Cognitive Research: Principles and Implications

ORIGINAL ARTICLE

Open Access

The message matters: changes to binary Computer Aided Detection recommendations affect cancer detection in low prevalence search

Francesca Patterson[1]* and Melina A. Kunar[1]

Abstract

Computer Aided Detection (CAD) has been used to help readers find cancers in mammograms. Although these automated systems have been shown to help cancer detection when accurate, the presence of CAD also leads to an over-reliance effect where miss errors and false alarms increase when the CAD system fails. Previous research investigated CAD systems which overlayed salient exogenous cues onto the image to highlight suspicious areas. These salient cues capture attention which may exacerbate the over-reliance effect. Furthermore, overlaying CAD cues directly on the mammogram occludes sections of breast tissue which may disrupt global statistics useful for cancer detection. In this study we investigated whether an over-reliance effect occurred with a binary CAD system, which instead of overlaying a CAD cue onto the mammogram, reported a message alongside the mammogram indicating the possible presence of a cancer. We manipulated the certainty of the message and whether it was presented only to indicate the presence of a cancer, or whether a message was displayed on every mammogram to state whether a cancer was present or absent. The results showed that although an over-reliance effect still occurred with binary CAD systems miss errors were reduced when the CAD message was more definitive and only presented to alert readers of a possible cancer.

Keywords Mammogram, Artificial intelligence, Low prevalence, Computer Aided Detection (CAD), Over-reliance, Binary CAD, Automation

Introduction

Observers often miss rare targets in visual search at disproportionately high rates (Wolfe et al., 2005). Previous laboratory research involving participants searching for a target amongst a set of distractors suggests that as the prevalence rate of the target decreases, the proportion of miss errors (failing to notice the target) increase (e.g., Godwin et al., 2015; Hout et al., 2015; Kunar et al., 2010, 2021; Mitroff & Biggs, 2014; Rich et al., 2008; Russell & Kunar, 2012; Van Wert et al., 2009; Wolfe et al., 2005, 2007). This low prevalence (LP) effect has also shown to be robust against real-world tasks where targets are rare, such as breast screening, where radiologists inspect mammogram images for breast cancers. For example, it is estimated that 20–40% of cancers are missed in initial screening (Bird et al., 1992; see also Evans et al., 2013) and research by Evans et al. (2013) has highlighted LP as a cause of miss errors for expert mammographers in a clinical setting.[1] Failing to detect a cancer in radiology poses serious health risks, and as such, it is vital to find ways to help improve cancer detection.

*Correspondence:
Francesca Patterson
frankiepatterson16@gmail.com
[1] Department of Psychology, The University of Warwick, Coventry CV4 7AL, UK

[1] This relates to routine screening mammography, for women who have no signs or symptoms of breast cancer. The prevalence of breast cancer is higher in symptomatic screening of mammograms.

Computer Aided Detection (CAD) has been developed to aid operators in the difficult perceptual task of cancer detection by using computer algorithms to highlight suspicious features of a mammogram (Castellino, 2005). CAD typically works by overlaying a salient visual cue on the breast tissue to indicate the location of a potentially suspicious area which radiologists would then need to verify. CAD has been approved for use in radiology by the Food and Drug Administration in the USA, with a primary goal of increasing cancer detection, providing a more efficient workflow and reducing demands on radiologists (Castellino, 2005). Standard practice involves the radiologist first viewing the mammogram in the absence of CAD, then activating CAD and re-evaluating the image before issuing their final conclusion (Castellino, 2005). In laboratory studies, this reading mode has been shown to provide the optimal outcome in terms of cancer detection in comparison to conditions where readers were simultaneously presented with CAD cues on first reading of the mammogram (Kunar, 2022).

Whilst many studies have investigated the effectiveness of CAD, historically its assets and liabilities have remained controversial. For example, Lehman et al. (2015) compared the effect of CAD on digital screening mammography performance in terms of sensitivity, specificity, and cancer detection rate in a large-scale multi-screening centre study. They found that there was no improvement in screening performance when mammograms were read with CAD, compared to when CAD was not used. Furthermore, Bennett et al. (2006) conducted a literature review to compare single reading with CAD to double reading procedures (where two radiologists read the mammographic images), but differences in methodology produced indefinite conclusions. Using pooled estimates of effect sizes from two meta-analyses, Taylor and Potts (2008) found that there was no significant difference in cancer detection rates between single reading with CAD and double reading. However, differences in screening programmes used in the meta-analyses may have affected these results. While a double reading procedure remains an effective method for cancer detection (Kunar et al., 2021; Taylor & Potts, 2008), it may not be a feasible long-term approach due to the increasing number of women needing screening and the demands this place on an already limited workforce (Chen et al., 2023; Guerriero et al., 2011; James et al., 2010). Instead, recent developments in automation and Artificial Intelligence (AI) suggest that there could be improved efficacy of CAD for use in medical diagnostic imaging with systems using AI deep learning models (e.g., Fujita, 2020; Salim, et al., 2020). For example, recent research using AI as an independent supporting reader in breast cancer screening, was found to be a comparable (and in some cases superior) method to human double reading (Ng et al., 2023).

With this advancement in AI capabilities, Ng et al. (2023) also note that there is a need to evaluate new strategies for using AI technology, as supporting readers, alongside humans. Previous research has shown that the way CAD is presented to humans affects their ability to detect cancers (Kunar, 2022; Kunar & Watson, 2023). For example, it has been shown that the presence of a CAD prompt can result in an *over-reliance effect* which biases reader judgements depending on the accuracy of CAD (Kunar, 2022; Kunar & Watson, 2023; Kunar et al., 2017; Zheng et al., 2004). The over-reliance effect shows that while the use of a CAD prompt that correctly highlights a cancer decreases the amount of miss errors, there is a large increase in miss errors when the CAD system fails. For example, if the CAD cue fails to highlight a cancer or the cancer falls outside of the highlighted area, miss errors are increased in comparison to if no CAD system is used (Drew et al., 2020; Kunar et al., 2017). Furthermore, incorrect CAD cues lead to higher false alarms and subsequently recall rates where women are incorrectly recalled for further assessment (e.g., Fenton et al., 2007; Kunar, 2022; Kunar & Watson, 2023; Kunar et al., 2017). Although one could argue that miss errors, where a cancer goes undetected has potentially more serious consequences for the women involved, an increase in false alarms also has its own problems. Women who have been recalled due to a false alarm have been shown to experience psychological distress in relation to their experience (Aro, 2000) and are more likely to delay participation, or not participate at all, in future mammography screening (Kahn & Luce, 2003). Furthermore, an increase in false alarms creates unnecessary demands on healthcare systems which are already over-burdened and in crisis due to a shortage of healthcare workers (e.g., Darzi & Evans, 2016; Konstantinidis, 2023).

The over-reliance effect of using CAD is robust and is difficult to remove (although it can be mitigated in some circumstances; Kunar & Watson, 2023). One of the reasons proposed for this robust over-reliance effect is that CAD cues often use salient exogenous cues to alert the location of a cancer (Drew et al., 2020; Kunar & Watson, 2023). Exogenous cues are thought to receive higher weightings in attentional priority maps and thus capture attention automatically (Remington et al., 1992; Theeuwes, 2004; Wolfe, 2021). Even when people are told to specifically ignore these CAD cues, they still seem to elicit attentional priority leading to beneficial results when they cue the cancer, but over-reliance effects when they do not (Kunar & Watson, 2023). Having the CAD cue be a salient marker and physical presence on a mammogram may also result in other issues. For example, having a salient marker overlayed onto the mammogram may occlude parts of the breast tissue and interfere with the global regularities or 'gist' statistics of the mammogram. This is important as the ability to process the global image statistics of breast tissue is known to be an influential factor in detecting abnormalities in mammography (Evans et al., 2016; Raat et al., 2023).

In light of this, other CAD systems have been proposed. For example, Goldenberg and Peled (2011) discussed the advantages of CAD systems that output a simple binary recommendation of 'positive' or 'negative', indicating the presence or absence of an anomaly. Here, instead of a salient CAD cue being presented on a mammogram, a message indicating the presence or absence of a cancer would be presented elsewhere on the screen which radiologists would then have to verify (for example the message would state that a cancer is present or that a cancer is absent). This type of CAD system has recently been implemented in some AI systems within mammography where the AI system will result in a binary message to recall a woman (as a suspected cancer is present) or to not recall a woman (as either there is no cancer present or the AI has failed to detect a possible malignancy; Ng et al., 2023). CAD systems that generate binary Recall/No Recall recommendations have an advantage as they provide assistance and recommendations to human readers without the need for salient and exogenous CAD markers that are overlaid on the mammogram. However, sparse research has been conducted investigating how this type of binary message affects human decision making and in particular whether there is an over-reliance effect with this type of CAD prompt. As binary CAD systems do not use exogenous, salient cues to highlight a potential region of interest on a mammogram, then there will not be complications from over-laying a strong salient attentional cue on the breast tissue. Therefore, without these exogenous cues capturing attention the over-reliance effect previously observed with CAD markers may be reduced or eliminated. In contrast as humans are susceptible to biases, particularly in relation to recommendations given by technology (e.g., Salim Jr et al., 2023; Wysocki et al., 2023) it may be that an over-reliance effect is still observed even in the absence of these salient cues. We investigated this here.

The present study was the first investigation (at least, that the authors are aware of) into the human–computer interaction of a binary CAD system in search for a low prevalence cancer. We investigated whether users demonstrated an over-reliance on binary CAD, whereby a simple message indicated that a cancer may be present or absent. We also investigated whether there was an optimal way to present these binary CAD cues as previous research has shown that the way CAD is presented affects cancer detection. In particular, we varied the CAD messages, first, by their degree of certainty and, second, by whether the CAD message was presented only on some of the mammogram images to indicate the possible presence of a cancer or whether it was presented on every mammogram, stating that a cancer was either present or absent.

In relation to CAD certainty, it has been found that framing a CAD system to be more fallible led to a reduction in the over-reliance effect (Kunar & Watson, 2023). That is, people were more likely to perform a more exhaustive search if they were told that the CAD system was less accurate. This is particularly important in relation to search for LP targets where it is found that people often terminate their search prematurely, before they have searched the display in full (Wolfe & Van Wert, 2010). Therefore, we investigated whether manipulating the CAD message to either be definitive (i.e., a cancer *is present*—Experiments 1 and 3) or instead to be more probabilistic and less certain (i.e. a cancer *is likely*—Experiments 2 and 4) would lead to a difference in target detection.

Furthermore, we manipulated whether it was better to present CAD messages in situations where it only alerted readers to the possible presence of a cancer or whether it would be better to present messages on all mammograms to state that a cancer was either present or absent. We predicted that giving a message on some of the mammogram images versus giving a message on all mammogram images would lead to a difference in the over-reliance effect. For example, miss errors may be more pronounced under conditions where it was *explicitly* stated that a cancer was absent in comparison to when no message was shown. In this latter condition we hypothesised that, with no explicit recommendation, people would be more likely to perform an exhaustive search of the mammogram, which would result in fewer miss errors. Accordingly, in Experiments 1 and 2, we only presented the CAD message on some of the mammogram images to alert people to the possible presence of a cancer. In these experiments, if participants were shown a CAD message it was a *recall* message where we indicated the potential presence of a cancer (i.e., that a cancer was 'present' in Experiment 1 or 'likely' in Experiment 2). In a clinical setting, a 'recall' message would suggest indications of a possible cancer and that the woman should return to the clinic for further assessment. In contrast, a 'no recall' message would suggest no further assessment was needed. In Experiments 3 and 4 we presented a CAD message on *all* images of the mammograms to indicate whether a cancer was present or not. In these experiments a CAD message would either contain a *recall* or *no-recall* message (i.e., stating that a cancer was 'present' or 'absent', respectively in Experiment 3, and 'likely' or 'not likely', respectively in Experiment 4).

Throughout this study we used laboratory-based experiments where we recruited non-expert observers to search for a low prevalence simulated cancer in a mammogram. Lab-based experiments have been found to successfully examine search behaviour in applied settings (Cunningham et al., 2017; Drew et al., 2012; Drew et al., 2020; Kunar et al., 2017; Kunar et al., 2021; Kunar, 2022; Kunar & Watson, 2023; Raat et el., 2023) and are a legitimate way to test a variety of conditions that would not be otherwise feasible. Lab-based experiments using non-expert observers enable the presentation and testing of a range of CAD display options, which would not be practical in Randomised Clinical Trials (RCTs) and difficult to test with radiologists, given the demands on their time. For example, RCTs are often expensive and time-consuming. By the point of RCTs it would be important to know which are the optimal ways to present CAD and which to rule out. This information can be determined from lab-based studies, which can then be used to inform trials in a clinical setting. Furthermore, research findings that have been found to be observed in the lab have also been observed in clinical settings (Evans et al., 2013) and given that all humans share the same underlying search principles, it has been established that the search behaviour of non-expert participants is similar to those of clinicians (Drew et al., 2012; Taplin et al., 2006; Wolfe et al., 2016). Therefore, these lab-based experiments provide a valid way to investigate how binary CAD affects cancer detection.

Method

Transparency and openness

The data can be found on the Open Science Framework (https://osf.io/jgxz5/). All data were compiled in Microsoft® Excel® for Microsoft 365 MSO (Version 2112 16.0.14729.20254) and imported into SPSS (Version 27, Release 27 0.1.0) and JASP (Version 0.16; JASP Team, 2021) for statistical analysis. The experimental programs were written in PsychoPy (Peirce et al., 2019). The study design, hypotheses and analytic plan were not pre-registered. All manipulations, data exclusions and measures are reported.

Participants

All participants were recruited via the University of Warwick's Research Experience panel and received course credit as compensation for their participation in the study. A G* power analysis (alpha = 0.05, effect size = 0.5) determined that a minimum sample size of 34 participants per experiment was required to achieve a power of 0.8. Participant numbers varied slightly across experiments due to participants opting out of completing the experiment and choosing to partake in a separate activity for course credit. Thirty-nine participants took part in Experiment 1, thirty-nine participants took part in Experiment 2, thirty-eight participants took part in Experiment 3, and thirty-six participants took part in Experiment 4. None of the participants took part in more than one experiment.

Stimuli

Stimuli included 338 images of mammograms sourced at random from the volume of 695 normal mammograms (those not containing a cancer) on the Digital Database for Screening Mammography (DDSM; Heath et al., 2001). Images were presented in the centre of a computer display and subtended approximately 10.7 degrees by 18.6 degrees at a viewing distance of 57 cm in size (please note that the actual size of each image varied as they were of real mammograms). They were categorised into 'present' images (an image of a cancerous mass sourced from the cancer volume of the DDSM was transposed onto the normal mammogram images using image editing software) and 'absent' images where no cancer was shown.

Procedure

The experiment was created using PsychoPy, presented on a PC, and took approximately 30 min to complete. Participants were instructed to search for a cancer in the mammogram images being presented to them and were presented with example images of both mammograms containing a cancer and mammograms that did not. To familiarise participants with the task, participants were then asked to complete a training set, whereby they were presented with 10 present images and 10 absent images and were required to respond whether a cancer was present or absent via a two-alternative forced choice task and pressing the 'm' or 'z' key, respectively. Participants were required to respond correctly on 70% of the training set trials in order to continue to the experiment proper. If they failed to do so, they were given four attempts at re-completing the training block until they got (at least) 70% correct. The training block ensured that participants were able to recognise the appearance of a cancer. Following this, participants were given the experimental instructions and informed that during the experiment a cancer would be rarely present within the mammogram displays. Example images of the four trial types for that experiment were presented (manipulating both the presence of the cancer and the CAD message), followed by a practice block. Once completed, participants proceeded to the experimental block.

For each experimental block, there were 300 trials: 30 present images and 270 absent images (to give a 10%

Table 1 Summary of experimental conditions

Experiment	*N*	Recall—CAD message	No recall—CAD message	CAD occurrence
1	39	'Cancer Present'	No message	Some trials
2	39	'Cancer Likely'	No message	Some trials
3	38	'Cancer Present'	'Cancer Absent'	All trials
4	36	'Cancer Likely'	'Cancer Not Likely'	All trials

prevalence rate). For the present trials, 20 trials contained a 'Recall' CAD prompt explicitly indicating the presence of a cancer. The remaining 10 trials contained no CAD prompt in Experiments 1 and 2 or an explicit 'No Recall' CAD prompt indicating the absence of a cancer in Experiments 3 and 4. For absent trials, 180 images contained no CAD prompt in Experiments 1 and 2 or contained an explicit 'No Recall' CAD prompt in Experiments 3 and 4. The other 90 mammogram images contained a Recall CAD prompt. The CAD accuracy rate in these experiments was chosen to reflect CAD accuracy in a clinical setting, which is estimated to vary from 57% (Soo et al., 2005) to 85% (Obenauer et al., 2006; see also Henriksen et al., 2019, who report a CAD accuracy of between 65 and 77%). Therefore, CAD accuracy of 67% in these experiments falls within this range. Participants were not informed of the CAD accuracy rate but told that in some trials the CAD cue would give accurate information and in some trials it would not.

For Experiments 1 and 3 the Recall CAD prompt gave the message 'Cancer Present'. For Experiments 2 and 4 the Recall CAD prompt showed the less definitive message of 'Cancer Likely'. In Experiments 1 and 2 the CAD prompt (i.e., a Recall CAD message) appeared on only some of the trials. For Experiments 3 and 4 a CAD prompt appeared on all the trials. Therefore, for Experiments 1 and 2, on 'No Recall' trials there was no CAD prompt shown, whereas in Experiment 3, the No Recall message was 'Cancer Absent' and in Experiment 4, the No Recall message was 'Cancer Not Likely'. Participants were made aware of what the CAD message would say before each experiment (e.g., 'Cancer Present' or 'Cancer Likely'). Tables 1 and 2 give a summary of the different experimental conditions. Example stimuli can be found in Fig. 1.

In both the practice and experimental blocks, for each trial participants were asked to respond whether a cancer was present or absent by pressing the 'm' or 'z' key, respectively. Images were presented to each participant in a random order and remained on the screen until participants gave a response. Reaction Times (RTs) and error rates were recorded. In accordance with Fleck and Mitroff's (2007) theory that the LP effect could be due to response-execution motor errors (where participants made motor errors by responding too fast) participants were asked to confirm their response on each trial by pressing the 'm' or 'z' key for target present and absent responses respectively. This allowed participants to self-correct any motor mistakes if they realised they had pressed the wrong key by accident. This confirmed response was used to calculate final error rates for analysis. For both the training and practice trials, feedback was provided, however none was provided on the experimental trials, mimicking conditions in a clinical setting where readers receive no immediate feedback. RTs over 10,000 ms and those less than 200 ms were considered outliers and removed from data analysis.

The over-reliance effect was measured as the difference in error rates when the CAD message indicated a cancer (i.e., in Recall trials) compared to when it did not (No Recall trials). That is, when a cancer went unprompted by CAD, were participants more likely to miss it? Furthermore, on trials where no cancer was present were participants more likely to report a false alarm with a Recall message compared to a No Recall message.

Results

The outlier procedure removed 1.02%, 1.33%, 0.39% and 0.90% of all data in Experiments 1, 2, 3 and 4 respectively. Error rates for all conditions are presented in Figs. 2 and 3. In line with previous work, we were concerned with how CAD affected miss errors and false alarm rates independently from each other (Alberdi et al., 2004; Drew et al., 2020; Kunar, 2022; Kunar & Watson, 2023; Kunar et al., 2017). Thus, data were analysed accordingly throughout.

Table 2 Trial numbers of conditions for each experiment

	Number of recall trials	Number of no recall trials	Total number of trials
Cancer present trials	20	10	30
Cancer absent trials	90	180	270

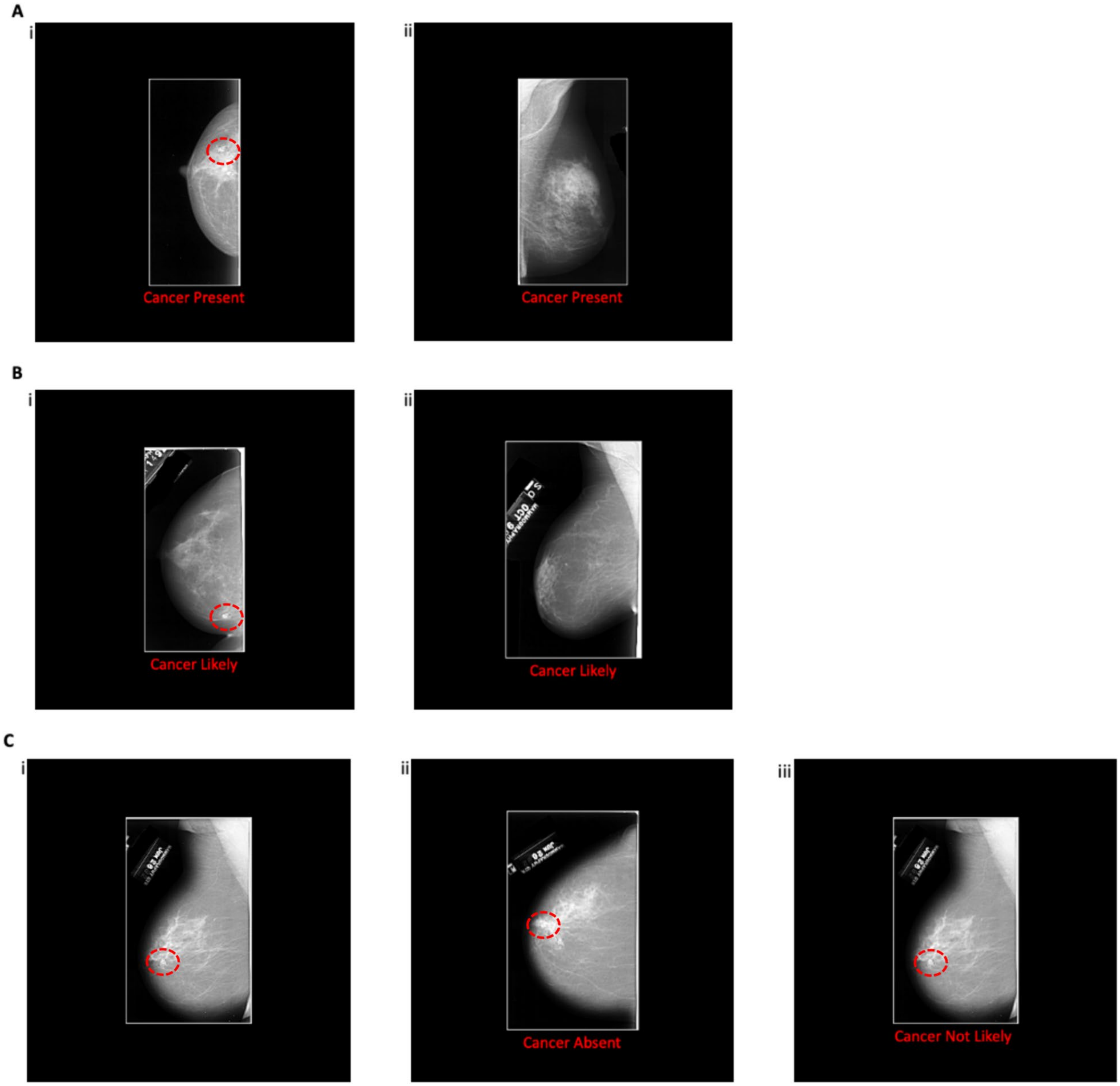

Fig. 1 Examples of the images used in the Experiments. *Note*. Examples **A**(i) and A(ii) show a Recall CAD prompt with a 'Cancer Present' message (used in Experiments 1 and 3). Example **A**(i) contains a cancer, Example **A**(ii) does not. Examples **B**(i) and **B**(ii) show a Recall CAD prompt with a 'Cancer Likely' message (used in Experiments 2 and 4). Example **B**(i) contains a cancer, Example **B**(ii) does not. Examples C show mammograms where a cancer was present, however in the Experiments there were also images where there was no cancer. Example **C**(i) shows a No Recall trial with no CAD prompt given (used in Experiment 1 and 2). Example **C**(ii) shows a No Recall CAD prompt giving a 'Cancer Absent' message (used in Experiment 3). Example **C**(iii) show a No Recall CAD prompt giving a 'Cancer Not Likely' message (used in Experiment 4). In these examples a red dotted line highlights the position of the cancer. The red dotted line did not appear in the experiment proper

Miss errors

Miss Errors were examined using a 2×2×2 repeated measures ANOVA with within participant factors of Recall (whether a CAD 'Recall' message was presented vs. 'No Recall' message) and between experiment factors of CAD Message (Present vs. Likely) and CAD Occurrence (Some vs. All trials). The results showed that there was a main effect of CAD Recall, $F(1, 148) = 18.32$, $p < 0.001$, $\eta^2_p = 0.11$. Participants missed more cancers when there was a 'No Recall' message compared to when there was a 'Recall' message, showing an over-reliance effect. There was no main effect of CAD Message, F(1,

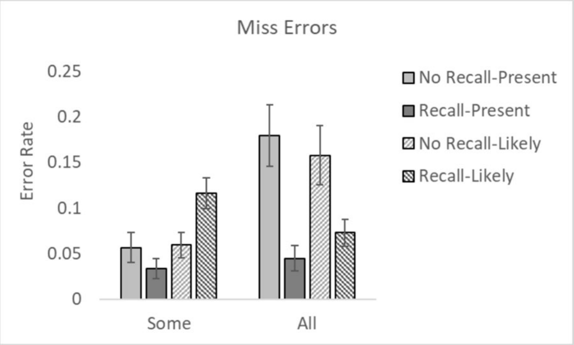

Fig. 2 Miss error rates for all conditions. *Note.* Error bars represent the standard error

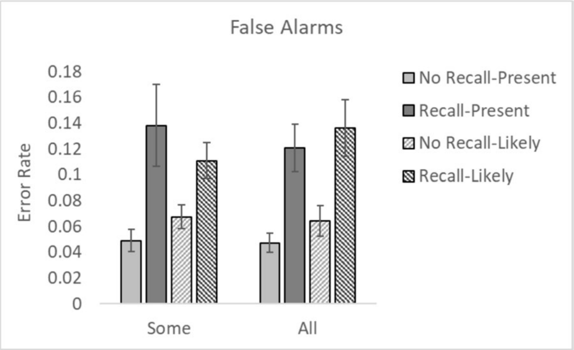

Fig. 3 False alarm rates for all conditions. *Note.* Error bars represent the standard error

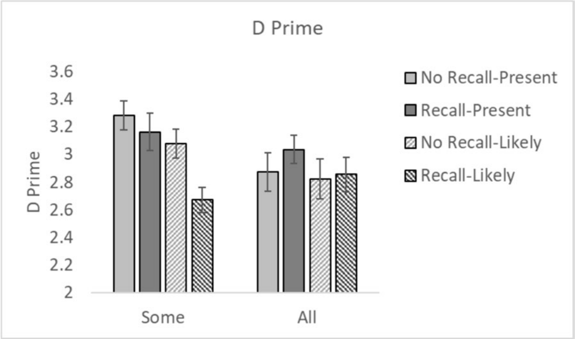

Fig. 4 D' values for all conditions. *Note.* Error bars represent the standard error

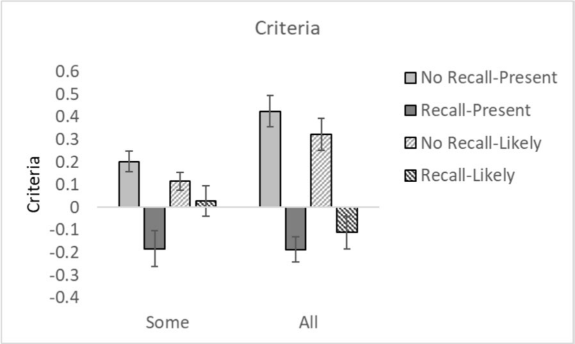

Fig. 5 C values for all conditions. *Note.* Error bars represent the standard error

148) = 1.74, $p = 0.19$, $\eta^2_p = 0.01$. However, the main effect of CAD Occurrence was significant, $F(1, 148) = 7.38$, $p = 0.007$, $\eta^2_p = 0.05$. Participants missed more cancers in conditions where CAD was shown on all trials compared to when it was only present on some of the trials. The Recall × CAD Message interaction was significant, $F(1, 148) = 9.03$, $p = 0.003$, $\eta^2_p = 0.06$, where the difference in miss errors between Recall and No Recall conditions (i.e., the 'over-reliance' effect) was greater when the CAD message said the cancer was present compared to when it was likely. The Recall × CAD Occurrence interaction was also significant, $F(1, 148) = 34.31$, $p < 0.001$, $\eta^2_p = 0.19$, where the difference in miss errors between Recall and No Recall conditions (the 'over-reliance' effect) was greater when the CAD message was present on all trials in comparison to some trials. None of the other interactions were significant (all Fs < 1.3, ps > 0.25).

False alarms

A 2×2×2 repeated measures ANOVA with within participant factors of Recall (Recall vs. No Recall) and between experiment factors of CAD Message (Present vs. Likely) and CAD Occurrence (Some vs. All trials) was conducted on False Alarms. There was a main effect of CAD Recall, $F(1, 148) = 58.75$, $p < 0.001$, $\eta^2_p = 0.28$. Participants made more false alarms when the CAD prompt indicated the presence of a cancer, compared to when it did not (or when no CAD prompt was given). There was no main effect of CAD Message, $F(1, 148) = 0.16$, $p = 0.69$, $\eta^2_p = 0.001$. Neither was there a main effect of CAD Occurrence, $F(1, 148) = 0.002$, $p = 0.96$, $\eta^2_p = 0.000$. None of the interactions were significant (all Fs < 1.8, ps > 0.19).

Signal detection theory

Signal Detection Theory (SDT, Green & Swets, 1966; Macmillan & Creelman, 2005) was used to calculate *d'* (sensitivity) and *c* (criterion) in each experiment.[2] Figures 4 and 5 shows the *d'* and *c* values, respectively.

[2] False alarm or miss error rates of 0 and 1 were adjusted using the formulas $1/2n$ and $1-(1/2n)$, where n = the number of trials (Macmillan & Kaplan, 1985, see also Russell & Kunar, 2012; Wolfe et al., 2007; Kunar et al., 2021; Kunar, 2020; Kunar & Watson, 2023, who used this procedure).

Sensitivity (*d'*)

A 2×2×2 repeated measures ANOVA with within participant factors of Recall (Recall vs. No Recall) and between experiment factors of CAD Message (Present vs. Likely) and CAD Occurrence (Some vs. All trials) was conducted on *d'*. There was no main effect of Recall, $F(1, 148) = 2.76$, $p = 0.10$, $\eta^2_p = 0.018$. Neither was there a main effect of CAD Occurrence, $F(1, 148) = 1.99$, $p = 0.16$, $\eta^2_p = 0.013$. However, there was a main effect of CAD Message, $F(1, 148) = 4.63$, $p = 0.03$, $\eta^2_p = 0.03$. D Prime was greater when the CAD message stated that a cancer was 'present' in comparison to when it was 'likely'. There was a significant Recall × CAD Message interaction, $F(1, 148) = 4.31$, $p = 0.04$, $\eta^2_p = 0.03$, in which the difference in *d'* between a Recall and No Recall CAD was greater when the CAD Message indicated a cancer was likely rather than a cancer was present. There was also a significant Recall x CAD Occurrence interaction, $F(1, 148) = 13.28$, $p < 0.001$, $\eta^2_p = 0.082$, in which the difference in *d'* between the Recall and No Recall conditions was greater when the CAD message was shown on some of the trials versus when it was shown on all of the trials. None of the other interactions were significant (all *F*s < 1.2, *p*s > 0.28).

Criteria, (*c*)

A 2×2×2 repeated measures ANOVA with within participant factors of Recall (Recall vs. No Recall) and between experiment factors of CAD Message (Present vs. Likely) and CAD Occurrence (Some vs. All trials) was conducted on *c*. There was a main effect of Recall, $F(1, 148) = 176.74$, $p < 0.001$, $\eta^2_p = 0.54$, in which participants were less willing to respond that a target was present in the No Recall condition in comparison to the Recall condition. There was no main effect of CAD Occurrence, $F(1, 148) = 1.60$, $p = 0.21$, $\eta^2_p = 0.011$. Neither was there a main effect of CAD Message, $F(1, 148) = 0.18$, $p = 0.67$, $\eta^2_p = 0.001$. There was a significant Recall × CAD Message interaction, $F(1, 148) = 17.58$, $p < 0.001$, $\eta^2_p = 0.11$, in which there was a bigger difference in response criteria between the Recall and No Recall condition when the CAD message said the cancer was present compared to when it said the target was likely. There was also a significant Recall × CAD Occurrence interaction, $F(1, 148) = 25{,}14$, $p < 0.001$, $\eta^2_p = 0.15$, in which there was a bigger difference in response criteria between the Recall and No Recall condition when the CAD message was presented on all trials compared to some of the trials. None of the other interactions were significant (all *F*s < 1.2, *p*s > 0.29).

Comparison of CAD systems: which system is best?

To compare CAD Systems across the experiments we calculated the mean overall miss errors, false alarms, the Recall Rate and the Positive Predictive Value (PPV) for Experiments 1–4 (see Figs. 6, 7 and Table 3). Recall Rate (i.e., the percentage of mammograms that were reported to have abnormal findings) and PPV (i.e., the percentage of women recalled for further tests who have cancer) are important clinical metrics within breast cancer screening (e.g., Norsuddin et al., 2015; Rauscher et al., 2021; Taylor-Phillips et al., 2024) and were calculated as follows, in which TP stands for True Positive, FP stands for False Positive (false alarms), TN stands for True Negative and FN stands for False Negative:

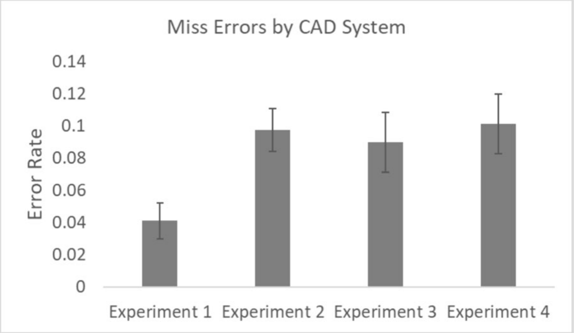

Fig. 6 Overall mean miss errors across CAD systems in experiments 1–4. *Note.* Error bars represent the standard error

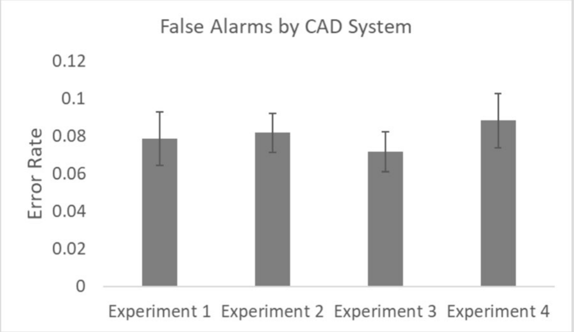

Fig. 7 Overall mean false alarms across CAD systems in experiments 1–4. *Note.* Error bars represent the standard error

Table 3 Mean recall rate and PPV across experiments

	Recall rate (%)	PPV (%)
Experiment 1	16.7 (1.30)	67.6 (3.44)
Experiment 2	16.5 (0.98)	61.0 (2.80)
Experiment 3	15.5 (0.99)	64.6 (3.03)
Experiment 4	16.9 (1.34)	61.6 (3.52)

Standard errors are shown in the parentheses

$$\text{Recall Rate} = \frac{\sum(\text{TP} + \text{FP})}{\sum(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \times 100$$

$$\text{PPV} = \frac{\sum \text{TP}}{\sum(\text{TP} + \text{FP})} \times 100$$

One way between experiment ANOVAs were used to analyse which of the CAD systems (if any) showed better performance in each of these measures. The results showed there to be a difference in the miss errors across CAD systems, $F(3, 148) = 3.22$, $p = 0.02$, $\eta^2_p = 0.061$, with the CAD system in Experiment 1 producing fewer overall miss errors compared to the other systems. However, there was no difference across CAD systems in false alarms, $F(3, 148) = 0.29$, $p = 0.83$, $\eta^2_p = 0.006$, Recall Rates, $F(3, 148) = 0.27$, $p = 0.85$, $\eta^2_p = 0.006$, or PPV,[3] $F(3, 148) = 0.68$, $p = 0.56$, $\eta^2_p = 0.014$.

General discussion

Previous research has shown that when a CAD system used exogenous cues to highlight a cancer, an over-reliance effect emerged where participants became overly dependent on the CAD cues. The current study investigated whether an over-reliance effect also occurred with CAD systems that presented binary CAD recommendations alongside the mammogram. Experiments 1 and 2 presented a CAD message on some of the trials to indicate the presence of a cancer. In Experiment 1 the message stated that a cancer was present, while in Experiment 2, the message stated that a cancer was likely. Experiments 3 and 4 presented a CAD message on every trial to indicate that a cancer was either present or absent (Experiment 3) or that a cancer was either likely or not likely (Experiment 4).

The data make several important points. First, even when using CAD as a binary system, an over-reliance effect emerged. For False Alarms, participants were more likely to (incorrectly) respond that a cancer was present when shown a CAD Recall message. For miss errors, participants were more likely to miss a cancer when there was no CAD message (Experiments 1 and 2) or when they were presented with a No Recall message (Experiments 3 and 4). This over-reliance effect was greater when a CAD message was presented on *all trials* compared to when it was only presented on *some* of the trials. Furthermore, the over-reliance effect was more pronounced when the CAD message said the cancer was 'present' compared to when it said a cancer was 'likely'. In a clinical setting, any increase in miss errors and false alarms have their own associated problems. Miss errors are obviously worrying as it means that an undiagnosed cancer will go untreated, having potentially serious health consequences for the women involved. False alarms frequently mean that recalled women undergo further tests which can be both invasive and costly (in terms of time and money) to both the women involved and the healthcare system. Furthermore, women who have been falsely recalled have been known to report feelings of psychological distress (Aro, 2000) and may delay participation or not participate at all in future screening programs (Kahn & Luce, 2003).

Please note that in Experiment 2, miss errors when a Recall message was present were higher than trials when there was no CAD message. This pattern was opposite of what would be predicted from the over-reliance effect. The reason for this was unclear. One could argue that the message 'Cancer Likely' led people to 'second guess' and dismiss the CAD cue more often compared to when the CAD cue was definitive (Cancer Present). However, this seems unlikely given that the same pattern was not observed in Experiment 4. It may be that miss errors in Experiment 2 were artificially inflated in this experiment, but as the mammogram stimuli were identical across experiments it is again unclear what would be driving this. Future research will be needed to investigate this further.

A comparison between binary CAD systems showed that while there was no difference in overall false alarms, Recall Rate or PPV across experiments, there were fewer miss errors for the CAD system tested in Experiment 1. Miss errors are an important metric that can be measured in the laboratory but cannot be easily determined in a clinical setting as by definition, a radiologist will only become aware that they have missed a cancer if the woman becomes symptomatic between routine breast screening checks. An explanation for the reduction in miss errors in Experiment 1 may be gleaned from the SDT data, which showed that participants' sensitivity to detect a target was greater when the CAD message read 'present' rather than 'likely'. Furthermore, on No Recall trials participants were less willing to commit to a response that a cancer was present when the CAD was shown on all trials, versus some of them. Based on these experiments we would recommend the binary CAD system in Experiment 1 be tested for use in a clinical setting (e.g., definitive messages presented only on mammograms where a cancer is suspected). Future research would, of course, be needed to ascertain whether the

[3] Please note that, although the PPV rates found in these experiments were similar to PPV rates found in Europe, they are higher than those typically found in the USA (which range from 15 to 30% approximately; Kopans, 1992). These differences are likely to be due to these experiments being run in a laboratory compared to a clinical setting. Further research would be needed to investigate how Recall Rates and PPV are affected by binary CAD presentation within a real-world setting.

same benefits would occur with medical readers. Despite this, there is compelling evidence to show that principles found in the lab can be applied to healthcare professionals. For example, the proportion of miss error rates in the lab are similar to radiologists reading mammograms in a clinical setting (Evans et al., 2013). Furthermore, over-reliance effects with salient CAD cues, also appear to occur with radiologists (Zheng et al., 2004). Given that similar search strategies have been found in both non-medical and medically-trained readers (Wolfe et al., 2016) we would predict that a similar benefit found in the binary CAD system of Experiment 1 would also be likely to occur to breast cancer screening in the real world.

Please also note that the over-reliance effect observed in these binary CAD systems may be affected by radiologist experience. It has been suggested that radiologists with more experience tend to interact with CAD less than those with less experience (Hupse et al., 2013). Goldenberg and Peled (2011) suggested that for binary CAD systems, outcomes that are positive in identifying a disease should be verified by more experienced readers, while those with a negative outcome could be considered by less experienced staff (particularly if CAD systems were used as a way to triage patients). However, we would suggest that for mammography, dividing cases based on staff experience would not result in best practice. Triaging acute medical conditions based on CAD outputs would be important in emergency situations where diagnosis is time critical as urgent cases could be treated by an experienced physician in a timely manner (Goldenberg & Peled, 2011). However, in cases where the disease is chronic, such as with cancer screening, the benefit of separating CAD outputs via reader experience would be negligible and possibly damaging. That is, if less experienced clinicians were more dependent on CAD, they would be more likely to miss a cancer if it was not flagged by the CAD prompt. Thus, consideration of how to allocate CAD outputs across readers with different experience needs to be considered by future research.

It is also worth noting that mammogram reading procedures differ globally. Double reading mammogram procedures are considered standard practice in the UK, (Chen et al., 2023) and across most other European countries (Balta et al., 2020), whereas single reading is the more common practice in the USA. The use of CAD in mammogram screening across different countries will therefore be affected by current practices and regulations in different parts of the world. Nevertheless, as double reading is considered labour intensive and there is continued concern about the number of radiologists currently available (Chen et al., 2023) there has been increased interest as to whether AI and automated aids can be feasibly used as a 'second reader' to help workflow within mammographic screening (Geras et al., 2019; Rodriguez-Ruiz et al., 2019). Data from experiments such as these can help inform clinicians of the optimal way to present AI prompts to readers.

There are, of course, limits to the conclusions that we can make based on this study given that the experimental procedure is very different to clinical procedures in mammography. First, in our experiments participants were only shown one mammogram image at a time with no control over how the image was presented. This is very different from normal clinical practice in which radiologists have a custom hanging protocol for how images are displayed. Furthermore, radiologists are able to view images from prior mammograms for comparison if they chose to whereas, participants were not offered that option in our experiments. This may have affected participant's over-reliance on the CAD prompt, as they had less control of what they were presented. Second, in Experiments 3 and 4 a CAD message was shown on all trials. For present purposes, we have compared errors in the Recall vs. No Recall conditions to measure the over-reliance effect. However, in future work it would be good to compare these data to a baseline condition in which No CAD message was presented. Third, in a clinical setting the prevalence rate of a cancer is much lower than the 10% prevalence rates exhibited in our experiments. Wolfe et al. (2005) have found that miss errors increase as the prevalence of the target decreases. Furthermore, search strategies change under low prevalence conditions. Wolfe and Van Wert (2010) proposed a Multiple-Decision Model stating that at low prevalence, the time spent searching a display before terminating search is decreased and there is also a shift in response bias so that people require more evidence before committing to a target present response. Further research would be needed to determine if the effects found in the current study also apply to a clinical setting when searching for a cancer with a lower prevalence rate.

CAD use in mammography has the potential to help save lives. The rise of AI in breast screening is promising in terms of helping readers with the mammography task and helping with the workload in healthcare systems. Of course, along with the visual output of CAD recommendations it is also important to consider where to add AI into the healthcare workflow. Having AI be fully automated and used as a standalone reader to make a clinical recommendation would be beneficial in relieving the workload for healthcare professionals (Raya-Povedano et al., 2021). However, it has been found participants feel more comfortable with AI acting as an additional reader in the workflow, acting alongside a human rather than it being used to make an independent clinical diagnosis outside of human input (Ng et al., 2023; Ongena et al., 2021). Ng et al. (2023) have suggested that adding AI into the workflow as a *supporting* reader may be the best approach to capture the benefits of CAD, while assuaging the concerns of patients. Here, the AI acts as a second reader. However, in situations where there is a disagreement between the human reader and the AI then the case is deferred to another human reader for assessment. Given the range of possibilities of presenting CAD and how the different strategies affect clinical outcomes it is clear that ever-more research is needed to investigate how humans optimally interact with CAD systems.

Conclusion

Previous research has found that over-trust in computer-aided systems can lead users to make diagnostic errors (Jorritsma et al., 2015). The present data add to the narrative that the way we present CAD to humans affects how we interact with it. Although Goldenberg and Peled (2011) hoped that binary CAD would be an advancement of traditional CAD systems, our work has shown that binary CAD systems are still susceptible to over-reliance behaviours. Showing that a similar effect also occurs in a clinical field will be important to establish going forward with the increased exploration of how AI and other automated decision systems will be rolled out across healthcare services.

Abbreviations

LP	Low prevalence
CAD	Computer Aided Detection
ANOVA	Analysis of variance
RTs	Reaction Times
USA	United States of America
AI	Artificial Intelligence
DDSM	Digital Database for Screening Mammography
RCT	Randomised Clinical Trial

Acknowledgements

For the purpose of open access, the author has applied a Creative Commons Attribution (CC-BY) licence to any Author Accepted Manuscript version arising from this submission.

Significance statement

The use of automation and Computer Aided Detection (CAD) systems to help radiologists find cancers in mammograms has the potential to save lives. Recent developments in Artificial Intelligence have been used to improve the efficacy of CAD systems with the aim of easing the workload of healthcare professionals. However, research is also needed to investigate the best ways for humans to interact with these systems. The current research investigated a CAD system that gave a binary recommendation to alert the reader to possible cancer presence. The results show that the way this system was presented to people affected how they interacted and depended on it. The results are important in determining how humans optimally interact with automated systems for medical screening purposes.

Author contributions

Frankie Patterson was responsible for designing and programming the experiments, data collection, analysis of data and writing up the results into manuscript form. Melina Kunar was responsible for designing and programming the experiments, analysis of data and writing up the results into manuscript form.

Funding

This work was not part of a funded grant.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the Open Science Framework repository, https://osf.io/jgxz5/.

Declarations

Ethics approval and consent to participate

Full ethical approval for this study was granted by the Department of Psychology Ethics Committee of the University of Warwick. All participants provided informed consent prior to completing the experiment.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 16 August 2023 Accepted: 9 July 2024
Published online: 02 September 2024

References

Alberdi, E., Povyakalo, A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology, 11*(8), 909–918.

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.

Balta, C., Rodriguez-Ruiz, A., Mieskes, C., Karssemeijer, N., & Heywang-Köbrunner, S. H. (2020). Going from double to single reading for screening exams labeled as likely normal by AI: What is the impact?. In *15th international workshop on breast imaging (IWBI2020)* (Vol. 11513, pp. 94–101). SPIE.

Bennett, R. L., Blanks, R. G., & Moss, S. M. (2006). Does the accuracy of single reading with CAD (computer-aided detection) compare with that of double reading? A review of the literature. *Clinical Radiology, 61*(12), 1023–1028.

Bird, R. E., Wallace, T. W., & Yankaskas, B. C. (1992). Analysis of cancers missed at screening mammography. *Radiology, 184*(3), 613–617.

Castellino, R. A. (2005). Computer aided detection (CAD): An overview. *Cancer Imaging, 5*(1), 17.

Chen, Y., James, J. J., Michalopoulou, E., Darker, I. T., & Jenkins, J. (2023). Performance of radiologists and radiographers in double reading mammograms: The UK National Health Service breast screening program. *Radiology, 306*(1), 102–109.

Cunningham, C. A., Drew, T., & Wolfe, J. M. (2017). Analog Computer-Aided Detection (CAD) information can be more effective than binary marks. *Attention, Perception, & Psychophysics, 79*, 679–690.

Darzi, A., & Evans, T. (2016). The global shortage of health workers: An opportunity to transform care. *The Lancet, 388*(10060), 2576–2577.

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*(10), 1260–1267.

Drew, T., Guthrie, J., & Reback, I. (2020). Worse in real life: An eye-tracking examination of the cost of CAD at low prevalence. *Journal of Experimental Psychology: Applied, 26*(4), 659–670.

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., Haygood, T. M., Cooper, J., Culpan, A. M., & Wolfe, J. M. (2016). A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast. *Proceedings of the National Academy of Sciences of the United States of America, 113*, 10292–10297.

Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E., & Elmore, J. G. (2007). Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine, 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.

Fujita, H. (2020). AI-based computer-aided diagnosis (AI-CAD): The latest review to read first. *Radiological Physics and Technology, 13*(1), 6–19.

Geras, K. J., Mann, R. M., & Moy, L. (2019). Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives. *Radiology, 293*(2), 246–259.

Godwin, H. J., Menneer, T., Riggs, C. A., Cave, K. R., & Donnelly, N. (2015). Perceptual failures in the selection and identification of low-prevalence targets in relative prevalence visual search. *Attention, Perception, & Psychophysics, 77*, 150–159.

Goldenberg, R., & Peled, N. (2011). Computer-aided simple triage. *International Journal of Computer Assisted Radiology and Surgery, 6*(5), 705–711.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1, pp. 1969–2012). New York: Wiley.

Guerriero, C., Gillan, M. G., Cairns, J., Wallis, M. G., & Gilbert, F. J. (2011). Is computer aided detection (CAD) cost effective in screening mammography? A model based on the CADET II study. *BMC Health Services Research, 11*(1), 1–9.

Heath, M., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, P. (2001). The digital database for screening mammography, IWDM-2000. In *Fifth international workshop on digital mammography* (pp. 212–218). Medical Physics Publishing.

Henriksen, E. L., Carlsen, J. F., Vejborg, I. M., Nielsen, M. B., & Lauridsen, C. A. (2019). The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: A systematic review. *Acta Radiologica, 60*(1), 13–18.

Hout, M. C., Walenchok, S. C., Goldinger, S. D., & Wolfe, J. M. (2015). Failures of perception in the low-prevalence effect: Evidence from active and passive visual search. *Journal of Experimental Psychology: Human Perception and Performance, 41*(4), 977.

Hupse, R., Samulski, M., Lobbes, M. B., Mann, R. M., Mus, R., den Heeten, G. J., Beijerinck, D., Pijnappel, R. M., Boetes, C., & Karssemeijer, N. (2013). Computer-aided detection of masses at mammography: Interactive decision support versus prompts. *Radiology, 266*, 123–129.

James, J. J., Gilbert, F. J., Wallis, M. G., Gillan, M. G., Astley, S. M., Boggis, C. R., Agbaje, O. F., Brentnall, A. R., & Duffy, S. W. (2010). Mammographic features of breast cancers at single reading with computer-aided detection and at double reading in a large multicenter prospective trial of computer-aided detection: CADET II. *Radiology, 256*(2), 379–386.

JASP Team. (2021). JASP (Version 0.16) [Computer software].

Jorritsma, W., Cnossen, F., & van Ooijen, P. M. (2015). Improving the radiologist–CAD interaction: Designing for appropriate trust. *Clinical Radiology, 70*(2), 115–122.

Kahn, B. E., & Luce, M. F. (2003). Understanding high-stakes consumer decisions: Mammography adherence following false-alarm test results. *Marketing Science, 22*(3), 393–410.

Konstantinidis, K. (2023). The shortage of radiographers: A global crisis in healthcare. *Journal of Medical Imaging and Radiation Sciences, 55*, 101333.

Kopans, D. B. (1992). The positive predictive value of mammography. *American Journal of Roentgenology, 158*(3), 521–526.

Kunar, M. A. (2022). The optimal use of computer aided detection to find low prevalence cancers. *Cognitive Research: Principles and Implications, 7*(1), 1–18.

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., & Watson, D. G. (2023). Framing the fallibility of Computer-Aided Detection aids cancer detection. *Cognitive Research: Principles and Implications, 8*(1), 30.

Kunar, M. A., Watson, D. G., & Taylor-Phillips, S. (2021). Double reading reduces miss errors in low prevalence search. *Journal of Experimental Psychology: Applied, 27*(1), 84.

Kunar, M. A., Watson, D. G., Taylor-Phillips, S., & Wolska, J. (2017). Low prevalence search for cancers in mammograms: Evidence using laboratory experiments and computer aided detection. *Journal of Experimental Psychology: Applied, 23*(4), 369.

Lehman, C. D., Wellman, R. D., Buist, D. S., Kerlikowske, K., Tosteson, A. N., Miglioretti, D. L., Breast Cancer Surveillance Consortium. (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine, 175*(11), 1828–1837.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide*, 2nd edn New York. NY: Lawrence Erlbaum Associates Publishers.

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*(1), 185.

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.

Ng, A. Y., Glocker, B., Oberije, C., Fox, G., Sharma, N., James, J. J., Ambrózay, É., Nash, J., Karpati, E., Kerruish, S., & Kecskemethy, P. D. (2023). Artificial intelligence as supporting reader in breast screening: A novel workflow to preserve quality and reduce workload. *Journal of Breast Imaging, 5*(3), 267–276.

Norsuddin, N. M., Reed, W., Mello-Thoms, C., & Lewis, S. J. (2015). Understanding recall rates in screening mammography: A conceptual framework review of the literature. *Radiography, 21*(4), 334–341.

Obenauer, S., Sohns, C., Werner, C., & Grabbe, E. (2006). Impact of breast density on computer-aided detection in full-field digital mammography. *Journal of digital imaging, 19*, 258–263.

Ongena, Y. P., Yakar, D., Haan, M., & Kwee, T. C. (2021). Artificial intelligence in screening mammography: A population survey of women's preferences. *Journal of the American College of Radiology, 18*(1), 79–86.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-01193-y

Raat, E. M., Kyle-Davidson, C., & Evans, K. K. (2023). Using global feedback to induce learning of gist of abnormality in mammograms. *Cognitive Research: Principles and Implications, 8*(1), 1–22.

Rauscher, G. H., Murphy, A. M., Qiu, Q., Dolecek, T. A., Tossas, K., Liu, Y., & Alsheik, N. H. (2021). The "sweet spot" revisited: Optimal recall rates for cancer detection with 2D and 3D digital screening mammography in the Metro Chicago Breast Cancer Registry. *American Journal of Roentgenology, 216*(4), 894–902.

Raya-Povedano, J. L., Romero-Martín, S., Elías-Cabot, E., Gubern-Mérida, A., Rodríguez-Ruiz, A., & Álvarez-Benito, M. (2021). AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: A retrospective evaluation. *Radiology, 300*(1), 57–65.

Remington, R. W., Johnston, J. C., & Yantis, S. (1992). Involuntary attentional capture by abrupt onsets. *Perception & Psychophysics, 51*, 279–290.

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), 15–15.

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Teuwen, J., Broeders, M., Gennaro, G., Clauser, P., Helbich, T. H., Chevalier, M., Mertelmeier, T., & Wallis, M. G. (2019). Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *European Radiology, 29*(9), 4825–4832.

Russell, N. C., & Kunar, M. A. (2012). Colour and spatial cueing in low-prevalence visual search. *Quarterly Journal of Experimental Psychology, 65*(7), 1327–1344.

Salim Jr, A., Allen, M., Mariki, K., Masoy, K. J., & Liana, J. (2023). Understanding how the use of AI decision support tools affect critical thinking and over-reliance on technology by drug dispensers in Tanzania. arXiv preprint arXiv:2302.09487

Salim, M., Wåhlin, E., Dembrower, K., Azavedo, E., Foukakis, T., Liu, Y., Smith, K., Eklund, M., & Strand, F. (2020). External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. *JAMA Oncology, 6*(10), 1581–1588.

Soo, M. S., Rosen, E. L., Xia, J. Q., Ghate, S., & Baker, J. A. (2005). Computer-aided detection of amorphous calcifications. *American Journal of Roentgenology, 184*(3), 887–892.

Taplin, S. H., Rutter, C. M., & Lehman, C. D. (2006). Testing the effect of computer-assisted detection on interpretive performance in screening mammography. *American Journal of Roentgenology, 187*(6), 1475–1482.

Taylor, P., & Potts, H. W. (2008). Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. *European Journal of Cancer, 44*(6), 798–807.

Taylor-Phillips, S., Jenkinson, D., Stinton, C., Kunar, M. A., Watson, D. G., Freeman, K., Mansbridge, A., Wallis, M. G., Kearins, O., Hudson, S., &

Clarke, A. (2024). Fatigue and vigilance in medical experts detecting breast cancer. *Proceedings of the National Academy of Sciences, 121*(11), e2309576121.

Theeuwes, J. (2004). Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review, 11*, 65–70.

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.

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review, 28*(4), 1060–1092.

Wolfe, J. M., Evans, K. K., Drew, T., Aizenman, A., & Josephs, E. (2016). How do radiologists use the human search engine? *Radiation Protection Dosimetry, 169*(1–4), 24–31.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature, 435*(7041), 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., & Van Wert, M. J. (2010). Varying target prevalence reveals two, dissociable decision criteria in visual search. *Current Biology, 20*, 121–124.

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2023). Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence, 316*, 103839.

Zheng, B., Swensson, R. G., Golla, S., Hakim, C. M., Shah, R., Wallace, L., & Gur, D. (2004). Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments1. *Academic Radiology, 11*(4), 398–406.

Publisher's Note