3 (153.2)	1423.6 (143.6)
No CAD - Cancer	1398.3 (120.3)	1587.5 (156.3)
Experiment 5		
Correct CAD - Benign	1546.3 (142.9)	n/a
Incorrect CAD - Benign	1777.1 (141.0)	n/a
No CAD - Benign	1625.6 (153.1)	n/a
Correct CAD - Cancer	1772.4 (188.0)	n/a
Incorrect CAD - Cancer	2408.2 (224.0)	n/a
No CAD - Cancer	1960.9 (212.6)	n/a

Table 5: Mean RTs (in ms) for absent trials across Experiments 1 - 5. Standard Errors are reported in the parentheses.

Condition	HP	LP
Experiment 1		
Mammogram	1044.6 (112.0)	931.9 (88.7)
Letter Visual Search	2871.0 (158.6)	2329.9 (156.6)
Experiment 2		
Incorrect CAD	1033.1 (93.8)	747.7 (68.2)
No CAD	872.3 (80.4)	669.2 (69.1)
Experiment 3		
All Absent trials	2069.2 (205.3)	1821.2 (251.3)
Experiment 4		
Incorrect CAD	1299.6 (151.5)	1364.4 (141.2)
No CAD	1410.4 (134.7)	1370.9 (115.4)
Experiment 5		
Incorrect CAD	2453.0 (329.3)	n/a
No CAD	2290.9 (325.9)	n/a

Table 6: Percentage of masses correctly identified across the different CAD conditions in Experiments 4 and 5. Standard Errors are reported in the parentheses.

	Incorrect CAD	**NO CAD**	**Correct CAD**
Experiment 4 - Benign	69.5 (4.7)	73.6 (4.4)	77.9 (4.4)
Experiment 4 - Cancer	60.1 (5.9)	73.7 (5.3)	72.3 (5.6)
Experiment 5 - Benign	74.1 (3.0)	77.2 (2.4)	82.1 (2.4)
Experiment 5 - Cancer	72.2 (4.4)	81.1 (3.4)	81.6 (3.5)

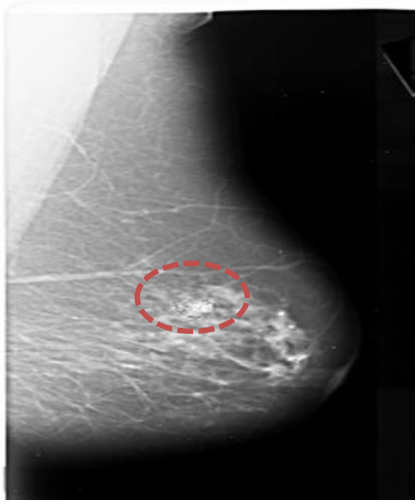
Figure Legends

Figure 1. Example displays of (a) mammogram search where participants have to search for a cancer in a mammogram (for reader clarity, the cancer in this image is in the dotted line. Please note the line did not appear in the experiment proper) and (b) letter visual search task where participants have to search for a T among Ls.

Figure 2. Example displays of CAD use in Experiment 2. In Figure 2a the CAD prompt correctly highlights the cancer. In Figure 2b the cancer falls outside the CAD region.

Figure 3. Example displays of a benign mass (Figure 3a) and a cancerous mass (Figure 3b) and in Experiment 3. For reader clarity, the masses are highlighted by the dotted line. Please note the line did not appear in the experiment proper

(a)



(b)

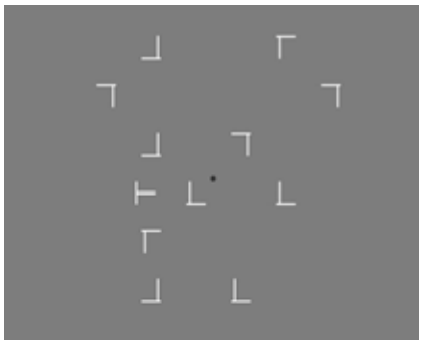
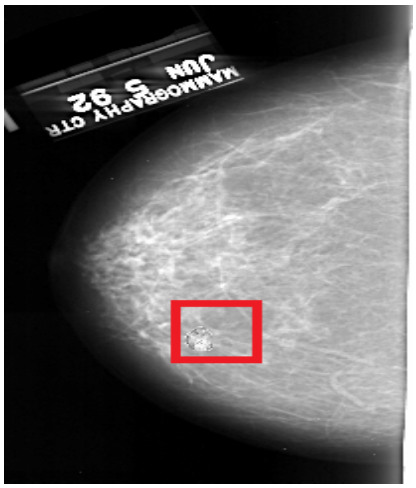


Figure 1

(a)



(b)

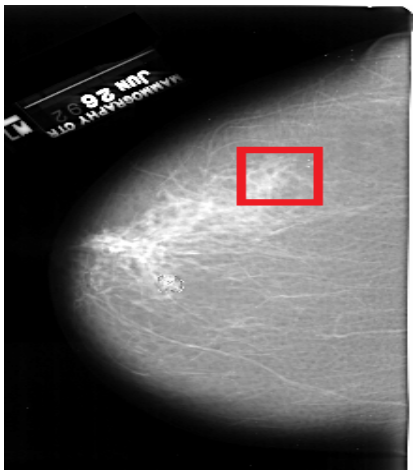
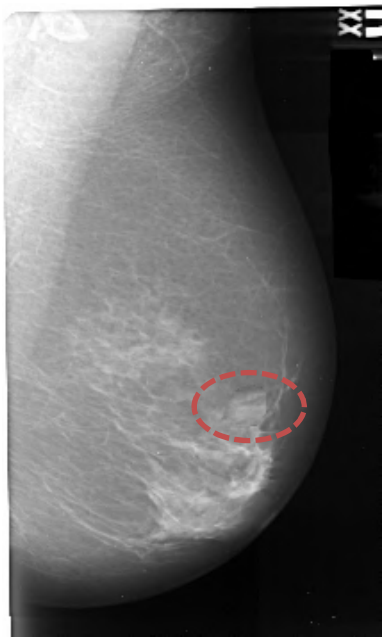


Figure 2



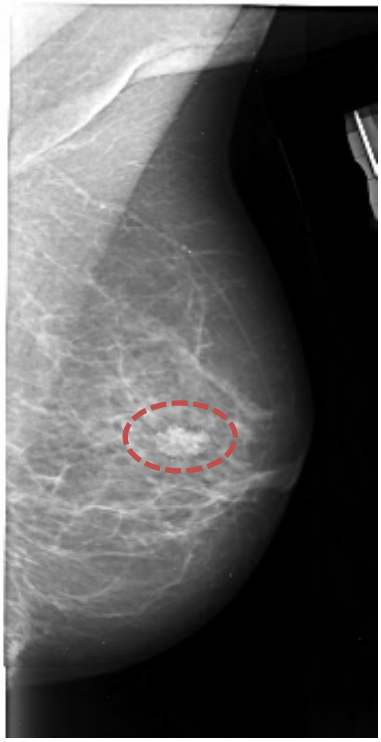


Figure 3