

Analog Computer-Aided Detection (CAD) information can be more effective than binary marks

Corbin A. Cunningham¹ · Trafton Drew² · Jeremy M. Wolfe^{3,4}

Published online: 7 December 2016
© The Psychonomic Society, Inc. 2016

Abstract In socially important visual search tasks, such as baggage screening and diagnostic radiology, experts miss more targets than is desirable. Computer-aided detection (CAD) programs have been developed specifically to improve performance in these professional search tasks. For example, in breast cancer screening, many CAD systems are capable of detecting approximately 90% of breast cancer, with approximately 0.5 false-positive detections per image. Nevertheless, benefits of CAD in clinical settings tend to be small (Birdwell, 2009) or even absent (Meziane et al., 2011; Philpotts, 2009). The marks made by a CAD system can be “binary,” giving the same signal to any location where the signal is above some threshold. Alternatively, a CAD system presents an analog signal that reflects strength of the signal at a location. In the experiments reported, we compare analog and binary CAD presentations using nonexpert observers and artificial stimuli defined by two noisy signals: a visible color signal and an “invisible” signal that informed our simulated CAD system. We found that analog CAD generally yielded better overall performance than binary CAD. The analog benefit is similar at high and low target prevalence. Our data suggest that the form of the CAD signal can directly influence performance. Analog

CAD may allow the computer to be more helpful to the searcher.

Keywords Computer-aided detection · Visual search · Attention · CADE · Medical image perception

Introduction

Visual search in the real world is a difficult process. We know from personal experience that searching for a friend or family member in a huge crowd can quickly become a daunting task. The task is even more daunting for professional searchers, such as radiologists or airport baggage screeners. For these searches, the targets may be from a complex, ill-defined category like “threat” and the image may be ambiguous. The situation is made worse by the low prevalence of targets in many screening tasks (e.g., the prevalence of breast cancer is about 0.5% in a North American screening population (Rosenberg et al., 2006)). Low prevalence tends to elevate false-negative (“miss”) errors (Biggs, Adamo, Mitroff, 2014; Wolfe, Brunelli, Rubinstein, & Horowitz, 2013; Wolfe, Horowitz, & Kenner, 2005), even when the consequences of a missed target can be deadly.

Computer-aided detection (CAD) systems represent one effort to improve the situation. CAD systems are divided into CADE and CADx variants. CADE systems aid in the search for targets. CADx (computer-aided diagnosis) help to determine if a specific item is or is not a target. In radiological screening, CADE systems typically mark potential abnormalities on the image, indicating to the radiologist that they should further scrutinize those portions of the image (Fig. 1). In airport checked baggage screening, CADE systems mark all potential threats (generally explosives). These then must be individually dismissed by the screener or examined by hand.

✉ Corbin A. Cunningham
cunningham@jhu.edu

¹ Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA

² Department of Psychology, University of Utah, Salt Lake City, UT, USA

³ Brigham and Women’s Hospital, Boston, MA, USA

⁴ Harvard Medical School, Boston, MA, USA

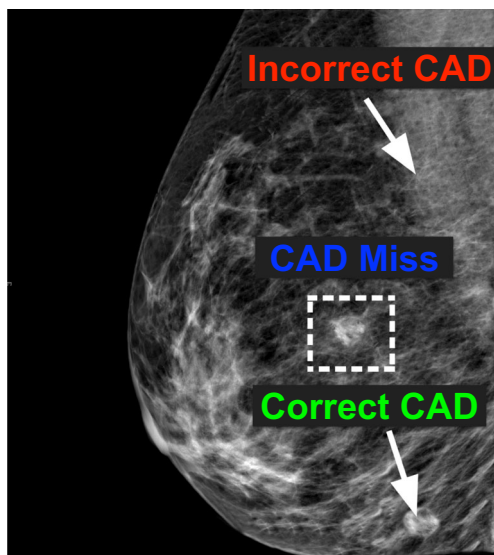


Fig. 1 Artificial example of a computer-aided detection program for mammography. CAD marks shown as white arrows. The reader would typically view the image without CAD and then CAD is utilized as a second reader. The dotted line square marks abnormality that the hypothetical CAD system failed to mark

In checkpoint baggage screening, the object recognition problem is extremely difficult because of the diversity of “threat” (e.g., a gun is a threat at the checkpoint but not in checked luggage). Some detection algorithms have been designed to detect not only explosives, but other threat objects, such as knives and guns (Singh & Singh, 2003), but at this writing, CAD is not used in US checkpoint screening. CADE systems can be quite accurate. In breast cancer screening, good CADE systems might detect 90% of breast cancer, with a false-positive rate of approximately 0.5 false detections per image (Zheng et al., 2001). Nevertheless, performance of radiologists using CADE is not as good as one might expect. Overall, the added benefit of CADE in screening mammography is small (Birdwell, 2009; Gilbert et al., 2006) or nonexistent (Meziane et al., 2011; Philpotts, 2009). It has been suggested using CADE should be like having two radiologists reading a case (Gilbert et al., 2008). However, a meta-analysis demonstrated that screening situations with two radiological readers enhances screening more so than a single reader with CAD (Taylor & Potts, 2008). Some studies have suggested that CADE does not increase hit rate in clinical settings (Brem & Schoonjans, 2001; Gur et al., 2007b). Others have indicated that CADE increases false alarm rate (Fenton et al., 2011). In contrast, there are studies that have shown substantial benefits. For instance, Freer and Ulissey (2001) found that the use of CADE increased cancer detection rates by roughly 20%. Finally, researchers have demonstrated that differences between radiologists in their decision-making with CADE can influence its utility (Alberdi et al., 2014).

Thus, there is something of a mystery; CADE systems are quite proficient and yet the combination of a good CADE

system and a good human expert does not produce the benefits that one might expect. As evidence that at least part of the problem lies in the interaction of human expert and computer, consider that while several studies have shown that CADE can detect up to 77% of cancers missed by radiologists (Brem et al., 2003; Karssemeijer et al., 2003; Nishikawa et al., 2012; Warren Burhenne et al., 2000), Nishikawa and colleagues (2012) have shown that radiologists only respond to a correct CADE prompt 30% of the time.

How might this high rate of unsuccessful use of CADE be explained? Approximately 75% of all mammograms in America are read with the help of CADE programs (Rao et al., 2010). Typically, CADE is used as a “second reader,” invoked after the radiologist has examined of the image. Because the prevalence of both disease in radiology and dangerous objects in baggage screening is extremely low in a screening population, we can imagine why the interaction between CAD systems and the operator is problematic. Imagine 1,000 images, containing 5 cancers/dangerous objects. Let us suppose that the CADE system marks all five. If it is making 0.5 marks per image that makes (at minimum) 500 marks, of which 5 are true positives. The CADE advice, therefore, would have a positive predictive value of approximately 1%. It makes sense to ignore advice that is useful only 1 in 100 times.

Perhaps if each CADE mark were more informative, users would find more utility in the information that CAD systems can provide. One approach that partially addressed this issue was recently outlined by Samulski and colleagues (2010). They evaluated a system that only offers an opinion if asked. When the radiologist actively queried a specific location, that CADE system offered a graded assessment of the features at that location. A follow-up study by Hupse et al. (2013) found that when radiologists used an interactive CAD system such as this, reader performance significantly increased. Their study can be considered to be a computer-aided diagnosis (CADx) system (Fraiola, Serra, & Passariello, 2010). In this paper, we will focus on a related approach to CADE. Specifically, we will consider how CADE information is conveyed to the observer. CADE systems can utilize a binary approach to signaling an abnormality to the user (e.g., putting a mark next to a potential tumor). However, these binary approaches are all-or-nothing, meaning that they do not convey any form of *likelihood* from the CAD probability distribution. Alternatively, information could be presented in the form of an analog signal that specifically indicated the likelihood that a location contained a target. Would observers make more effective use of CADE information if that information was presented as an analog signal at all locations rather than as binary, discrete marks on locations that exceed a likelihood threshold?

Before it would be reasonable to suggest field studies in radiology, airport security, or other such venues, it is important to work out the basic principles in the lab, typically with

nonexpert observers. Thus, in this paper, we compare binary and analog presentation of information in an artificial task, designed to make it possible to study these questions in naïve observers. Findings from lab studies using artificial stimuli certainly can predict search behavior in professional search settings. Previous work has shown that naïve participants and radiologists use CAD in similar ways during search (Drew et al., 2012; Taplin et al., 2006). Similarly, a large body of research in radiology investigates Satisfaction of Search: the successful detection of one lesion appears to hinder the detection of subsequent additional pertinent findings (Berbaum, El-Khoury, Franken, Kuehn, Meis, Dorfman, et al., 1994; Berbaum, Franken, Caldwell, & Schartz, 2009; Berbaum, Franken, Dorfman, Caldwell, & Krupinski, 2000; Berbaum, Franken, Dorfman, Miller, Caldwell, Kuehn, et al., 1998; Samuel, Kundel, Nodine, & Toto, 1995). This body of research has been paralleled by research with naïve observers (Biggs & Mitroff, 2014; Cain, Dunsmoor, LaBar, & Mitroff 2011; Fleck, Samei, & Mitroff, 2010). Similarly, as mentioned, low prevalence raises miss error rates—a finding that was described and explored in a series of studies with nonexpert populations (Wolfe, Horowitz, & Kenner, 2005; Wolfe, Horowitz, Van Wert, Kenner, Place, & Kibbi, 2007; Wolfe, & VanWert, 2010; Rich, Kunar, Van Wert, Hidalgo-Sotelo, Horowitz, & Wolfe, 2008) before being confirmed with cytologists (Evans, Tambouret, Wilbur, Evered, & Wolfe, 2011), radiologists (Evans, Birdwell, & Wolfe 2013), and airport screeners (Wolfe, Brunelli, Rubinstein, & Horowitz, 2013).

The current study focuses on whether we can improve observer performance by altering how CAD information is conveyed to the observer. We created an artificial task where naïve observers were asked to detect the presence of difficult to diagnose “targets” with or without the help of a simulated CAD system. Because we are not specifically interested in the image processing algorithm of a CAD system, this simulated CAD system simply has access to a signal about target presence that is uncorrelated with visible signal, seen by the observer. The critical variable was the mode in which this artificial CAD signal was presented. In Experiment 1, we compared discrete, binary CAD signals to an analog CAD signal in which a graded signal gave the CAD system’s estimate of the likelihood that each item was or was not a target.

Method

Participants

Fifteen observers (mean age = 22.8, standard deviation [SD] = 8.1; 6 females) were tested in Experiment 1. One participant was eliminated from the analysis because that observer responded that a target was present on >99% of trials, where actual target prevalence was 50%. Twelve observers (mean

age = 24.42, SD = 9.34; 7 females) were tested in Experiment 2. Finally, 15 observers were tested in Experiment 3. Observers ranged in age from 19–52 years (average = 25.66, 8 females, SD = 9.28). All observers gave informed consent and were paid \$10/hr to participate in this experiment. All had at least 20/25 vision with correction, all passed the Ishihara Color Test, and all were fluent speakers and readers of English.

Apparatus, stimuli, and experimental setup

All experimental sessions across all three experiments were performed on a Macintosh G4 computer running Mac OS 10.5. Experiments were written in Matlab 7.5 (The Mathworks) using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997), version 3. Stimuli were presented on 20” CRT monitor (Mitsubishi Diamond Pro 91TXM) with resolution set to 1280 × 960 pixels, and an 85-Hz refresh rate. Observers were placed so that their eyes were 57.4 cm from the monitor. At this viewing distance, 1 cm subtends 1° of visual angle. Observers were placed in a silent dimly lit room by themselves for the duration of the task.

Stimuli

Observers were looking for targets based on their color. On each trial, six ambiguously colored patches of dots were presented in a circular array in the center of the screen (Fig. 2). Each dot cluster consisted of a patch made up of 200 smaller dots (approximately 2 pixels in diameter) randomly dispersed around the center of the solid color circle. The color of dot was drawn from one of two normal distributions separated by a d' of 2.5 (Fig. 3). Dots in a target circle were drawn from a

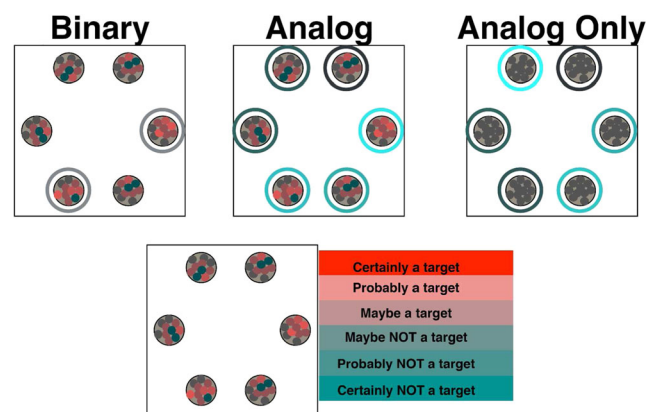


Fig. 2 Dot cluster stimuli. Observers are making a judgment about the average color of the dots. The ring around the dots is the CAD signal. Four conditions are shown from Experiment 1: binary CAD with color signal; analog CAD with color signal, and analog only without a color signal. In the second row, a color signal only accompanied by the response confidence scale. Observers clicked on the stimulus item they thought was “most like a target” and rated their confidence, using the 6-point response scale

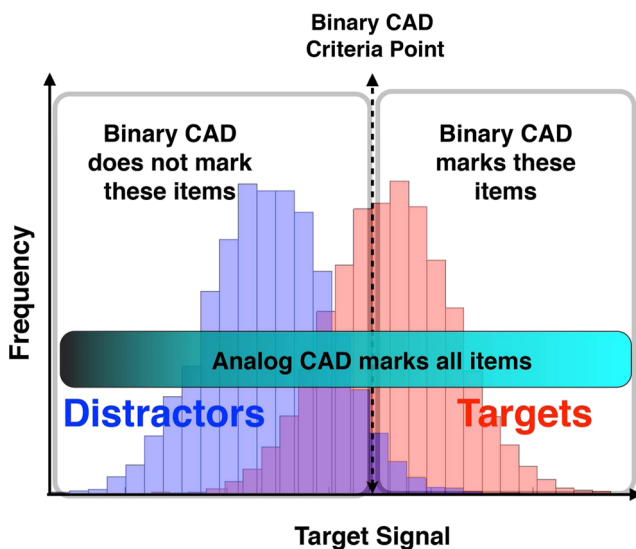


Fig. 3 Distributions of the targets and distractors associated with the color signal and how the different CAD signals would flag those items. Note that the CAD markings were based on an independent signal, distributed in the same way but invisible to the observer

distribution biased toward red. Dots in a distractor circle were drawn from a distribution biased toward blue-green. The d' of 2.5 set an upper bound on the performance of either a perfect CAD system or a perfect human observer. In one condition, the color signal was eliminated and all dots were achromatic and uninformative.

In all cases when CAD was present, the CAD was based on a signal that was independent of the color signal that was visible to the observer. Specifically, the signal for the dot stimuli (e.g., targets and distractors) and the signal for the CAD stimuli were drawn from two separate independent distributions with the same d' . Thus, the combination of CAD and human had more information than was available to either alone. Three simulated CAD systems were used: binary CAD, analog CAD, and analog only. Binary CAD consisted of a grey ring that was placed around the outside of any dot clusters with a CAD signal that were above a threshold. This threshold was set so that CAD would mark 90% of target items and false alarm on 10%. In “analog CAD” conditions, a ring was presented around every dot cluster. The ring indicated the strength of its CAD signal by its color. Analog CAD marked all of the six disks. Items that were more likely to be targets would have CAD rings of a more saturated turquoise color and items that were more likely to be distractors had rings that were closer to a desaturated grey. The color, of course, is arbitrary and is intended merely to produce a gradient of CAD information. Importantly, the Analog CAD signal is a mapping of the independent CAD signal, mentioned above, to a range of colors. This means that the color of the CAD signal was not dependent on the prevalence of targets and distractors. In the analog only condition, the dots were neutral in color so observers had to base their response entirely

on the CAD signal. This is simply a control to show that observers could successfully evaluate the CAD color code. All other experimental conditions included the color signal in the dots.

Prior to initiating each trial, observers were shown a row of 18 examples of randomly generated target dot clusters and another row of 18 distractor clusters. These were given as a reference for the observer to get a sense of the appearance of target clusters and distractor clusters. On the right side of the screen was a 6-point rating scale (Fig. 2). The six levels were labeled: “Certainly a Target”; “Probably a Target”; “Maybe a Target”; “Maybe NOT a Target”; “Probably NOT a Target”; and “Certainly NOT a Target.” The scale went from red to green, red being “Certainly a Target” and “Certainly NOT a Target” in green.

Experimental procedures and design

The Experiment 1 consisted of four blocks with a different condition for each: analog CAD only (no dot color signal), analog CAD with the color signal, binary CAD with the color signal, and color signal only (no CAD). Observers were asked to complete 20 practice trials and 150 experimental trials for each block. The block order was randomized. Target prevalence was 50%. On each trial, the observers were shown an array of six dot clusters. They were instructed to choose the dot cluster that looked most like a target item. After clicking on the item, observers then rated how likely the item was a target item. For example, in some cases all of the dot clusters were actually distractors, so the observer would pick the distractor that appeared most like a target but might rate it as “certainly NOT a target.” Additionally, reaction times were recorded for each trial. Participants were instructed to try to find the correct answer as quickly as possible and told to slow down if they made any inadvertent errors where they quickly selected the wrong option. Medical image perception studies frequently use this method of forcing observers to mark the most suspicious location, even if the case is deemed to be negative (Seltzer, et al. 1995, 1998).

Simulating the conditions of experiment 1

To determine if the analog condition could produce better performance, then the binary conditions, we simulated a variety of possible versions of our basic paradigm. Similar to Experiments 1-3, our model simulation considered virtual search displays that had a visual set size of six items. For target present trials, five of the six items were designated as distractors. These distractor items were assigned a “visual” color signal that was randomly chosen from the distractor color distribution (i.e., a normal distribution). The remaining

target item's "visual" color signal was drawn from a normal distribution with the same standard deviation but a higher mean that could be varied between simulated conditions. For target absent trials, all six items were treated as distractors. The independent CAD signal (for both analog and binary) was simulated in the same way. The strength (d') of the simulated CAD signal could be varied independently of the strength of the simulated visual signal. To simulate the analog condition, the CAD and visual signals were added together. To simulate the binary condition, the CAD signal was recoded to zero if the value fell below threshold and to the mean strength of the signal of the CAD if it fell above threshold. The binary CAD signal was then added to the visual signal. The relative weights of visual and CAD signals can be parametrically varied. We will present examples where the two signals are equally weighted.

To simulate a response, we determined if the largest signal of the six simulated stimuli was above a criterion value. On target present trials, if the signal was above criterion, the trial was scored as a hit or true positive; if not, it was a miss (false negative). For absent trials, if the signal was above criterion, the trial was scored as a false alarm (false positive); if not, it was a correct rejection (true negative). By sweeping the criterion value from low to high, it is possible to create receiver operating characteristic (ROC) curves for each condition. To produce the ROC plots shown in Fig. 4, we simulated 10,000 trials each condition for each set of parameters.

Figure 4 shows simulation results for three different stimulus strengths. The analog (green) curve lies above the binary curve in all conditions. Thus, our simple simulation predicts that performance with analog CAD should be superior to binary CAD as long as our human observers combine CAD information and visual information in a manner similar to the decision criteria outlined above.

Behavioral results and discussion

Experiment 1: Are analog CAD signals better than binary CAD signals?

ROC curves can be derived from rating scale data and these are shown in Fig. 5. The area under the curve (AUC) was computed as the figure of merit for performance on this task. As shown in the figure, AUC for the analog condition is greater than the AUC for the binary conditions ($t(14) = 2.7, p = 0.017$). As should be the case, the Analog CAD combination of analog signal and colored dot signal is better than either analog alone ($t(14) = 4.18, p < 0.001$) or no CAD (color only) ($t(14) = 7.37, p < 0.001$). AUC is a better measure than d' in this case, because the ROC curves are not symmetrical, indicating that underlying, internal signal, and noise distributions do not have equal variance. The zROC functions ($z(\text{false alarm}) \times z(\text{hit})$) have slopes less than 1.0, indicating that the noise variance is less than the signal variance. Interestingly, this is seen in true radiological search tasks as well (Kundel, 2000) and in other complex search tasks (Figure 18 of Wolfe et al., 2007). In the present case, the zROC functions also deviate from linearity making alternatives to D' like $D(a)$ unreliable. AUC seems to be the best choice for assessing the difference between tasks, although the same pattern is seen if other signal detection measures are analyzed.

To determine whether our different CAD formats had an impact on the time required to make a decision, we analyzed reaction time (RT) on each trial. We removed RTs larger than 3.5 standard deviations from the mean for each condition and responses shorter than 100 milliseconds. Figure 6 shows the RTs for the four conditions. There are no significant differences among the conditions, with the exception of the analog only (no color signal) control condition, which is faster, presumably because observers do not have to evaluate the

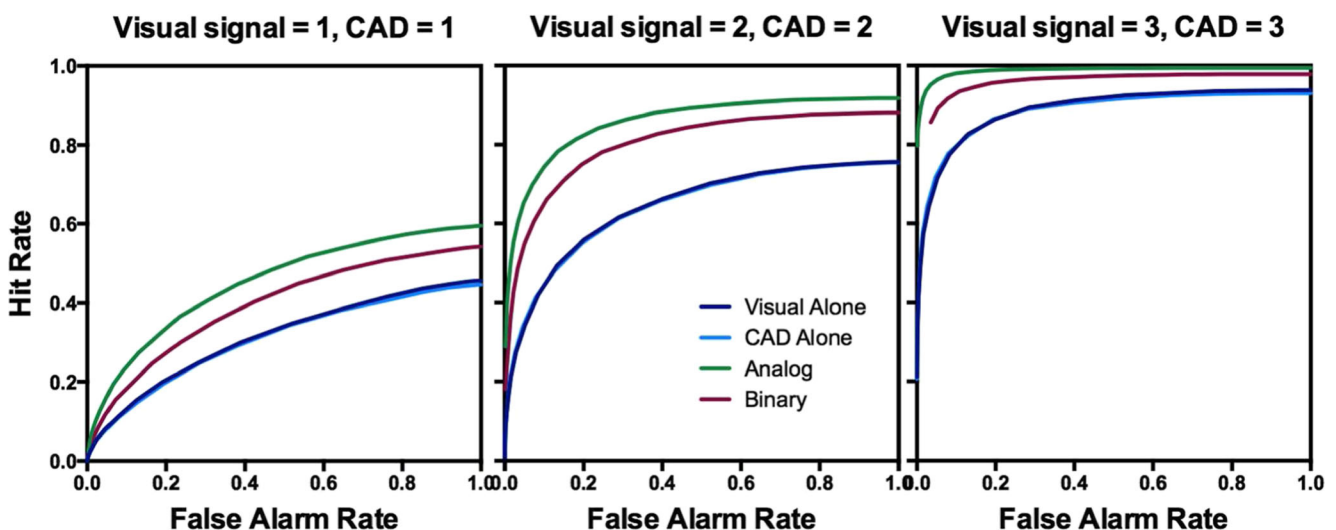


Fig. 4 ROC curves comparing simulations of visual, CAD, analog, and binary conditions over a range of signal strengths

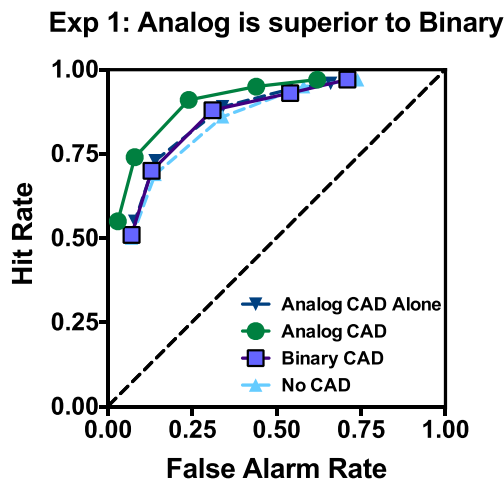


Fig. 5 ROC curve plot for all observers across all conditions in Experiment 1. The black dashed line is chance performance. Each point on the condition lines represents a criterion point. We can see that performance in the Analog CAD condition was better (higher) compared with all of the other conditions

average dot color. A repeated measures ANOVA comparing the reaction times across all four conditions revealed a significant main effect of CAD condition ($F(3, 39) = 4.81, p < 0.05$, Greenhouse-Geisser corrected, partial $\eta^2 = 0.27$). Contrasts revealed that the analog only condition was the only condition that was statistically distinct from the other conditions ($p < 0.01$).

Discussion

Overall, the behavioral of Experiment 1 results are consistent with our simulation: When observers are asked to detect a target among distractors, different methods of conveying the same information influence the ability to perform the task accurately. Specifically, the analog CAD format produced higher AUC values than binary CAD. Although these tasks were created to capture some aspects of the difficult visual searches that

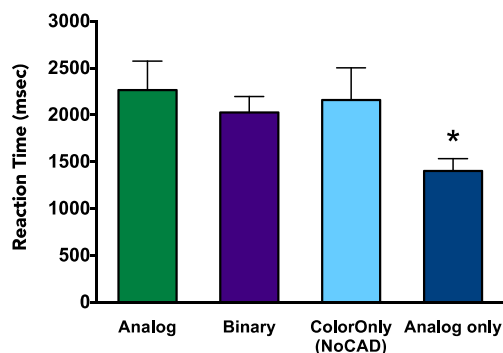


Fig. 6 Reaction times for the four conditions. The critical analog CAD with color condition (green bar) is not statistically different from the other conditions with color signals. The only condition that is statistically different is the analog CAD only (no color signal) condition. Error bars show ± 1 standard error of the mean

professional searchers might encounter, the 50% target prevalence rate was much higher in than it would be in most professional search environments. In Experiment 2, we examine the effects of prevalence on the analog CAD advantage.

Experiment 2: Do prevalence or criterion point matter?

In Experiment 1, we compared observers' performance on a target detection task where observers were instructed to search for the item that looked most like a target. On some trials, they were provided with simulated CAD information either in the form of binary CAD or analog CAD signal. Results from Experiment 1 revealed that observer's performance was higher when they were given analog CAD compared with binary CAD. Experiment 2 investigates whether target prevalence modulates the relative utility of binary CAD and analog CAD systems. Additionally, we were interested in whether there would be effects of manipulating the criterion of the binary CAD. That is, like a human, a binary CAD algorithm is set for one point on an ROC. If one is concerned about a low true positive rate, one can increase that rate by moving criterion to a point where the CAD systems marks more items. Of course, this also increases the false positive rate. It is possible that the improvement, seen with analog CAD, could be seen with a binary CAD operating at a different criterion, if we could modulate performance by simply shifting criterion point of the CAD system, that would be a very simple way to potentially improve radiologist performance. In Experiment 1, we had binary CAD mark 90% of target items and false alarm on 10%. For Experiment 2, we created two separate binary CAD conditions: liberal binary CAD and conservative binary CAD. In the conservative binary CAD condition, the system marked 80% of targets and 5% of distractors. The liberal binary CAD condition marked 95% of target dot clusters and 20% of the distractors. At 50% prevalence, this corresponds to a d' of 2.5 in both conditions with a criterion of -0.4 for the liberal binary CAD and 0.4 for the conservative.

Two prevalence levels were used: 50% and 10%. The validity of CAD marks varies with prevalence and with the liberal/conservative manipulation. Consider 100 trials at a prevalence of 50%. The liberal CAD will mark 47.5 of the targets and 10 distractors for a validity of $47.5/(47.5 + 10) = 0.825$. The conservative CAD yields 0.94 validity at 50% prevalence. An indication of the difficulty faced by CAD at lower prevalence is revealed by performing the same calculation at 10% prevalence. Now the liberal CAD validity is 0.34 and the conservative CAD validity is 0.64. CAD signal validity would be worse, of course, at the prevalence levels of 1% or less, which are seen in medical screening or airport security tasks.

Note that it is something of a misnomer to call our criterion values of ± 0.4 "conservative" and "liberal" when the

prevalence is 10%. Conservative and liberal are defined relative to a neutral criterion where the probability of a positive response is the same as the probability of a negative response (where the two normal curves intersect). For 50% probability, that neutral point lies at zero. For 10% prevalence, it is a z-score of 0.9, because the area under the target curve is so much smaller than the area under the distractor curve. Thus, criteria at -0.4 and $+0.4$ both are liberal relative to the neutral point. At 10% prevalence, it is better to think of 0.4 as “more conservative” than -0.4 but not as absolutely conservative relative to a neutral point.

At 10% prevalence, the positive predictive value (PPV) of liberal CAD mark will mark a target goes down because of the increased chance that the CAD will mark a nontarget. For our stimuli, the PPV for the more conservative CAD criterion is 0.64. For the more liberal criterion, it is 0.35. PPV is high if we make the CAD criterion truly conservative, relative to the low prevalence neutral point of 0.9. At a CAD criterion of 1.3 ($0.9 + 0.4$), the PPV would be 0.91. CAD would mark only 48% of targets, although it would incorrectly mark only 0.5% of the absent trials. The high positive predictive value of 0.91 is good but the failure to mark half of the targets would be a problem. In any case, in Experiment 2, we used the same criterion in low and high prevalence conditions, understanding that at low prevalence, the criteria are both liberal. We will refer to them as “more liberal” and “more conservative” below.

The stimuli for this experiment were otherwise similar to the previous experiment. To incentivize observers to use the full 6-point rating scale, a point system was implemented. Observer would receive 3 points for correct responses at the extremes of the scale (1 and 6). They received 2 points for a 2 or 5 and 1 point for a 3 or 4. Observers lost 1 point for an incorrect response regardless of placement on the scale. The following four experimental conditions were used: more conservative binary CAD condition, more liberal binary CAD condition, analog CAD with color signal condition, and the no CAD color signal only condition. These four conditions were crossed with the two target prevalence levels: high prevalence and low prevalence, yielding eight conditions in a 2×4 design. The procedure was otherwise the same as the previous experiment. Observers were instructed to choose the item that looked most like a target and rate how confident they were that it was a target item on a 6-point scale. The high prevalence blocks consisted of 15 practice trials and 80 experimental trials and low prevalence blocks consisted of 15 practice trials and 400 experimental trials to have adequate numbers of target present trials; 40 target present trials at each prevalence level. Each observer completed high and low prevalence blocks for all of the CAD conditions (8 blocks total). The order of the blocks was randomized. Additionally, the presentations of trials that contained true targets were randomly intermixed with the trials that did not contain targets.

Results

Figure 7 shows average AUC values for each of the eight conditions, calculated as in Experiment 1. There is a modest advantage of Analog CAD over the other CAD conditions. We conducted a 2×4 repeated measures ANOVA across all four CAD conditions and the two prevalence conditions. Results revealed a significant main effect of CAD condition ($F(3, 33) = 3.5, p < 0.035$, Greenhouse-Geisser corrected, partial $\eta^2 = 0.24$), suggesting that performance varied as a function of CAD condition. There was no main effect of the prevalence condition ($F(1, 11) = 1.2, p = 0.30$, partial $\eta^2 = 0.10$). The interaction between target prevalence and CAD condition was not statistically reliable ($p > 0.05$). *A priori* contrasts revealed that analog CAD led to significantly better performance than the more liberal binary CAD conditions ($p < 0.03$). The analog advantage over the more conservative binary CAD was weaker ($p = 0.10$). The two binary CAD conditions did not differ significantly ($p > 0.05$). Additionally, we averaged the more conservative binary CAD and more liberal binary CAD conditions and compared the performance on binary CAD, *in general*, to the analog CAD performance, which resulted in a significant difference ($t(11) = 2.38, p = 0.04$).

Discussion

Results from Experiment 2 weakly replicated the advantage of analog CAD over binary CAD systems. The added information from this experiment is that this pattern of results is essentially the same for low and high prevalence. We also found that criterion point for the CAD system had no effect on observer performance regardless of target prevalence. This argues against the idea that the positive predictive value of each CAD mark strongly influences how the observer uses the

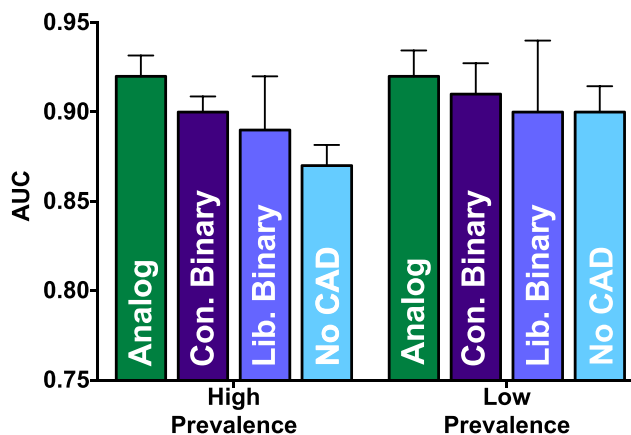


Fig. 7 Results from Experiment 2. Performance was better when observers were given analog CAD compared to both conservative (Con.) and liberal (Lib.) binary CAD. This was consistent across both high and low prevalence environments. Error bars show ± 1 standard error of the mean

CAD system. The difference in PPV between our “More Liberal” and “More Conservative” CAD systems strongly interacts with target prevalence, yet this strong manipulation had no observable effect on performance. Despite the predictive value of each CAD mark in the more conservative CAD block being much higher than in the more liberal CAD block, there was no evidence that the observers used the CAD systems any differently. Furthermore, the main effect of prevalence is on criterion and, thus, on the pattern of errors rather than on measures of overall performance like the AUC (Gur et al., 2007a; Gur et al., 2003). This standard pattern of increased miss errors and reduced false alarms at low prevalence is seen Fig. 8.

For the data in Fig. 8, ratings of 1–3 were taken to be “absent” ratings, whereas 4–6 were scored as present. For each condition, rates of miss (false negative) errors are higher at low prevalence and rates of false alarm (false positive) errors are lower (all paired $t(11) > 3.2$, all $p < 0.01$). It would have been interesting if the analog condition had ameliorated the prevalence effect. However, the basic pattern is the same in each condition. Prevalence effects remain a robust aspect of visual search. For present purposes, the main conclusion of Experiment 2 is that the patterns of results are similar at high and low prevalence and that analog presentation of CAD information preserves a weak advantage over binary. Efforts to change the effectiveness of binary CAD by changing the criterion of the CAD (more liberal vs. more conservative) had no effect in this experiment. It is, of course, possible that a larger manipulation of CAD criterion or a much larger population of observers would reveal an effect but there is no hint of an effect of CAD criterion here.

Experiment 3: What makes Analog CAD better?

Experiments 1 and 2 demonstrated that Analog CAD systems produce somewhat better performance than binary CAD.

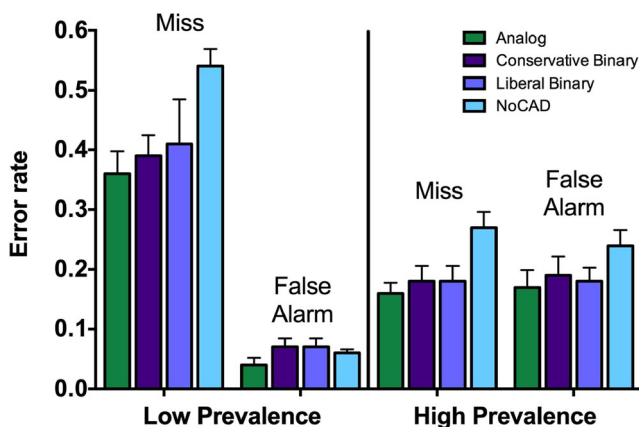


Fig. 8 Effects of prevalence on the pattern of errors in Experiment 2. Low (10%) prevalence produces high rates of miss (false negative) errors and low rates of false alarm (false positive) errors. Error types are more balanced at high (50%) prevalence. Error bars show ± 1 standard error of the mean

What is the source of this advantage? It could be that analog is beneficial because it provides richer information about all likely target items. Alternatively, because analog CAD also provides information about low probability items, it might be useful to allow observers to rule out unlikely items. In Experiment 3, we evaluated these two possibilities. We also investigated whether the analog advantage would survive a change in the form of the CAD information.

The stimuli for this experiment were similar to Experiment 2 unless otherwise noted. Prevalence was 50% in Experiment 3. There was no low prevalence condition. One notable difference was a change in the presentation of the CAD information. Previously, we used a circular ring around the dot clusters. In Experiment 3, we used an asterisk (*) next to each dot cluster to deliver the CAD signal. In the no CAD condition, all asterisks remained white. In other conditions, the color of the asterisks conveyed CAD information. In Experiment 3, there were four different conditions: no CAD, binary CAD, analog CAD, and thresholded analog CAD. In the binary CAD condition, the asterisk was white if the CAD signal was below threshold and blue if the CAD signal was above threshold. Binary CAD marked 90% of target items and would false alarm on 10%. Analog CAD was the same as analog CAD in Experiment 1; however, instead of a ring around the dot cluster, the asterisk color varied continuously as a function of the CAD signal from black (likely nontargets) to blue (likely targets). Finally, the new, thresholded analog CAD condition showed an analog signal (again, black to blue) but only for the 90% of targets and 10% of nontargets that would be marked by a binary CAD. For the other items in the display, the asterisks were white. For example, if the binary CAD conditions marked two dot clusters in a display, the threshold analog CAD also would mark those items but with an analog coloration of the relevant asterisks while the other asterisks remained white.

The procedure was the same as the previous experiment. Observers were instructed to choose the item that looked most like a target and rate how confident they were that it was a target item on a 6-point scale. Observers completed one block for each condition, with 4 blocks total. Each block consisted of 20 practice trials and 180 experimental trials. Block order was randomized across observers.

Results

Figure 9 shows average AUC values for each of the four conditions, calculated as in Experiment 1 and Experiment 2.

Overall, the analog CAD condition produces the highest average AUC followed by the thresholded analog and binary CAD conditions. No CAD, as expected, is the worst. A repeated measures ANOVA across all four CAD conditions revealed a significant main effect of CAD condition ($F(3, 42) =$

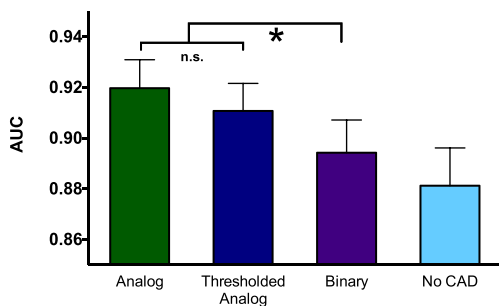


Fig. 9 Average AUC from Experiment 3. Performance was better when observers were given analog CAD or thresholded analog CAD compared with binary CAD. All CAD conditions produced better performance compared to the color only (no CAD) condition. While there was no significant (n.s.) difference between the analog and the thresholded analog conditions ($p > 0.05$), both conditions were significantly different from the Binary condition. Error bars show ± 1 standard error of the mean

5.98, $p < 0.006$, Greenhouse-Geisser corrected, partial $\eta^2 = 0.299$), indicating that performance varied as a function of CAD condition. As in the other experiments, *a priori* contrasts revealed that performance in the analog condition was higher than performance in the binary condition ($p = 0.02$). Additionally, performance in the thresholded analog condition was higher than performance in the binary condition ($p = 0.036$). Finally, although its average AUC was somewhat lower, the thresholded analog condition did not statistically differ from the analog condition ($p = 0.139$).

Discussion

Based on these results, we can reject the hypothesis that the analog CAD advantage requires that graded information be provided about every item in the image, because thresholded CAD is better than binary CAD. We cannot conclude that the graded marking of likely distractors is useful, because the difference between analog and thresholded analog conditions is small and not statistically significant. This could be a power issue. The roughly 1% difference in average AUC would be statistically significant with approximately twice the 15 observers tested or with approximately twice the number of trials per observer. In any case, any advantage of analog over thresholded analog presentation of CAD information appears to be small.

General discussion

This study investigated how different forms of a simulated computer-aided detection systems influence performance in a challenging visual search task. In previous work, we have shown that there are conditions under which CAD can produce a decrement in performance (Drew et al.,

2012). In the present work, we were interested in determining whether different modes of presenting CAD information might be differentially effective. Experiment 1 compared binary CAD and analog CAD systems and found that, when CAD systems provided analog information, overall performance was significantly improved. Experiment 2 showed that these results generalized to lower (10%) target prevalence, although the analog advantage was relatively weak in this experiment. Finally, Experiment 3 replicated the advantage of analog over binary CAD and showed that the analog advantage persisted even if the CAD only marked items above some threshold. This last finding may have practical consequences in considering how analog CAD might be implemented. In our experiment, there were six discrete items to evaluate. A lung or breast is a more continuous stimulus. One way to provide analog information would be a “heat map” of some sort that produced a CAD likelihood value at every location. A different method would be to provide graded, analog CAD information only for a set of locations above some threshold of suspicion. That would be similar to the thresholded analog condition of Experiment 3, and the results suggest that such a method might still be superior to a binary CAD mark. Many current CAD systems provide this graded information when queried (CADx). Our results suggest that conveying this information as part of a CAD detection (CADE) system may improve performance.

Obviously, these experiments are a long distance from the clinic. The experiments presented are, at best, abstract laboratory approximations of the use of CAD in professional search situations. The observers are not trained experts and the stimuli are entirely artificial. However, we believe that results from nonexpert observers are informative nonetheless. Experts and nonexperts all use the same human search engine. Experts can do things with that search engine that amateurs cannot, but expertise does not seem to change the basic rules of visual search. For example, prevalence effects look similar in expert and nonexpert populations (Evans, Birdwell, & Wolfe 2013; Evans, Tambouret, Wilbur, Evered, & Wolfe, 2011; Wolfe, Brunelli, Rubinstein, & Horowitz, 2013) and experts, like nonexperts, are subject to limits like those producing inattentional blindness (Chabris & Simons, 2011; Drew, Vo, & Wolfe, 2013; Mack & Rock, 1998; Simons & Chabris, 1999). The differences in stimuli are more problematic. Discrete clusters of colored dots are a rather schematic approximation to the sorts of complex perceptual decisions made by experts. Consequently, it would be valuable to test the analog/binary difference with a more realistic set of stimuli, ideally with an expert population.

These experiments present one of the first inquiries into how visual search and decision-making are influenced by differential CAD systems. Whereas this initial inquiry addresses some critical questions about the utility of binary and analog

CAD signals in visual search, several future experiments would greatly improve upon this work. For example, in many real-world radiological tasks, there often are multiple cancers presented in a given search display. How would decision-making change when this is the case? More specifically, would the benefits of analog CAD systems change as a function of the number of possible targets in a given display? We know from a number of Satisfaction of Search studies that the presence of multiple targets can influence target detection rates in visual search (Adamo, Cain, & Mitroff, 2013; Berbaum, El-Khoury, Franken, Kuehn, Meis, Dorfman, et al., 1994; Berbaum, Franken, Caldwell, & Schartz, 2009; Berbaum, Franken, Dorfman, Caldwell, & Krupinski, 2000; Berbaum, Franken, Dorfman, Miller, Caldwell, Kuehn, et al., 1998; Biggs & Mitroff, 2014; Biggs & Mitroff, 2015; Fleck, Samei, & Mitroff, 2010; Samuel, Kundel, Nodine, & Toto, 1995). Would decision making change differentially for different CAD signals? Additionally, CAD systems in the real world rely heavily on image processing information from a given search display. In our experiments, we were interested in whether analog CAD was more useful compared with binary CAD, given that the two system were on equal footing (i.e., CAD signals did not incorporate the likelihood of a target on a given trial). Future work should investigate how CAD systems influence decision making and search behavior when information about the search display is incorporated into the CAD signal. For example, does incorporating prevalence information, under a Bayesian framework, into Analog and Binary CAD signals influence observers' reliance on those systems? Overall, it appears that this body of work would benefit greatly from further studies.

Whereas there may be a benefit to analog CAD signals, the modest size of the effects in the present set of experiments suggests that analog CAD is not going to be the “magic bullet” that can make CAD clearly valuable in a clinical setting. Although the analog CAD seems to offer a small improvement, it is worth noting that small statistically reliable improvements, spread over thousands of patients, can represent real progress. The problem of the low positive predictive value of CAD in clinical and other screening situations seems likely to be a much more substantial challenge. One solution is continued improvement of CAD algorithms. If the system were perfect or nearly so, prevalence would cease to be a problem. Until the computer scientists produce those algorithms, it will fall to the behavioral sciences to determine how to deliver more effectively CAD information to the humans who will be using it. While CAD algorithms are adept increasingly at pulling signal out of the noise in complex images, the expert human searchers remain the final arbiters of the information conveyed by most CAD systems. Under these circumstances, behavioral research can provide valuable insight towards optimizing the CAD-expert searcher interaction.

Acknowledgements We thank Preeti Verghese who, in her role as a reviewer, provided valuable motivation and guidance for our modeling work. This work was supported by NIH – NEI EY017001 to JMW and 1F32EB011959-01 to TD.

Compliance with ethical standards

Funding NIH – NEI EY017001 to JMW, 1F32EB011959-01 to TD

References

- Adamo, S. H., Cain, M. S., & Mitroff, S. R. (2013). Self-induced attentional blink a cause of errors in multiple-target search. *Psychological Science*, *24*(12), 2569–2574.
- Alberdi, E., Povyakalo, A. A., Strigini, L., Ayton, P., Hartswood, M., Procter, R., & Slack, R. (2014). Use of computer-aided detection (CAD) tools in screening mammography: A multidisciplinary investigation. *The British journal of radiology*.
- Berbaum, K. S., El-Khoury, G. Y., Franken, E. A., Jr., Kuehn, D. M., Meis, D. M., Dorfman, D. D., ... Kathol, M. H. (1994). Missed fractures resulting from satisfaction of search effect. *Emergency Radiology*, *1*, 242–249.
- Berbaum, K. S., Franken, E. A., Jr., Caldwell, R. T., & Schartz, K. M. (2009). Satisfaction of search in traditional radiographic imaging. In E. Samei & E. A. Krupinski (Eds.), *The handbook of medical image perception and techniques*. Cambridge: Cambridge University Press.
- Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Caldwell, R. T., & Krupinski, E. A. (2000). Role of faulty decision making in the satisfaction of search effect in chest radiography. *Academic Radiology*, *7*, 1098–1106.
- Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Miller, E. M., Caldwell, R. T., Kuehn, D. M., & Berbaum, M. L. (1998). Role of faulty visual search in the satisfaction of search effect in chest radiography. *Academic Radiology*, *5*, 9–19.
- Biggs, A. T., Adamo, S. H., & Mitroff, S. R. (2014). Rare, but obviously there: Effects of target frequency and salience on visual search accuracy. *Acta Psychologica*, *152*, 158–165.
- Biggs, A. T., & Mitroff, S. R. (2014). Different predictors of multiple-target search accuracy between nonprofessional and professional visual searchers. *The Quarterly Journal of Experimental Psychology*, *67*(7), 1335–1348.
- Biggs, A. T., & Mitroff, S. R. (2015). Improving the efficacy of security screening tasks: A review of visual search challenges and ways to mitigate their adverse effects. *Applied Cognitive Psychology*, *29*(1), 142–148.
- Birdwell, R. L. (2009). The preponderance of evidence supports computer-aided detection for screening mammography. *Radiology*, *253*(1), 9–16. doi:10.1148/radiol.2531090611
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. doi:10.1163/156856897x00357
- Brem, R. F., Baum, J., Lechner, M., Kaplan, S., Souders, S., Naul, L. G., & Hoffmeister, J. (2003). Improvement in sensitivity of screening mammography with computer-aided detection: A multiinstitutional trial. *AJR American Journal of Roentgenology*, *181*(3), 687–693. doi:10.2214/ajr.181.3.1810687
- Brem, R. F., & Schoonjans, J. M. (2001). Radiologist detection of microcalcifications with and without computer-aided detection: A comparative study. *Clinical Radiology*, *56*(2), 150–154.

- Cain, M. S., Dunsmoor, J. E., LaBar, K. S., & Mitroff, S. R. (2011). Anticipatory anxiety hinders detection of a second target in dual-target search. *Psychological Science*, *22*(7), 866–871.
- Chabris, C. F., & Simons, D. J. (2011). You do not talk about Fight Club if you do not notice Fight Club: Inattention blindness for a simulated real-world assault. *i-Perception*, *2*(online).
- 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., Vo, M. L.-H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattention blindness in expert observers. *Psychological Science*, *24*(9), 1848–1853. doi:10.1177/0956797613479386
- 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. doi:10.1371/journal.pone.0064366
- Evans, K. K., Tambouret, R., Wilbur, D. C., Evered, A., & Wolfe, J. M. (2011). Prevalence of abnormalities influences Cytologists' error rates in screening for cervical cancer. *Archives of Pathology & Laboratory Medicine*, *135*(12), 1557–1560. doi:10.5858/arpa.2010-0739-OA
- Fenton, J. J., Abraham, L., Taplin, S. H., Geller, B. M., Carney, P. A., D'Orsi, C., ... Barlow, W. E. (2011). Effectiveness of computer-aided detection in community mammography practice. *Journal of the National Cancer Institute*, *103*(15), 1152–1161.
- Fleck, M. S., Samei, E., & Mitroff, S. R. (2010). Generalized "satisfaction of search": Adverse influences on dual-target search accuracy. *Journal of Experimental Psychology: Applied*, *16*(1), 60.
- Fraioli, F., Serra, G., & Passariello, R. (2010). CAD (computer-aided detection) and CADx (computer aided diagnosis) systems in identifying and characterising lung nodules on chest CT: Overview of research, developments and new prospects. *Radiologia Medica*, *115*(3), 385–402. doi:10.1007/s11547-010-0507-2
- Freer, T. W., & Ulissey, M. J. (2001). Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center 1. *Radiology*, *220*(3), 781–786.
- Gilbert, F. J., Astley, S. M., Gillan, M. G., Agbaje, O. F., Wallis, M. G., James, J., ... & Duffy, S. W. (2008). Single reading with computer-aided detection for screening mammography. *New England Journal of Medicine*, *359*(16), 1675–1684.
- Gilbert, F. J., Astley, S. M., McGee, M. A., Gillan, M. G., Boggis, C. R., Griffiths, P. M., & Duffy, S. W. (2006). Single reading with computer-aided detection and double reading of screening mammograms in the united kingdom national breast screening program 1. *Radiology*, *241*(1), 47–53.
- Gur, D., Bandos, A. I., Fuhrman, C. R., Klym, A. H., King, J. L., & Rockette, H. E. (2007). The prevalence effect in a laboratory environment changing the confidence ratings. *Academic Radiology*, *14*(1), 49–53.
- Gur, D., Rockette, H. E., Armfield, D. R., Blachar, A., Bogan, J. K., Brancatelli, G., ... Warfel, T. E. (2003). Prevalence effect in a laboratory environment. *Radiology*, *228*(1), 10–14. doi:10.1148/radiol.2281020709
- Hupse, R., Samulski, M., Lobbes, M. B., Mann, R. M., Mus, R., den Heeten, G. J., ... & Karssemeijer, N. (2013). Computer-aided detection of masses at mammography: Interactive decision support versus prompts. *Radiology*, *266*(1), 123–129.
- Karssemeijer, N., Otten, J. D., Verbeek, A. L., Groenewoud, J. H., de Koning, H. J., Hendriks, J. H., & Holland, R. (2003). Computer-aided detection versus independent double reading of masses on mammograms. *Radiology*, *227*(1), 192–200. doi:10.1148/radiol.2271011962
- Kundel, H. L. (2000). Disease prevalence and the index of detectability: A survey of studies of lung cancer detection by chest radiography. In E. A. Krupinski (Ed.), *Medical imaging 2000: image perception and performance* (vol. 3981, pp. 135–144).
- Mack, A., & Rock, I. (1998). *Inattention blindness*. Cambridge: MIT Press.
- Meziane, M., Obuchowski, N. A., Lababede, O., Lieber, M. L., Philips, M., & Mazzone, P. (2011). A comparison of follow-up recommendations by chest radiologists, general radiologists, and pulmonologists using computer-aided detection to assess radiographs for actionable pulmonary nodules. *American Journal of Roentgenology*, *196*(5), W542–549.
- Nishikawa, R. M., Schmidt, R. A., Linver, M. N., Edwards, A. V., Papaioannou, J., & Stull, M. A. (2012). Clinically missed cancer: How effectively can radiologists use computer-aided detection? *AJR American Journal of Roentgenology*, *198*(3), 708–716. doi:10.2214/AJR.11.6423
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. doi:10.1163/156856897x00366
- Philpotts, L. E. (2009). Can computer-aided detection be detrimental to mammographic interpretation? 1. *Radiology*, *253*(1), 17–22.
- Rao, V. M., Levin, D. C., Parker, L., Cavanaugh, B., Frangos, A. J., & Sunshine, J. H. (2010). How widely is computer-aided detection used in screening and diagnostic mammography? *Journal of the American College of Radiology*, *7*(10), 802–805.
- Rich, A. N., Kunar, M. A., VanWert, 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, article 15), 1–17.
- Rosenberg, R. D., Yankaskas, B. C., Abraham, L. A., Sickles, E. A., Lehman, C. D., Geller, B. M., & Ballard-Barbash, R. (2006). Performance benchmarks for screening mammography. *Radiology*, *241*(1), 55–66. doi:10.1148/radiol.2411051504
- Samuel, S., Kundel, H. L., Nodine, C. F., & Toto, L. C. (1995). Mechanisms of satisfaction of search: Eye position recordings in the reading of chest radiographs. *Radiology*, *194*, 895–902.
- Samulski, M., Hupse, R., Boetes, C., Mus, R. D., den Heeten, G. J., & Karssemeijer, N. (2010). Using computer-aided detection in mammography as a decision support. *European Radiology*, *20*(10), 2323–2330. doi:10.1007/s00330-010-1821-8
- Seltzer, S. E., Judy, P. F., Adams, D. F., Jacobson, F. L., Stark, P., Kikinis, R., ... Feldman, U. (1995). Spiral CT of the chest: Comparison of cine and film-based viewing. *Radiology*, *197*(1), 73–78.
- Seltzer, S. E., Judy, P. F., Feldman, U., Scarff, L., & Jacobson, F. L. (1998). Influence of CT image size and format on accuracy of lung nodule detection. *Radiology*, *206*(3), 617–622. doi:10.1148/radiology.206.3.9494475
- Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattention blindness for dynamic events. *Perception*, *28*(9), 1059–1074.
- Singh, S., & Singh, M. (2003). Explosives detection systems (EDS) for aviation security. *Signal Processing*, *83*(1), 31–55.
- 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.
- Warren Burhenne, L. J., Wood, S. A., D'Orsi, C. J., Feig, S. A., Kopans, D. B., O'Shaughnessy, K. F., & Castellino, R. A. (2000). Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology*, *215*(2), 554–562.
- Wolfe, J. M., Brunelli, D. N., Rubinstein, J., & Horowitz, T. S. (2013). Prevalence effects in newly trained airport checkpoint screeners:

- Trained observers miss rare targets, too. *Journal of Vision*, 13(3:33), 1–9. doi:10.1167/13.3.33
- Wolfe, J. M., Horowitz, T. S., VanWert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *JEP: General*, 136(4), 623–638. doi:10.1037/0096-3445.136.4.623
- Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare targets are often missed in visual search. *Nature*, 435(7041), 439–440. doi:10.1038/435439a
- Wolfe, J. M., & VanWert, M. J. (2010). Varying target prevalence reveals two, dissociable decision criteria in visual search. *Current Biology*, 20(2), 121–124. doi:10.1016/j.cub.2009.11.066
- Zheng, B., Ganott, M. A., Britton, C. A., Hakim, C. M., Hardesty, L. A., Chang, T. S., et al. (2001). Soft-copy mammographic readings with different computer-assisted detection cuing environments: Preliminary findings. *Radiology*, 221(3), 633–640.