Increasing Transparency of Computer Aided Detection Impairs Decision Making in Visual Search

Melina A. Kunar[1], Giovanni Montana[2] and Derrick G. Watson[1]

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

Corresponding Author: Melina Kunar

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

Tel: +44 (0)2476 522133

Running Title: Transparency of CAD Impairs Decision Making

Abstract

Recent developments in Artificial Intelligence (AI) have led to changes in healthcare. Government and regulatory bodies have advocated the need for transparency in AI systems with recommendations to provide users with more details about AI accuracy and how AI systems work. However, increased transparency could lead to negative outcomes if humans become over-reliant on the technology. This study investigated how changes in AI transparency affected human decision making in a medical screening visual search task. Transparency was manipulated by either giving or withholding knowledge about the accuracy of an 'AI system'. We tested performance in seven simulated lab mammography tasks, in which observers searched for a cancer which could be correctly or incorrectly flagged by Computer Aided Detection (CAD) 'AI prompts'. Across tasks the CAD systems varied in accuracy. In the 'Transparent' conditions participants were told the accuracy of the CAD system, in the 'Not Transparent' they were not. The results showed that increasing CAD transparency impaired task performance producing an increase in false alarms, decreased sensitivity, an increase in Recall Rate and a decrease in Positive Predictive Value. Along with increasing investment in AI this research shows that it is important to investigate how transparency of AI systems affect human decision making. Increased transparency may lead to over-trust in AI systems, which can impact clinical outcomes.

Key words: Artificial Intelligence, Computer Aided Detection (CAD), Transparency, Low Prevalence, Over-reliance, Visual Search.

Introduction

In recent years developments in Artificial Intelligence (AI) have led to important changes in healthcare (Kerasidou et al., 2022). It has been proposed that AI may help with tasks such as workflow and clinical administration (Mello-Thoms & Mello, 2023), the optimisation of clinical trials (Askin et al., 2023), and medical imaging with predictions that AI use in this area will grow dramatically in future years (Allen et al., 2021)

In medical screening, Computer Aided Detection (CAD) uses computer algorithms to alert readers to the presence of suspicious entities, such as cancers in mammograms. Historically, there have been conflicting results on the benefit of this technology. Although CAD was approved by the Food and Drugs Administration (FDA) in the late 1990s and rolled out at great financial cost (estimated at over $400 million per annum, Lehman et al., 2015) little research was conducted to examine how these automatic aids affected human decision making. Subsequent studies showed mixed benefits in relation to CAD use - some positive with increased detection of early-stage malignancies (Freer and Ulissey, 2001), whereas other studies showed little benefit (Lehman et al., 2015). Furthermore, some studies showed harmful results, where CAD led to reduced accuracy of mammogram interpretation (Fenton et al., 2007) and clinicians missed more cancers if not marked by CAD (Zheng et al., 2004; Taplin et al., 2006).

Recent advancements in AI, however, have shown promising results in CAD use. AI acting as a supporting reader produced similar performance to double reading procedures (where two human readers read the same mammogram) while also reducing the number of cases that humans had to read (McKinney, et al., 2020; Ng et al., 2023). Given that there is a shortage of

available healthcare workers (e.g., Konstantinidis, 2023) the use of AI as a second reader in tasks such as mammography could provide significant benefits.

Despite the potential advantages of AI use in medical screening there are also disadvantages. Although, the FDA has already approved several AI systems (Benjamens, et al., 2020), the clinical and cognitive costs of human interaction with these systems are still not fully understood and Human-AI interactions remain largely under-researched. Furthermore, there is considerable evidence showing that people can become over-reliant on this technology (e.g., Bucina et al., 2021; Bussone et al., 2015; Jacobs et al., 2021; Kunar et al., 2017). This is particularly problematic when the CAD systems either fail to flag a cancer or incorrectly predict the presence of a cancer when there is not one (Kunar et al., 2017). In the former case, cancers that are not prompted by CAD, are more likely to go unnoticed, meaning that women will not receive appropriate and timely medical care. The latter means that women will be needlessly recalled for further tests, that can be worrying and also add an extra and unnecessary burden to healthcare systems (Aro, 2000). Other work has found ways to keep the benefits of CAD while mitigating the costs, for example by changing the way AI prompts are presented to readers (Kunar, 2022; Patterson & Kunar, 2024) or how the CAD systems are framed (Kunar & Watson, 2023). Crucially, these studies have shown that what people know about AI systems affect how they are used and their influence on decision making. Given that there is growing investment and support to integrate AI into medical screening (Alexander et al., 2024) it is critical that we examine how such systems influence human decision making.

With this growth in AI development, government and regulatory bodies have stipulated the importance of transparency within AI systems (Gov, U.K., 2023; Kerasidou et al., 2022; Kingsman et al., 2022). AI transparency can refer to making sure humans are aware of how AI

systems operate and clarity in their accuracy. Transparency can be achieved through a number of ways. Kiseleva et al., (2022) suggested that factors such as information about an AI system and its interpretability can affect transparency. Lekadir et al. (2021) proposed a set of guidelines for AI use in medical screening, which include that readers should be informed about the errors that AI systems make (for example, by showing uncertainty estimates) to increase user trust and usability. However, increasing transparency in AI is often complex and can sometimes lead to negative outcomes if people become too dependent on the system (Bucinca et al., 2021). If increased transparency leads to over-trust in the technology this could have significant consequences for clinical outcomes. For example, increasing transparency about CAD may lead to clinicians accepting the CAD recommendation even if it disagrees with their initial judgement.

We investigated whether increasing transparency of CAD affects search performance in a simulated mammogram task where participants searched for a cancer using seven different CAD conditions. Previous work has shown that laboratory experiments are a reliable way to investigate human reliance on CAD systems (see Kunar et al., 2017, for full details). For example, lab testing allows use of designs that are often not practical, or ethical, in clinical settings, due to the shortage of radiologists and the expense of Randomised Control Trials (RCTs). Furthermore, lab studies allow measures such as miss errors to be observed, which are difficult to determine in a clinical setting given that, by definition, radiologists will be unaware that they have missed a cancer. We can use this measure, along with False Alarms and other breast screening metrics such as Recall Rate (the percentage of mammograms that were reported to have abnormal findings) and Positive Predictive Value (PPV, the percentage of women recalled for further tests who have cancer) to investigate the effects of transparency on decision making with CAD.

We manipulated transparency across a range of CAD systems that differed in their accuracy to predict a cancer[1]. CAD transparency was manipulated by explicitly telling people the accuracy of the CAD system before use. In Conditions 1-3 participants were asked to interact with a non-transparent CAD system which either accurately predicted the target on 33% of trials, 67% of trials or 83% of trials. In these conditions participants were not told the accuracy of the CAD algorithm. In Conditions 4-6, participants were shown these same systems but explicitly told in advance the CAD accuracy to make performance more transparent. A final seventh condition acted as a No CAD control, to allow comparison to a baseline where CAD was never used. To preview the results, although making the systems more transparent did not affect miss errors, more transparent systems led to a decrease in performance in relation to false alarms, sensitivity (as measured by *d'*), Recall Rate and PPV.

General Method

Transparency and Openness

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

[1] In a clinical setting the accuracy of CAD is thought to be in the range of 57 – 85% but this can vary across systems (Soo et al., 2005; Obenauer et al., 2006). In these experiments a CAD accuracy range of 33% to 83% was used to test whether there were differential effects of transparency on CAD systems that had low or high accuracy rates.

Participants

Six-hundred and forty-five participants were tested, in which one hundred participants took part in Conditions 1-6 and 44 participants took part in Condition 7[2]. A G*Power calculation determined that this number of participants resulted in an experimental power above of 0.95, (the minimum number of participants needed per condition to achieve this power was 42, F-tests, fixed effects and interactions, alpha = 0.05, effect size = 0.25). Participants were aged 18 or above, recruited via Prolific, and were only able to take part in one experiment. Ethical approval for all studies was granted by the Humanities and Social Sciences Research Ethics Committee at the University of Warwick.

Stimuli and Procedure:

Seven conditions were used to examine performance with different CAD systems. The conditions varied by CAD accuracy and whether people were informed of this. In all conditions, participants were asked to search and respond to a mass presented on a mammogram. Two hundred mammogram images were obtained from the Digital Database for Screening Mammography (DDSM, Heath et al., 2001, 1998). All original images were randomly selected from the image group that had been confirmed to be cancer free. One hundred and eighty-eight of these images made up the target absent trials (where there was no cancer). The other 12 mammogram images were edited so that they contained a 'cancer'. To do this, four cancers were chosen at random from cancer cases on the DDSM. For each 'target present' image one of these cancers was transposed onto a mammogram that previously contained no cancer. Across the experiment all four cancers were transposed onto the mammograms equally often with the premise that only one cancer would appear on each

[2] 101 participants took part in condition 2 due to an error with Prolific – for ethical reasons we chose to analyse all data sets collected.

mammogram. Four cancers were chosen so that there would be a degree of perceptual variability across trials increasing the complexity of search, but the target would remain identifiable by medically naïve readers. The mass could appear on any area of the breast tissue, chosen at random (mimicking conditions in a clinical setting), provided that it was clearly distinguishable once fixated. These stimuli and procedure were chosen as they have previously shown to be successful to use in search tasks with participants who have no formal medical training (e.g., Kunar et al., 2017, Kunar et al., 2021; Kunar, 2022; Kunar & Watson, 2023; Patterson & Kunar, 2024). At the beginning of the experiment participants were given a training session, where they were shown example images of the mass and mammograms and asked to detect a mass in a mammogram. They could only continue to the experiment proper, once they had passed the training phase in which they had to correctly respond to whether a mammogram contained a cancer or not at a level above 70% accuracy. If participants failed the training session, they repeated it until their accuracy was above 70%. If participants did not reach this level after their fourth attempt, they were still allowed to proceed, however their data were removed from analysis. However, all participants successfully completed the training phase by this point.

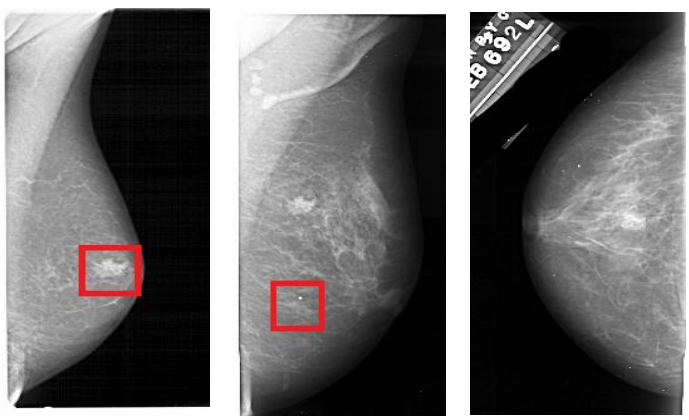
For each condition, the prevalence rate of the cancer was 6%. Conditions 1-6 had participants search for the cancer with the use of a 'CAD system' while Condition 7 acted as a baseline where no 'CAD system' was used. In Conditions 1-6, participants were informed that they may be shown a CAD prompt in the form of a red box. They were also informed that the CAD cue could be accurate and highlight the target item (Correct CAD), but sometimes it could highlight a non-cancerous area even when a cancer was present (Incorrect CAD) or could contain a cancer which was not flagged by CAD (No CAD). For target absent trials, there would either be an Incorrect CAD cue, which would highlight an area that did not contain a cancer, or no

CAD cue would be presented. For target absent trials, the lack of CAD would correctly indicate there was no cancer in the display. Example displays are shown in Figure 1. Across different experiments, the accuracy of CAD's ability to correctly highlight the location of the cancer varied, so that it correctly identified a cancer on either 33%, 67% or 83% of times. Table 1 shows the accuracy rate and number of trials, that either correctly or incorrectly contained a CAD cue, for each condition. All trials were presented in a randomly generated order for each participant. In Conditions 1-3 participants were given no explicit knowledge of the CAD accuracy rate. In Conditions 4-6, participants were explicitly told how accurate the CAD system was (e.g., "In this session the CAD cue (red box) will highlight the cancer 83% of the time"). In Condition 7 participants were asked to search for cancers without the use of a 'CAD system'. In this Condition, participants responded to whether a mammogram contained a cancer or not. None of the mammogram images contained a CAD cue and CAD was not mentioned in the instructions.

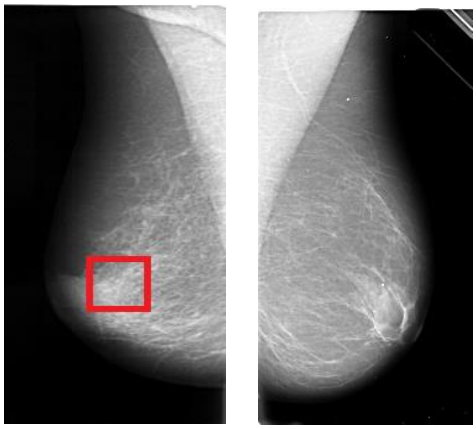
Figure 1.

Examples of mammogram displays with Correct CAD, Incorrect CAD and No CAD for cancer present trials and Incorrect CAD and No CAD (correct) for cancer absent conditions.

<u>Cancer Present Trials - Correct CAD, Incorrect CAD and No CAD</u>



Cancer Absent Trials – Incorrect CAD and No CAD



/

Table 1

Summary of Accuracy Rates and Trial Numbers for each Condition

Condition	Overall CAD Accuracy Rate	Transparent Knowledge of CAD accuracy	Absent trials with CAD	Absent trials with No CAD	Present trials with Correct CAD	Present trials with Incorrect CAD	Present trials with No CAD
1	33%	No	47	141	4	4	4
2	67%	No	47	141	8	2	2
3	83%	No	47	141	10	1	1
4	33%	Yes	47	141	4	4	4
5	67%	Yes	47	141	8	2	2
6	83%	Yes	47	141	10	1	1
7	n/a	n/a	0	188	0	0	12

Note. CAD Accuracy refers to accurate detection of a cancer when it is present. Condition 7 acted as a No CAD control.

For each trial, participants were asked to respond whether there was a cancer in the mammogram image. If they, believed there was a cancer they pressed the key 'm'. If they believed there was no cancer they pressed with the key 'z'. To ensure that the results were not affected by motor errors (Fleck and Mitroff, 2007), participants had to respond a second time, to confirm their response. This was done by again pressing the 'm' key for target present responses and 'z' key for target absent responses. This ensured that participants could correct their initial response if they accidently pressed the wrong button. Participants were given a short practice session before the start of the experiment.

Data Analysis

Incomplete data sets were removed from analyses. This led to the removal of six participants (one participant in Conditions 1, 3, 6 and 7, two participants in Condition 4). To avoid motor errors, the confirmed responses were used to calculate miss errors and false alarms. If performance was negatively affected by transparency we would expect to see a greater proportion of miss errors and/or a greater proportion of false alarms in the Transparent versus the Not Transparent CAD conditions.

To understand the reason for any differences in error rates across experiment we examined how Sensitivity (as measured by *d'*) and Response Bias (measured by *c*) changed across CAD Systems using Signal Detection Theory (SDT, Green and Swets, 1967, Macmillan & Creelman, 2005). If performance was negatively affected by transparency we would expect to see a decrease in *d'* in the Transparent compared to the Not Transparent CAD conditions. A change

in criteria across transparency would also suggest a shift in response bias, with a higher criteria reflecting that participants were less willing to respond that a cancer was present.

The data were also analysed to see how transparency affected Recall Rates and PPV. Recall Rate and PPV 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:

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

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

If adding transparency affected performance negatively, we would expect to see higher recall rates and lower PPV in the Transparent conditions compared to the Not Transparent conditions.

For each of these metrics we conducted a 3 x 2 ANOVA with between factors of CAD Accuracy (33%, 67% and 83%) and Transparency (Transparent vs Not Transparent). Significant interactions were further analysed using planned t-tests, where we also include Bayesian analyses, as supportive evidence (Wagenmakers et al., 2018a). We only include Bayesian analysis for these planned t-tests rather than ANOVAs as the latter is still an ongoing topic of research (Wagenmakers et al., 2018b). For our Bayesian analyses, we adopt the recommendations of Jeffreys (1961) in which a BF_{10} of 1 to 3 provides *anecdotal* evidence for the alternative, a BF_{10} of 3 to 10 provides *substantial* evidence for the alternative, a BF_{10} of 10

to 30 provides *strong* evidence for the alternative, a BF_{10} of 30 to 100 provides *very strong* evidence for the alternative and a BF_{10} of greater than 100 provides *decisive* evidence for the alternative. The inverse of these numbers (BF_{01}) provide evidence in support the null hypothesis (Jarosz & Wiley, 2014).

Lastly, data from each CAD Condition (1 – 6) were compared to the No CAD control (Condition 7). This enabled us to determine whether there was an overall benefit or cost of CAD, in relation to when no CAD system was used. For each metric six t-tests were run. To compensate for multiple comparisons, we used the Bonferroni correction for with the adjusted alpha levels of 0.008 per test (.05/6).

Results

Figure 2 shows the data for all conditions.

Miss Errors

For miss errors, the 3 x 2 ANOVA revealed a main effect of CAD Accuracy, $F(2, 590) = 22.64$, $p < .001$, $\eta_p^2 = 0.07$, in which miss errors decreased with increasing CAD accuracy. Neither the main effect of Transparency, $F(1, 590) = 0.02$, $p = .90$, $\eta_p^2 < 0.001$ nor the CAD Accuracy x Transparency interaction were significant, $F(2, 590) = 0.005$, $p = .995$, $\eta_p^2 < 0.001$.

Comparisons of individual conditions with the No CAD Control showed that there were no significant differences in miss errors (see Table 2 for details of all comparisons).

Figure 2 *Mean Values Across Conditions*

Note. Error bars represent the standard error.

Table 2

Comparisons of each CAD Condition with the No CAD Control

Condition Compared with the No CAD Control	Metric	CAD Accuracy Rate	Transparent Knowledge of CAD accuracy	t	df	p
1	Miss Errors	33%	Not Transparent	1.57	142	.13
2		67%	Not Transparent	0.10	140	.92
3		83%	Not Transparent	2.17	140	.03
4		33%	Transparent	1.42	141	.16
5		67%	Transparent	0.12	139	.90
6		83%	Transparent	2.17	140	.03
1	False Alarms	33%	Not Transparent	2.84	142	.005**
2		67%	Not Transparent	2.51	140	.01
3		83%	Not Transparent	1.54	140	.13
4		33%	Transparent	0.77	141	.45
5		67%	Transparent	0.19	139	.85
6		83%	Transparent	1.74	140	.09
1	D Prime	33%	Not Transparent	1.35	142	.18
2		67%	Not Transparent	2.47	140	.015
3		83%	Not Transparent	3.15	140	.002**
4		33%	Transparent	0.39	141	.70
5		67%	Transparent	0.19	139	.85
6		83%	Transparent	3.02	140	.003**
1	Criteria	33%	Not Transparent	4.02	142	< .001**
2		67%	Not Transparent	2.85	140	.005**
3		83%	Not Transparent	1.40	140	.16
4		33%	Transparent	2.45	141	.02
5		67%	Transparent	0.29	139	.77
6		83%	Transparent	1.30	140	.20
1	Recall Rate	33%	Not Transparent	3.03	142	.003**
2		67%	Not Transparent	2.51	140	.01
3		83%	Not Transparent	1.41	140	.16
4		33%	Transparent	0.89	141	.37
5		67%	Transparent	0.20	139	.85
6		83%	Transparent	1.60	140	.11
1	PPV	33%	Not Transparent	2.86	142	.005**
2		67%	Not Transparent	3.07	140	.003**
3		83%	Not Transparent	2.66	140	.009
4		33%	Transparent	1.37	141	.17
5		67%	Transparent	0.28	139	.78
6		83%	Transparent	2.45	140	.02

Note. P-values labelled as ** are significant using the adjusted Bonferroni correction alpha level of 0.008 per test (.05/6).

False Alarms

For false alarms, the 3 x 2 ANOVA revealed no main effect of CAD Accuracy, $F(2, 590) = 0.98$, $p = .374$, $\eta_p^2 = 0.003$. There was a main effect of Transparency $F(1, 590) = 9.94$, $p = .002$, $\eta_p^2 = 0.017$ 0.001, with more false alarms in the transparent CAD conditions. The CAD Accuracy x Transparency interaction was also significant, $F(2, 590) = 3.42$, $p = .03$, $\eta_p^2 = 0.01$. Planned t-tests showed that False Alarms were higher for Transparent CAD systems when the CAD accuracy rate was 33%, $t(199) = 2.28$, $p = .02$, $d = 0.32$, with anecdotal evidence in support of the alternative, $BF_{10} = 1.70$, and when the CAD accuracy was 66%, $t(195) = 3.17$, $p = .002$, $d = 0.45$, with strong evidence in support of the alternative, $BF_{10} = 15.73$. There was no effect of transparency when the CAD accuracy was 83%, $t(196) = 3.18$, $p = .85$, $d = 0.03$, with substantial evidence in support of the null, $BF_{10} = 0.16$.

Comparisons of individual conditions with the No CAD Control showed fewer false alarms in the 33% Not Transparent CAD condition compared to the No CAD Control. There were no other significant differences.

Sensitivity (*d'*)

The 3 x 2 ANOVA revealed a main effect of CAD Accuracy, $F(2, 590) = 6.67$, $p = .001$, $\eta_p^2 = 0.02$, in which *d'* increased with increased CAD accuracy. There was also a significant main effect of Transparency, $F(1, 590) = 5.71$, $p = .02$, $\eta_p^2 = 0.01$, in which *d'* was lower in the Transparent CAD conditions. The CAD Accuracy x Transparency interaction was not significant, $F(2, 590) = 1.77$, $p = .17$, $\eta_p^2 = 0.006$.

Comparisons of individual conditions with the No CAD Control showed that *d'* was higher in both the 83% accuracy Not Transparent and Transparent conditions in comparison to the No CAD control. There were no other significant differences.

Criterion *(c)*

The 3 x 2 ANOVA revealed a main effect of CAD Accuracy, $F(2, 590) = 8.68$, $p < .001$, $\eta_p^2 = 0.03$, in which people were more willing to respond that a target was present as CAD accuracy increased. There was also a significant main effect of Transparency, $F(1, 590) = 8.07$, $p = .005$, $\eta_p^2 = 0.01$, in which people were more willing to respond that a cancer was present in the Transparent CAD conditions. The CAD Accuracy x Transparency interaction was not significant, $F(2, 590) = 2.10$, $p = .12$, $\eta_p^2 = 0.007$.

Comparisons of individual conditions with the No CAD Control showed that participants were more willing to say a cancer was present in the No CAD control compared to the 33% Not Transparent CAD condition and to the 67% Not Transparent CAD condition. There were no other significant differences.

Recall Rate

The 3 x 2 ANOVA on Recall Rate showed no main effect of CAD Accuracy, $F(2, 590) = 1.21$, $p = .30$, $\eta_p^2 = 0.0043$. There was a significant main effect of Transparency, $F(1, 590) = 10.11$, $p = .002$, $\eta_p^2 = 0.02$, in which Recall Rate was higher in the Transparent CAD conditions. The CAD Accuracy x Transparency interaction was also significant, $F(2, 590) = 3.45$, $p = .03$, $\eta_p^2 = 0.01$. Planned t-tests showed that Recall Rate was higher in the Transparent CAD conditions when the CAD accuracy was 33%, $t(199) = 2.31$, $p = .02$, $d = 0.33$, with anecdotal evidence in

support of the alternative, $BF_{10} = 1.8$, and when the CAD accuracy was 66%, $t(195) = 3.18$, $p = .002$, $d = 0.45$, with strong evidence in support of the alternative, $BF_{10} = 16.32$. However, there was no reliable difference across Transparency when the CAD accuracy was 83%, $t(196) = 3.18$, $p = .86$, $d = 0.03$ with substantial evidence in support of the null, $BF_{10} = 0.16$.

Comparisons of individual conditions with the No CAD Control showed that the Recall Rate was lower in the 33% Not Transparent CAD condition. There were no other significant differences.

Positive Predictive Value (PPV)

The 3 x 2 ANOVA on PPV showed no main effect of CAD Accuracy, $F(2, 590) = 1.40$, $p = .25$, $\eta_{p//}^2 = 0.005$. There was a significant main effect of Transparency, $F(1, 590) = 9.11$, $p = .003$, $\eta_p^2 = 0.02$, in which PPV was lower in the Transparent CAD conditions. The CAD Accuracy x Transparency interaction was not significant, $F(2, 590) = 2.15$, $p = .12$, $\eta_p^2 = 0.007$.

Comparisons of individual conditions with the No CAD Control showed that the PPV was higher in the 33% Not Transparent CAD condition and the 67% Not Transparent CAD condition. There were no other significant differences.

General Discussion

This study examined the effect of transparency on a range of CAD systems that varied in their predictive accuracy. Not surprisingly, the more accurate CAD systems led to better target detection. This increase in accuracy was mostly observed in fewer miss errors given that the CAD accuracy manipulation was specific to target present trials only. More importantly

Transparent CAD conditions showed an impairment in performance with increased false alarms, decreased sensitivity, a less conservative response threshold, an increase in Recall Rate (more women being unnecessarily recalled in a clinical setting) and a decrease in PPV (fewer women being recalled who actually had a cancer). Overall, the data showed that increasing transparency by informing people about the accuracy of the CAD system, led to negative performance across a number of metrics. This is of concern given recent government and regulatory body recommendations that AI systems should show increased transparency (Kerasidou et al., 2022).

The shift in false alarms, recall rate and PPV can be explained by the SDT data, which indicated that participants showed a decrease in sensitivity in the transparent conditions and were more likely to report a cancer was present. Wolfe and Van Wert (2010) proposed a Multiple-Decision Model (MDM) to account for visual search data based on two factors: (i) the amount of time spent searching an image before concluding a target is not there (the 'quitting threshold') and (ii) the amount of evidence above which a target is deemed as present (the response threshold). These factors can be affected by target prevalence (e.g., Wolfe et al., 2007) and by the addition of CAD (e.g., Kunar, 2022). Our data add to this model by showing that CAD Transparency also affects a person's response threshold above which they are willing to accept a target as present.

Comparing performance of the CAD conditions to the No CAD control we see mixed results. Somewhat surprisingly, evidence that performance in the CAD conditions was superior to the No CAD condition was under-whelming. This is particularly true of the Transparent CAD conditions, which showed little difference compared to the No CAD control. One exception was that sensitivity in the Transparent condition, was higher than the No CAD control with a

CAD accuracy rate of 83%. However, this did not translate to better performance in the other metrics. In the Not Transparent conditions there was some improvement over the No CAD baseline. However, this mostly occurred when the CAD accuracy rate was lower[3]. Sensitivity was improved in the 83% CAD accuracy condition, but again this did not translate to improvement in other measures. Given that CAD systems in a clinical setting would be expected to show a high degree of accuracy it is interesting that these systems only showed little improvement in performance compared to No CAD conditions. However, given that other research has shown beneficial effects of CAD over No CAD systems (e.g., Drew et al., 2020), future research would be needed to investigate this further.

One reason why giving people explicit knowledge about CAD accuracy changes their decision outcomes may be because it affects their dependency on those systems. If transparency leads to over-trust in the CAD system, participants would be more likely to accept the CAD recommendation, even if it disagrees with their own judgement (see Felzmann et al., 2020, for a discussion on the link between transparency and trust in AI). Given that explainable AI (XAI) is a complex and difficult field of research (Biran & Cotton, 2017; Scharowski et al., 2023) the above results question whether the need to produce transparent AI is always necessary. Further research needs to be conducted on this but for present purposes our data clearly show that increased transparency about AI accuracy can lead to negative effects.

Lastly, it may be that participants chose not to use the CAD system if they were informed of its inaccuracy. There is some evidence that this may be the case if we examine Figure 2, which shows differences in metrics at the lower CAD accuracy rates (33% and 67%) across

[3] With decreased False Alarms, lower Recall Rate and increased PPV in the 33% accuracy condition, and increased PPV in the 67% accuracy condition

Transparent and Not Transparent conditions. This may suggest that when participants were informed of a lower accuracy rate in the Transparent conditions that the CAD systems were being *under-used*[4]. Given that we did not measure people's perceptions of the CAD systems in these experiments we are unable to determine this here. Nevertheless, this is an important avenue of research for future work.

The current results are important for medical screening in clinical settings. However, there are differences between lab-based studies and medical screening which need further exploration. For example, the prevalence rate of breast cancer would be lower in the clinical setting, and radiologists, of course, have greater expertise than our participants. Furthermore, we operationalised transparency as the level of explicit knowledge available about a system's accuracy. However, there are other factors that affect transparency (Kiseleva et al., 2022). Future research is needed to examine these factors further. However, for now the data suggest that the regulatory goal of making AI systems more transparent, may not always lead to positive outcomes.

[4] We thank Todd Horowitz for this suggestion.

Acknowledgements

The work was supported by a National AI Strategy Award funded by the Alan Turing Institute.

Authors' contributions

Melina Kunar was responsible for the conceptualization, methodology and programming of the experiments. She was also responsible for data collection, analysis of data and writing up the results into manuscript form. Derrick Watson was responsible for programming the experiments, analysis of data and reviewing and editing the manuscript. Giovanni Montana was responsible for reviewing and editing the manuscript. All authors were involved in funding acquisition.

Open Practices Statement

The data and materials for all experiments are available at https://osf.io/zrafu/. None of the experiments were pre-registered.

References

Alexander, A., Jiang, A., Ferreira, C., & Zurkiya, D. (2020). An intelligent future for medical imaging: a market outlook on artificial intelligence for medical imaging. *Journal of the American College of Radiology*, *17*(1), 165-170.

Allen, B., Agarwal, S., Coombs, L., Wald, C., & Dreyer, K. (2021). 2020 ACR Data Science Institute artificial intelligence survey. *Journal of the American College of Radiology*, *18*(8), 1153-1159.

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.

Askin, S., Burkhalter, D., Calado, G. *et al.* (2023). Artificial Intelligence Applied to clinical trials: opportunities and challenges. *Health Technol.* 13**,** 203–213.

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ digital medicine*, *3*(1), 118.

Biran, O., & Cotton, C. (2017, August). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (Vol. 8, No. 1, pp. 8-13)

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1-21.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015, October). The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics* (pp. 160-169). IEEE.

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. https://doi.org/10.1037/xap0000277

Gov, U. K. (2023). A pro-innovation approach to AI regulation. *GOV. UK*. Retrieved from https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

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

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and engineering ethics*, *26*(6), 3333-3361.

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. N Engl J Med, 356, 1399-1409.

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

Freer, T. W. & Ulissey, M. J. (2001) Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology, 220, 781-786.

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., Medical Physics Publishing* (pp. 212-218).

Jacobs, M., Pradier, M. F., McCoy Jr, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment

selections: the example of antidepressant selection. *Translational psychiatry*, *11*(1), 108.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*, 2–9.

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

Jeffreys, H. (1961). Theory of probability (3rd edt.) oxford university press. *MR0187257*, *432*.

Kerasidou, C. X., Kerasidou, A., Buscher, M., & Wilkinson, S. (2022). Before and beyond trust: reliance in medical AI. *Journal of medical ethics*, *48*(11), 852-856.

Kingsman, N., Kazim, E., Chaudhry, A., Hilliard, A., Koshiyama, A., Polle, R., ... & Mohammed, U. (2022). Public sector AI transparency standard: UK Government seeks to lead by example. *Discover Artificial Intelligence*, *2*(1), 2.

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

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., 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.

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.

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.

Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., ... & Martí-Bonmatí, L. (2021). FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medicalimaging. *arXiv preprint arXiv:2109.09658*.

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

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N.,Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A.,Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M.,Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D.,...Shetty, S. (2020). International evaluation of an AI system forbreast cancer screening.Nature,577(7788),

Mello-Thoms, C., & Mello, C. A. (2023). Clinical applications of artificial intelligence in radiology. *The British Journal of Radiology*, *96*(1150), 20221031.

Ng, A. Y., Glocker, B., Oberije, C., Fox, G., Sharma, N., James, J. J., ... & 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.

Patterson, F. & Kunar, M.A. (2024). The Message Matters: Changes to Binary Computer Aided Detection Recommendations Affect Cancer Detection in Low Prevalence Search. *Cognitive Research: Principles and Implications, 9*, 59

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.* 10.3758/s13428-018-01193-y

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.

Taplin, S. H., Rutter, C. M. & Lehman, C. D. (2006) Testing the Effect of ComputerAssisted Detection on Interpretive Performance in Screening Mammography. Am. J. Roentgenol., 187, 1475-1482.

Taylor-Phillips, S., Jenkinson, D., Stinton, C., Kunar, M. A., Watson, D. G., Freeman, K., ... & Clarke, A. (2024). Fatigue and vigilance in medical experts detecting breast cancer. *Proceedings of the National Academy of Sciences*, *121*(11), e2309576121.

Scharowski, N., Perrig, S. A., Svab, M., Opwis, K., & Brühlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. *Frontiers in Computer Science*, *5*, 1151150.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., Morey, R. D. (2018a). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, 25, 35-57.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018b). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 25, 58–76.

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

Zheng, B., Richard, G. S., Sara, G., Christiane, M. H., Ratan, S., Luisa, W. & David, G. (2004) Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments1. Academic radiology, 11, 398-406.