

Clemson University

TigerPrints

All Theses

Theses

5-2024

Detection and Assessment of Targets in Mock Digital Breast Tomosynthesis Images Using Computer-aided Detection Systems

Katharine Sabo
sabo2@g.clemson.edu

Follow this and additional works at: https://tigerprints.clemson.edu/all_theses



Part of the [Human Factors Psychology Commons](#)

Recommended Citation

Sabo, Katharine, "Detection and Assessment of Targets in Mock Digital Breast Tomosynthesis Images Using Computer-aided Detection Systems" (2024). *All Theses*. 4219.
https://tigerprints.clemson.edu/all_theses/4219

This Thesis is brought to you for free and open access by the Theses at TigerPrints. It has been accepted for inclusion in All Theses by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.

DETECTION AND ASSESSMENT OF TARGETS IN MOCK
DIGITAL BREAST TOMOSYNTHESIS IMAGES USING
COMPUTER-AIDED DETECTION SYSTEMS

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirement for the Degree
Master of Science
Applied Psychology

by
Katharine Sabo
May 2024

Accepted by:
Dr. Dawn Sarno, Ph.D., Committee Chair
Dr. Mark Neider, Ph.D.
Dr. Patrick Rosopa, Ph.D.

ABSTRACT

Radiologists use computer-aided detection (CAD) systems to assist with detecting and assessing breast cancers in mammograms, digital breast tomosynthesis (DBT), and other types of breast imaging; however, the usefulness of such automation aids has been debated since their implementation. Initial studies on CAD systems had mixed results, while more recent studies have shown that they can improve diagnostic outcomes (i.e., greater sensitivity and fewer false alarms) and decrease the reading time required for images. Three CAD types are currently in use: binary, analog, and interactive CAD. However, studies rarely explore the differences between the CAD methods. Additionally, recent work has suggested the benefit of including Breast Imaging Reporting and Database System (BI-RADS) ratings in CAD systems to help standardize assessments and improve diagnostic performance. The present study adds to the literature by comparing the three existing types of CAD (binary, analog, and interactive) and including a novel CAD system with BI-RADS ratings. Fifty undergraduate students completed a visual search task with mock DBT images and were assigned to either one of the CAD conditions or a control. Participants also completed surveys regarding their propensity to trust automation in general and their perceptions of the system's usability and trustworthiness. Results suggested that the binary and analog CAD systems improved participants' hit rate and sensitivity (d'). Additionally, participants in the analog and interactive CAD conditions appeared to trust their CAD aids more than those in the binary and BI-RADS conditions. Regarding perceived usability, participants in the binary, analog, and interactive CAD conditions rated their CAD aids with higher usability

scores than those in the BI-RADS CAD condition. Exploratory analyses provided support for the trust and usability findings in the present study and suggested that participants in the BI-RADS condition were more conservative than those in the other CAD conditions. Together, the results from the present study provide further support that the use of CAD aids, particularly those that provide additional information, can improve hit rate and sensitivity when assessing DBT images. Although the novel BI-RADS CAD aid did not perform as strongly as the other CAD systems, it is promising that this variation did not harm reader performance. More research is required to explore how different implementations of BI-RADS ratings within CAD systems may be beneficial in the assessment of DBT images. Findings from this research contribute to the development of more user-centered CAD systems to ultimately improve radiologists' diagnostic performance, particularly for DBT images.

ACKNOWLEDGMENTS

I have been fortunate to have the support of so many wonderful people as I have tackled this significant step in my academic journey. First, I would like to thank my advisor, Dr. Dawn Sarno, for guiding me through the process of creating and running a study rooted in the research and sharing her enthusiasm for exploring this area with me. Our discussions throughout this process have helped me grow as a researcher and writer; I am proud of the work we have done and look forward to continuing to learn from her.

Next, I would like to extend my thanks to my committee members for their support. To Dr. Patrick Rosopa, for his patience, encouragement, and assistance with the statistical analyses. To Dr. Mark Neider, for his avid interest in the project and for his guidance with the visual search aspects. They helped make this experience a pleasure.

From coast to coast, my friends and family have been essential sources of support and encouragement throughout this process. I would like to thank them for their interest in my studies, their advice, their empathetic listening, their company, and for adding light to my life. Special thanks to my wonderful lab partner Jeff Black for helping with the programming and to Kalvry Cooper for her positivity and support. I am extremely lucky to have so many amazing people in my corner, and I deeply appreciate every one.

Finally, I would like to thank Dr. Gabriella Hancock, Nick Roome, and Professor Jaye Van Kirk. Without their support and guidance, I would not have found my home in human factors. Professor Van Kirk introduced me to it and Dr. Hancock and Nick have been instrumental in fostering my love of human factors through our discussions, collaborations, and community activities. I am immensely grateful to each of them.

TABLE OF CONTENTS

TITLE PAGE	i
ABSTRACT.....	ii
ACKNOWLEDGMENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	ix
CHAPTER	
I. INTRODUCTION.....	1
Detection and Assessment of Potential Anomalies in Breast Images	3
Computer-aided Detection (CAD) Systems in Breast Cancer Imaging	10
The Rise of Digital Breast Tomosynthesis (DBT)	19
The Present Study	24
II. METHODS.....	27
Participants.....	27
Stimuli and Procedures	28
III. RESULTS.....	37
Differences in Search Performance by Condition and Prevalence	37
Activation of iCAD.....	47
Trust in the Specific CAD System.....	47
Perceptions of the Usability of CAD Systems.....	49

Exploratory Analyses.....	51
IV. DISCUSSION	57
Differences in Search Performance by Condition and Prevalence	57
Activation of iCAD.....	62
Perceived Usability and Trust in the Specific CAD Systems	64
SDT: Response Criterion	70
V. LIMITATIONS AND FUTURE DIRECTIONS	72
VI. CONCLUSION	76
REFERENCES	78
APPENDICES	104
A. Reliability-inspired Programming by Prevalence and CAD Conditions...	105
B. Propensity to Trust	106
C. System Usability Scale (SUS).....	107
D. Trust Between People and Automation.....	108
E. Qualitative Usability Questionnaire	109
F. Assumptions for the 5x2 Mixed Design ANOVAs.....	110
G. Responses to the Qualitative Usability Questionnaire	113
H. Circle Trial Information	116
I. Descriptive Statistics for Total iCAD Activation.....	120

LIST OF TABLES

Table		Page
3.1	Descriptive Statistics for Control Condition vs. CAD Conditions	40
3.2	Descriptive Statistics for Control Condition vs. CAD Conditions, Low Prevalence	41
3.3	Descriptive Statistics for Hit Rate by Automation Condition	43
3.4	Descriptive Statistics for False Alarm Rate by Automation Condition	44
3.5	Descriptive Statistics for Sensitivity (d') by Automation Condition	46
3.6	Descriptive Statistics for Target Absent Response Times by Automation Condition	47
3.7	Descriptive Statistics for Average Specific Trust by CAD Condition	49
3.8	Descriptive Statistics for Average SUS Score by CAD Condition	51
3.9	Correlations Between SUS and Specific Trust Scores by CAD Condition	51
A.1	Reliability-inspired Programming by Prevalence and CAD Conditions ...	105
F.1	Normality Assumption for the 5x2 Mixed Design ANOVAs	110
F.2	Homoscedasticity Assumption for the 5x2 Mixed Design ANOVAs	112
G.1	Themes from Responses to the Qualitative Survey by CAD Condition....	113
H.1	Circle Trial Information	116
H.2	Screengrabs of Off-Center Circled Items in the Binary CAD Condition ..	117
H.3	Screengrabs of Off-Center Circled Items in the aCAD and iCAD Conditions	117

H.4	Circle Trial Information for Low-accuracy Trials	118
H.5	Screengrabs of Circles for Identified Trials in the Binary CAD Condition.....	118
H.6	Screengrabs of Circles for Identified Trials in the aCAD and iCAD Conditions	119
I.1	Descriptive Statistics for Total Activation of iCAD Overlay	120

LIST OF FIGURES

Figure		Page
2.1	Navigation Through Image Layers	29
2.2	User Interfaces of Each Condition.....	32
3.1	Comparing Hit Rate for Control Condition vs. CAD Conditions.....	40
3.2	Comparing Low Prevalence Block Hit Rate for Control Condition vs. CAD Conditions.....	41
3.3	Hit Rate by Automation Condition.....	43
3.4	Sensitivity (d') by Automation Condition	45
3.5	Average Trust in the Specific System by CAD Condition	48
3.6	Average SUS Score by CAD Condition	50
3.7	Average Response Criterion by CAD Condition.....	55

CHAPTER ONE

INTRODUCTION

The American Cancer Society estimates that there will be 300,590 new cases of breast cancer in the United States in 2023, with breast cancer projected to make up 31% of all new cancer cases in women and retain its rank as the second leading cause of female cancer deaths (Siegel et al., 2023). Despite breast cancer incidence rates steadily increasing since the early 2000s, improved screening and treatment options have led to a decline in breast cancer mortality rates (Berry et al., 2005; Giaquinto et al., 2022; Myers et al., 2015; Narayan et al., 2020; Nelson et al., 2016; Oeffinger et al., 2015; Siegel et al., 2023). While breast cancer is typically thought of as a woman's disease, men are also at risk of developing breast cancer: men have roughly 1% of the breast cancer diagnoses in America, with a lifetime risk of getting breast cancer at about 1 in 833 (Division of Cancer Prevention and Control, 2023; Siegel et al., 2023). With breast cancer becoming more prevalent in the population, finding ways to improve early breast cancer detection continues to be a pressing issue.

Early breast cancer detection has primarily been accomplished in two ways: regular clinical breast exams (CBEs) and annual or biannual imaging screening after a woman reaches 40 or 50 years old (potentially earlier if she is deemed high-risk). While CBEs have long been recommended by doctors to identify potentially abnormal physical changes to the breasts that may indicate the presence of breast cancer, there is a lack of empirical support for this method as a reliable screening tool (Oeffinger et al., 2015). In contrast, imaging screening for breast cancer, primarily done via mammography or

digital breast tomosynthesis (DBT), is a highly recommended tool for catching breast cancer in the early stages, often before symptoms even begin to appear (Coleman, 2017; Siu & U.S. Preventive Services Task Force, 2016; Smith & Oeffinger, 2020). Research continues to show that participation in regular imaging screening, particularly for women after age 50, leads to substantial reductions in advanced and fatal breast cancer rates by catching the cancer early enough to begin treatment (Duffy et al., 2020; Jhangiani et al., 2023; Myers et al., 2015; Narayan et al., 2020; Nelson et al., 2016; Oeffinger et al., 2015; Pace & Keating, 2014). Unfortunately, it appears likely that there will soon be additional observational support that regular screening is crucial to decreasing breast cancer mortality rates, due to the coronavirus disease 2019 (COVID-19) pandemic. Delays of breast cancer screening and diagnosis caused by disruptions in medical services during COVID-19 have been estimated to result in over 2,000 excess deaths caused by breast cancer by 2030 (Alagoz et al., 2021). As breast imaging technology continues to improve, it will remain an essential component of proactive care.

Of course, early detection of breast cancer has benefits beyond reducing mortality rates. Catching signs of breast cancer early can have significant financial implications. Later-stage breast cancer treatments (e.g., chemotherapy and mastectomy) mean a greater financial burden on patients as well as a higher likelihood of job and wage loss post-treatment, particularly for those patients who are part of a minority group, low-income, or self-employed (Blinder & Gany, 2020; Sun et al., 2018). The earlier breast cancer can be detected, diagnosed, and treated, the less financial impact it will have on the patient and their loved ones. Improvements in breast imaging technology will not only save lives, but

they will also ultimately reduce the financial burden for patients whose breast cancer would otherwise have progressed to the more harmful late stages.

Detection and Assessment of Potential Anomalies in Breast Images

When radiologists examine mammograms and DBT images, they are looking for abnormalities in the breast tissue such as calcifications, masses, asymmetries, architectural distortion, and other suspicious findings (American Cancer Society, 2022). Calcifications typically appear as small white spots that can be classified as either macrocalcifications (larger and typically non-cancerous) or microcalcifications (much smaller and can indicate cancer depending on shape and layout). Masses appear as abnormal areas in the breast tissue that can be classified as cysts (fluid-filled sacs that can indicate cancer depending on their size, shape, and edges) or solid masses (usually non-cancerous). Asymmetries typically appear as white areas that have a different pattern than the surrounding normal breast tissue. Architectural distortion appears as pulled or otherwise contorted breast tissue. Distortions may be due to positioning during the imaging procedure, prior injury, previous breast procedures (e.g., surgery), or cancer. Although asymmetries and distortions are usually not cancer, it is likely that patients will require additional imaging. The ability to correctly distinguish between cancerous and non-cancerous abnormalities is essential to avoid unnecessary recalls of the patient for additional breast imaging that can be painful, costly, and expose the patient to additional radiation.

The detection and assessment of anomalies in imaging from mammography and

DBT can be greatly impacted by breast density. Breast density is determined by the proportion of fibrous and glandular tissue to fatty tissue, with density increasing as the percentage of fibrous and glandular tissue rises. Dense breasts are normal but can have a slightly higher risk of cancer and can create more challenges when it comes to detecting signs of cancer in mammogram and DBT images (American Cancer Society, 2022). The fibrous and glandular tissue that is characteristic of dense breasts appears white on mammogram and DBT imaging, which can act as camouflage for some types of suspicious findings (Indian Radiologist, 2021). Density is such an important factor in the detection and assessment of potential anomalies that it has its own method of assessment and categorization in breast imaging reporting. The reports generated by radiologists and other breast imaging readers are used to further guide patient care.

The Breast Imaging Reporting and Database System (BI-RADS) Scoring System

The Breast Imaging Reporting and Database System (BI-RADS) is used by radiologists for patient case management. The original intent of BI-RADS was to improve standardization for reporting results for breast cancer imaging screening through a “living” document that can be adapted as technology evolves (Burnside et al., 2009; Magny et al., 2023). BI-RADS has overall classifications for breast imaging that include: “Category 0,” incomplete; “Category 1,” negative; “Category 2,” benign; “Category 3,” probably benign; “Category 4,” suspicion of malignancy; “Category 5,” highly suggestive of malignancy; and “Category 6,” known biopsy-proven malignancy (Lieberman & Menell, 2002; Sickles et al., 2013). BI-RADS Category 4 is further divided into three

subcategories based on how likely it is that cancer is present: “4A,” low chance of cancer (2-10%); “4B,” moderate chance of cancer (10-50%); and “4C,” high chance of cancer (50-95%). While these three subcategories of Category 4 exist, their implementation varies depending on location, medical center, training, and other factors (Indian Radiologist, 2021; Mahfouz, 2016; Pijnappel et al., 2004). BI-RADS Category 3 has similar issues and is considered more proper to use after diagnostic rather than screening imaging to avoid unnecessary biopsies, delayed diagnoses, and confusion in both medical personnel and patients (Indian Radiologist, 2021; Pijnappel et al., 2004; Sickles et al., 2013; Trieu et al., 2020). BI-RADS Category 6 is only used after an abnormality has been biopsy-proven to be malignant (Magny et al., 2023; Sickles et al., 2013). Although there continues to be some debate about how to apply BI-RADS in practice, one of the benefits of it being a “living” document is that with more research and discussion, it can be updated to reflect evidence-based best practices.

There are several important components to imaging reports that support the overall BI-RADS classification for breast cancer images. Radiologists must account for patient history, breast composition, and any potential abnormalities that may be indicators of breast cancer to determine BI-RADS ratings for screening and diagnostic images (Barazi & Gunduru, 2023; Burnside et al., 2009). The BI-RADS rating provides further guidance for follow-up care, which can include recommendations for when the patient should come back for their next screening appointment and whether a biopsy should be performed. If an image set is classified incorrectly, the patient may be recalled for unnecessary further imaging or biopsies. Unfortunately, there is wide inter-observer

variability for BI-RADS ratings, and even intra-observer variability (Berg et al., 2000; Boumaraf et al., 2020; Geras et al., 2018; Lazarus et al., 2006; Melnikow et al., 2016; Obenauer et al., 2005; Pijnappel et al., 2004). The goals of implementing BI-RADS have only been partially realized: the structure and information to be included in reports has been standardized and the BI-RADS scores provide a quick way to communicate crucial details to other medical personnel, but radiologists and other breast imaging readers still need to develop a consensus for the determination of BI-RADS classifications in the interest of better patient outcomes.

To that end, it has been proposed that the BI-RADS scale should be integrated into computer-aided detection (CAD) systems to assist with assessment standardization (Qian et al., 2015). While very little progress has been made on this front, there has been some success using machine learning and natural language processing (NLP) approaches to determine BI-RADS ratings for radiology reports (Banerjee et al., 2019; Bozkurt et al., 2015; Bozkurt et al., 2016; Castro et al., 2017; Sippo et al., 2013). Although these NLP systems have largely been successful, they are not in wide use and are still in the initial exploration phase. It has been suggested that such NLP systems could be an integral part to automated decision support systems, continuing education, standardization of reporting, and the overall management of breast cancer screening (Bozkurt et al., 2016; Castro et al., 2017; Sippo et al., 2013). With this initial success in the automation of BI-RADS ratings from written reports, as well as other improvements in automated detection and assessment technology, it is time to start preparing for how best to implement suggested BI-RADS ratings in CAD user interfaces for image assessment.

The Prevalence Effect

While it is important to know what radiologists look for and how they assess breast screening images for potential malignancies, it is equally important to observe that an extraordinary amount of these images are negative for signs of cancer. Studies have found that less than 0.5% of breast screening images are positive for signs of breast cancer, making the detection and assessment of potential abnormalities in breast cancer imaging a rare-target visual search task (Biggs et al., 2014; Chan et al., 2019; Chen & Howe, 2016; Evans et al., 2013; Gur et al., 2004; Lehman et al., 2017; Nishikawa & Bae, 2018; Pisano et al., 2005; Wolfe et al., 2005). Low target prevalence has been shown to impact visual search performance, via the prevalence effect, by making it more likely for participants to fail to detect a target when it is present, leading to an increase in errors and decrease in accuracy (Biggs et al., 2014; Chen & Howe, 2016; Evans et al., 2013; Wolfe et al., 2005). When applied to breast cancer imaging search tasks, the prevalence effect can have deadly consequences.

Evaluating Performance: Diagnostic Performance

Signal detection theory (SDT; Green & Swets, 1988; Tanner & Swets, 1954) is one of the most common ways to assess the performance of radiologists. SDT is typically applied to reader performance for breast cancer screening using the following definitions: hits are considered to be when a reader accurately detects cancer is present; correct rejections are considered to be when a reader accurately determines there is no cancer present; misses are considered to be when a reader does not detect that cancer is present;

and false alarms are considered to be when cancer is not present, but the reader determines that one has been detected. Misses are the least desirable outcome, as the stakes are high if a radiologist fails to detect signs of cancer, particularly when it is best to catch it early enough to be successfully treated (Duffy et al., 2020; Jhangiani et al., 2023; Myers et al., 2015; Narayan et al., 2020; Nelson et al., 2016; Oeffinger et al., 2015; Pace & Keating, 2014). False alarms are more often referred to as “false positives” in breast cancer screening and, while not fatal like misses, are still not desirable because they often result in the need for additional imaging or biopsies and an increase in patient anxiety (Dustler, 2020; Geras et al., 2018; Pace & Keating, 2014; Qian et al., 2015). To maximize their chances of catching cancer in screening imaging, radiologists are trained to have a more liberal criterion so that they are more likely to say that cancer is present, resulting in fewer misses but more false alarms (Alberdi et al., 2004). Ideally, radiologists should have a high proportion of hits and correct rejections and a low proportion of misses and false alarms.

To that end, and to simplify analyses for easier comparisons, sensitivity and specificity are two other important measures often used to assess radiologist performance. Sensitivity refers to the proportion of cancers that are correctly identified out of all cancers in the image set, and specificity refers to the proportion of normal cases (those without cancer) that are correctly identified out of all normal cases in the image set (Alberdi et al., 2004). Comparing these measures across studies allows researchers to evaluate how reader performance has changed over time, differs between populations or imaging methods, or is impacted by study manipulations (e.g., prevalence, types of

targets, number of readers, or the use of search aids). It should be noted that radiologists tend to have high specificity (i.e., they are good at correctly identifying normal cases) but have greater differences and deficits in sensitivity (i.e., ability to correctly identify cancers; Dustler, 2020; Kunar, 2022; Mann et al., 2020; Salim et al., 2020; Shoshan et al., 2022; Qian et al., 2015). Regarding breast cancer screening with mammography, radiologists generally have at least 90% specificity while sensitivity has been found to range from as low as approximately 40% to as high as approximately 97%, though it should be noted that studies evaluating sensitivity measures do not consistently have radiologists assess images using the same scale and radiologists usually have a sensitivity of about 90% (Dustler, 2020; Katzen & Dodelson, 2018; Mushlin et al., 1998; Pisano et al., 2005; Qian et al., 2015). While research exploring how to improve radiologist diagnostic performance will usually report both sensitivity and specificity, sensitivity improvements tend to be the primary area of interest.

Evaluating Performance: Reading/Response Time

An additional measure of performance that is of particular interest to radiologists, breast imaging manufacturers and software developers, hospitals, and other breast screening centers is reading or response time. Response time is how long it takes radiologists or other image readers to classify the imaging they are evaluating. Studies have found that radiologists can use their expertise to make correct global assessments of normal or abnormal scans at a rate above chance (~70%) within 0.2s (Kundel & Nodine, 1975; Drew et al., 2013), but it typically takes radiologists longer to complete their

evaluations of breast imaging in practice. However, there is wide variability in reported reading times: average reading time per case for digital mammography has been reported to take 33-240s and for DBT with 2D mammography the average reading time per case has varied from 64.1-168s, depending on radiologist specialty and experience (Bernardi et al., 2012; Conant et al., 2019; Dang et al., 2014; Haygood et al., 2009; Lee et al., 2022). With such wide variability in reading times in mind, methods for bringing down the high end of the range are being investigated.

The amount of breast imaging scans radiologists have to review is only increasing as the population gets older. In the United States, the U.S. Preventative Services Task Force recommends that women between the ages of 50 and 74 years get biennial screening mammography (Siu & U.S. Preventive Services Task Force, 2016). Breast cancer screening is firmly established as a regular part of medical care; exploring how to reduce response times while maintaining or improving diagnostic performance is essential for ensuring patients receive appropriate care in a timely fashion (Balleyguier et al., 2017; Benedikt et al., 2018; Chae et al., 2019; Conant et al., 2019; Gao et al., 2019; Mann et al., 2020; Shoshan et al., 2022). As breast imaging technology improves, particularly the development and implementation of automated aids, reading time can be expected to continue to be a measure of interest in radiologist performance studies.

Computer-aided Detection (CAD) Systems in Breast Cancer Imaging

The use of CAD systems to assist radiologists with identifying cancers in the screening of breast images such as mammograms and digital breast tomosynthesis (DBT)

imaging has been rapidly increasing since the United States Food and Drug Administration (FDA) approved the use of CAD technology for this purpose in 1998 (Fazal et al., 2018; Fenton et al., 2011; Gao et al., 2019; Katzen & Dodelzon, 2018; Keen et al., 2018; Lehman et al., 2015; Richman et al., 2019). Currently, CAD is approved for use as a second reader for the assessment of breast cancer screening imaging to confirm the radiologist's initial findings, though its implementation is not standardized so it is often used in single reader settings (Chan et al., 2019; Fazal et al., 2018; Henriksen et al., 2019; Masud et al., 2019; Nishikawa & Bae, 2018). CAD systems use algorithms, machine learning, deep learning, and artificial intelligence (AI) techniques to mark suspicious features in breast images to bring them to the attention of the reader. Today, the majority of breast screening imaging assessment is performed with the assistance of a CAD system.

SDT is one of the most common ways to assess the effectiveness of CAD as an automation aid for radiologists and other breast imaging readers. In particular, researchers are interested in how the use of CAD might influence measures of sensitivity (Alberdi et al., 2009; Chae et al., 2019; Conant et al., 2019; Fenton et al., 2011; Kunar, 2022; Lehman et al., 2015; Salim et al., 2020; Shoshan et al., 2022). It is important to note that while specificity is often also reported when evaluating radiologists' performance with the use of CAD, CAD systems were intended to aid radiologists in catching cancers they might otherwise miss; they were not designed with specificity as a priority (Nishikawa & Bae, 2018). For CAD systems, there is an additional concern as the number of false alarms for current CAD systems impedes radiologists' diagnostic

performance (Fazal et al., 2018; Gao et al., 2019; Katzen & Dodelzon, 2018). As with radiologists, CAD systems are trained to have a more liberal criterion, meaning that they present readers with relatively more false positive indications (Alberdi et al., 2004; Chan et al., 2019; Kunar, 2022; Nishikawa & Gur, 2014). Researchers are working on developing better algorithms for CAD to decrease the number of false alarms while not sacrificing their sensitivity (Cortez et al., 2021; Gandomkar & Mello-Thoms, 2019; Fazal et al., 2018; Geras et al., 2018; Kohli & Saurabh, 2018; Kyono et al., 2018; Qian et al., 2015). Currently, CAD systems and radiologists have approximately the same average accuracy (84%; Kohli & Saurabh, 2018), but advances in machine learning, AI, and image processing technology promise to lead to the development of CAD systems that will be able to identify underlying patterns and clues that radiologists, with the limitations of human performance capabilities, are blind to (Cortez et al., 2021; Fazal et al., 2018; Gao et al., 2019; Geras et al., 2018; Jairam & Ha, 2022; Kohli & Saurabh, 2018; Kyono et al., 2018; Qian et al., 2015). Whether currently or in the future, in the context of reader performance, the use of CAD should improve sensitivity.

Despite the employment of CAD in most breast cancer screening assessments, a critical study by Lehman et al. (2015) found there to be no benefit to CAD use on any measure. In contrast, they found that the use of CAD significantly decreased sensitivity for the interpretation of mammograms compared to reviewing the images without CAD. Many other studies have reached similar conclusions that the use of CAD in the assessment of breast cancer imaging has no benefit, negatively impacts performance, or has such a wide range of potential impact on performance that it is impossible to

determine if it is truly a useful tool (Cole et al., 2014; Drew & Reback, 2017; Fazal et al., 2018; Fenton et al., 2011; Jorritsma et al., 2015; Katzen & Dodelson, 2018; Keen et al., 2018; Kunar et al., 2017; Taylor & Potts, 2008). Some of this conflict in outcomes may be due to how the use of CAD can impact radiologists' confidence and how they deal with the false positive CAD information. Radiologists, especially novices, may become over-reliant on the CAD system instead of trusting their own judgment (Jorritsma et al., 2015; Kunar et al., 2017; Nishikawa & Bae, 2018). This over-reliance on the CAD system can exacerbate the negative effects of CAD on diagnostic performance, especially since CAD systems are not yet (and may never be) 100% reliable. In contrast, some radiologists may completely disregard CAD information that they disagree with if they are confident in their own expertise or do not trust the system (Cole et al., 2014; Jorritsma et al., 2015; Katzen & Dodelson, 2018; Nishikawa & Bae, 2018; Nishikawa & Gur, 2014). As mentioned previously, one of the major drawbacks of current CAD systems is that they have a high rate of false positives. False positives can cause readers to question themselves or to take extra time to determine that the false positive markings are incorrect (Katzen & Dodelson, 2018; Kohli & Saurabh, 2018). A majority of past research suggests that CAD systems may be more of a hindrance than aid for radiologists.

Continued CAD use in radiology may be surprising until observing that most studies showing little to no benefit or even harm from the use of CAD in the assessment of breast imaging were conducted at least five to ten years ago; the technology used for CAD systems has improved substantially since then. Recent research has shown that the use of CAD in breast cancer screenings can be at least as good as using a second reader

when considering recall rates, sensitivity, and cancer detection rates (Henriksen et al., 2019). Concurrent CAD use with DBT images has been found to decrease reading time without degrading diagnostic performance (Balleyguier et al., 2017; Benedikt et al., 2018; Chae et al., 2019; Conant et al., 2019; Gao et al., 2019). In fact, research exploring new machine learning, deep learning, and AI algorithms for CAD systems for breast imaging detection and assessment have found that not only is the CAD program improving in its diagnostic performance, its use is also resulting in improved performance for radiologists (Conant et al., 2019; Du-Crow et al., 2019; Dustler, 2020; Jairam & Ha, 2022; Kim et al., 2020; Lee et al., 2023). CAD use may also be able to mitigate the prevalence effect, though this largely depends on how information is presented and the accuracy (particularly in terms of lower false alarm rates) of the CAD systems (Drew et al., 2020; Kunar, 2022; Kunar et al., 2017; Tan et al., 2015). As the diagnostic performance of CAD systems continues to improve with advances in technology, it is becoming increasingly important to investigate how the presentation of pertinent information can impact reader performance as well.

Types of CAD

Diagnostic performance can also be influenced by the type of CAD implementation the radiologist uses in their work. The three most common types of CAD implementation are binary (traditional) CAD, analog CAD (aCAD), and interactive CAD (iCAD). Binary CAD automatically provides visual indicators of anomalies in the images if the area of interest passes a particular threshold of possibility to have a “target present”

(Cunningham et al., 2017; Du-Crow et al., 2020; Hupse et al., 2013; Samulski et al., 2010). A slightly more advanced version of traditional CAD is aCAD, which automatically provides a visual indicator of a possible anomaly and a percentage-based suggestion of whether the target is present (Cunningham et al., 2017; Du-Crow et al., 2020; iCAD, Inc., 2020). Quickly gaining traction is the third type of CAD, iCAD. While iCAD still provides a visual indicator of a possible anomaly with a percentage-based suggestion of whether the target is present, it only does so when the iCAD overlay is activated by the participant (Du-Crow et al., 2020; Hupse et al., 2013; Samulski et al., 2010). All three types of CAD are currently in use in hospitals and other breast cancer screening centers.

While new CAD algorithms continue to be developed to improve sensitivity, specificity, and false positive rates, little research has been done to examine the usability of these systems' user interfaces, or how best to present information to radiologists to maximize their potential performance gains (Cortez et al., 2021; Drew et al., 2020; Gandomkar & Mello-Thoms, 2019; Geras et al., 2018; Kyono et al., 2018; Nishikawa & Bae, 2018; Nishikawa & Gur, 2014; Qian et al., 2015). It has been suggested that giving radiologists the option to activate a CAD overlay, as with iCAD, may improve diagnostic performance and response times, particularly if the CAD displays classification or other pertinent information, as with aCAD, that could be helpful for reader interpretation of the images (Drew et al., 2020; Gao et al., 2019; Nishikawa & Bae, 2018; Nishikawa & Gur, 2014). Allowing users to activate the CAD overlay after they have had the opportunity to view the image by itself has also been found to mitigate the potential costs of the

distraction effect that traditional CAD can produce during low-prevalence target searches (Drew et al., 2020; Kunar, 2022). Studies comparing aCAD and iCAD systems to a binary CAD system have found that radiologists performed better when they used the aCAD and iCAD implementations (Cunningham et al., 2017; Drew et al., 2020; Du-Crow et al., 2020). Though CAD research has shown benefits to using all three types of CAD, what little research has been done comparing them has pointed to the most benefit coming from the use of iCAD, with aCAD close behind due to its presentation of additional pertinent information.

Integrating BI-RADS with CAD Systems

An essential step in developing CAD systems is choosing what information to show the reader and how to display that information. Researchers have suggested that providing radiologists with classification information may help with detection and assessment tasks (Nishikawa & Gur, 2014; Tan et al., 2017; Tan et al., 2015; Qian et al., 2015). Integrating suggested BI-RADS classifications into the CAD system may provide such useful information that radiologists can consider when they make their final reports about the breast screening images. Although BI-RADS ratings specifically have not yet been integrated into CAD systems, there is evidence that the addition of case-based risk assessment scores can help radiologists determine which image sets they may need to examine more carefully (Tan et al., 2017; Tan et al., 2015; Qian et al., 2015). The tested case-based CAD systems were able to increase sensitivity while also decreasing the negative impact of false positives. Another argument to make for adding suggested BI-

RADS classifications to CAD systems is that it could help address the inter- and intra-observer variability issues the BI-RADS implementation currently has (Berg et al., 2000; Boumaraf et al., 2020; Geras et al., 2018; Lazarus et al., 2006; Melnikow et al., 2016; Obenauer et al., 2005; Pijnappel et al., 2004; Qian et al., 2015). By using CAD to give radiologists a starting point for a BI-RADS category assessment, the BI-RADS goal of standardization may be more readily achieved. With case-based risk assessment scores having promise as a way to improve radiologist performance with CAD and the desire to increase inter-observer reliability with BI-RADS use, investigating how best to present suggested BI-RADS ratings appears to be the next step.

Trust in Automation

The presentation of information by a CAD system can only impact the performance of the radiologist if the radiologist uses the CAD in their search and assessment of the breast screening imaging. As mentioned previously, it has been observed that radiologists often either under-trust or over-trust CAD aids, leading to disuse and misuse of the systems (Du-Crow et al., 2019; Jorritsma et al., 2015; Kunar et al., 2017; Nishikawa & Bae, 2018; Nishikawa & Gur, 2014). Disuse of CAD can occur when radiologists ignore the information provided by the CAD system or otherwise underutilize the CAD system (Dzindolet et al., 2003; Jorritsma et al., 2015; Parasuraman & Riley, 1997). In contrast, when radiologists misuse CAD, they may stop searching the breast imaging earlier, change their initial correct target-present determinations to target-absent, and generally perform worse on diagnostic measures as a result of relying on the

CAD system too much (Du-Crow et al., 2019; Jorritsma et al., 2015; Kunar et al., 2017; Nishikawa & Bae, 2018; Parasuraman & Riley, 1997). The goal of CAD implementation is, therefore, for radiologists to trust the CAD aid enough to use it for the benefits it can provide while also trusting their own judgment enough to know which CAD cues may be unreliable and can be ignored.

Usability of CAD Systems

Understanding how radiologists perceive the usability of the CAD systems they partner with will be essential to the appropriate, long-term use of these automation aids. Trust in automation can be influenced by the design of the system and vice versa, with both elements influencing the adoption and use of CAD technology (Filice & Ratwani, 2020; Hoff & Bashir, 2015; Jorritsma et al., 2015). While evaluations of CAD systems have primarily focused on their functionality, the lack of usability in these systems has led to confusion, disuse, and poorer outcomes than expected based on the performance of CAD algorithms, suggesting that usability testing will be crucial to the success of future CAD and AI tools (Filice & Ratwani, 2020; Jorritsma et al., 2014; Jorritsma et al., 2015; Lam Shin Cheung et al., 2023; Lekadir et al., 2023; Lekadir et al., 2021; Nishikawa & Bae, 2018). Yet, there are few usability studies of CAD systems for breast imaging. Improvements in the detection and assessment abilities of CAD imaging software and algorithms are virtually meaningless if the users of CAD systems are unable to appropriately take advantage of the information these automation aids provide due to poor UI design. By investigating how best to present information to radiologists,

improvements can be made to the UI to allow for even better diagnostic performance and patient outcomes.

The Rise of Digital Breast Tomosynthesis (DBT)

Mammography vs. DBT

When people think about breast cancer screening imaging, what most people think about is mammography. Mammography is currently the most common type of breast cancer screening (Katzen & Dodelzon, 2018; Siu & U.S. Preventive Services Task Force, 2016). Recent mammography tends to be full-field digital mammography instead of film mammography and consists of 2D images of breast tissue. A full set of mammography images usually consists of two images per breast, a craniocaudal view and a mediolateral view. To take the images, the breast is compressed between two metal plates. A much less painful procedure for mammography called synthetic mammography has started to gain ground. This type of mammography is created by combining x-ray “slices” of the breast taken during DBT into one 2D image (Lowry et al., 2020; Rocha García & Mera Fernández, 2019; Zuckerman et al., 2016). Synthetic mammography taken with DBT removes the element of pain from mammography and substantially decreases the amount of radiation patients receive compared to the use of digital mammography with DBT, but there is not enough evidence yet to support it as a full replacement for digital mammography.

DBT can be considered an evolution of mammography in that it is a compilation of multiple 2D image “slices” taken by an x-ray tube that rotates in an arc around the

breast. Radiologists can use this compiled image to “drill” through imaging of the breast and get a better idea of the breast tissue and the structure of potential anomalies in the breast. Due to the 3D information provided by DBT imaging, DBT is often referred to as “3D mammography,” but this term is incorrect; DBT creates a pseudo-third dimension using planar data (Sickles et al., 2013; Rocha García & Mera Fernández, 2019). DBT has been in use for many years now but has not been formally adopted as a screening or diagnostic tool. Current guidelines in both the United States and Europe mention DBT but note that more research needs to be done before DBT can be officially recommended for screening, let alone as a diagnostic method (Schünemann et al., 2020; Siu, A. L., & U.S. Preventive Services Task Force). Despite this, DBT’s popularity continues to increase and, as of March 2023, 86% of certified breast cancer screening facilities had DBT capabilities (Lee & Moy, 2023). It is anticipated that DBT will soon replace digital mammography as the primary method for breast screening imaging, with the use of DBT as a screening tool steadily rising since its introduction (Lee & Moy, 2023; Richman et al., 2019). With the adoption of DBT only expanding throughout the United States, it is not unreasonable to suggest that we might see a statistic within the next ten years that near 100% of certified breast cancer screening facilities will have DBT access.

One of the areas where DBT is expected to perform well is in the screening of high-density breasts (Gur, 2007). Compared to mammography, DBT has significantly better sensitivity, cancer detection rates, and recall rates for high-density breasts (Chong et al., 2019; Houssami et al., 2023; Lee & Moy, 2023; Marinovich et al., 2018). Without DBT, patients with high-density breasts would need additional imaging with more

advanced screening equipment such as magnetic resonance imaging (MRIs) or ultrasound, both of which can be challenging to find (less facilities have access to these machines), have high costs, and are often considered uncomfortable for patients (Gur, 2007). As DBT continues to be implemented in hospitals and other breast cancer screening centers, breast imaging practitioners become more familiar with using and interpreting DBT imaging, and research continues to support the improvement of diagnostic performance, DBT is likely to replace mammography as the primary method for breast screening imaging (Chong et al., 2019; Lowry et al., 2020; Zuckerman et al., 2016). With the increasing use of DBT as a screening and diagnostic tool, it is crucial to ensure that radiologists are more helped than hindered by DBT.

Reasons to Switch

While studies generally indicate that one of the main benefits of mammography screening is a reduction of breast cancer mortality, it has been noted that screening mammography can also result in harms such as false positives, unnecessary recalls, overdiagnosis, and overtreatment (Friedewald et al., 2014; Geras et al., 2018; Jhangiani et al., 2023; Myers et al., 2015; Narayan et al., 2020; Pace & Keating, 2014; Schünemann et al., 2020). When a patient has more mammograms, especially when they begin regular screening at younger ages, their risk of a false positive increases. Although the impact of a false positive result can be highly individualized, studies have found that these results tend to exacerbate feelings of depression, anxiety, and worry as well as prompt unnecessary biopsies (Dustler, 2020; Geras et al., 2018; Jairam & Ha, 2022; Pace &

Keating, 2014; Qian et al., 2015). Overdiagnosis, though largely still under debate both in definition and in how it should be operationalized, can lead to women being diagnosed with breast cancer that would not have impacted their lives if it had not been detected and treated (Narayan et al., 2020; Pace & Keating, 2014; Ryser et al., 2022). Extraneous treatment, more commonly referred to as overtreatment, is also not well-studied but is estimated to occur in similar rates as overdiagnosis (Pace & Keating, 2014). Patients who undergo screening mammography are also exposed to radiation and pain during the procedure to take these breast images (Jhangiani et al., 2023; Qian et al., 2015). These physical risks can lead to a decrease in the likelihood that a patient will continue with appropriate screening over the course of their lifetime. Despite the potential harms of mammography screening, the benefits are considered to be great enough that it remains in common practice.

With advances in breast imaging technology, though, such harms can be reduced. Compared to mammography, using DBT for breast cancer screening has been shown to result in lower breast cancer mortality, small increases in quality-adjusted life-years, and fewer false-positives (Lowry et al., 2020). There has also been evidence that the use of DBT may reduce unnecessary recall rates, though some studies have found that recall rates may be higher when DBT is used to assess breasts with high density (Chong et al., 2019; Houssami et al., 2023; Marinovich et al., 2018; Melnikow et al., 2016). These differences may be attributed to differences in sensitivity and cancer detection rate measures. Multiple studies have found that sensitivity and cancer detection rates are significantly higher for DBT than mammography screening, particularly when evaluating

high-density breasts (Chong et al., 2019; Houssami et al., 2023; Kerlikowske et al., 2022; Lee & Moy, 2023; Marinovich et al., 2018). With the improvements that come from using DBT instead of mammography, the potential harms from false positives, unnecessary recalls, overdiagnosis, and overtreatment are reduced.

Current State of DBT Research

Compared to research on mammography, research on DBT is quite sparse. Much of the research investigating diagnostic performance outcomes with DBT does not solely look at DBT; rather, the researchers add or compare it to another method of breast cancer screening imaging. For example, several studies have explored the use of DBT either in tandem with or as opposed to digital mammography (Friedewald et al., 2014; Kerlikowske et al., 2022; Lee & Moy, 2023). When it comes to integrating CAD use with DBT and mammography imaging, several studies have found that radiologists have better diagnostic results using DBT with CAD than they do using mammography with CAD (Balleyguier et al., 2017; Katzen & Dodelzon, 2018). As far as CAD use in itself, the vast majority of the literature is focused on using CAD with mammography imaging rather than DBT. Similarly, for aspects specific to DBT such as search behavior when there are multiple layers to “drill” through, research is just beginning (Adamo et al., 2018; Aizenman et al., 2017). Overall, the consensus appears to be that more research is needed regarding the impact of DBT on diagnostic performance outcomes, with and without CAD and digital or synthetic mammography, with a longitudinal aspect, and including randomized trials (Schünemann et al., 2020; Siu, A. L., & U.S. Preventive Services Task

Force). To assist with gathering this data, the American College of Radiology (ACR) has begun recruiting for its Tomosynthesis Mammographic Imaging Screening Trial (TMIST), a randomized longitudinal study, and has already recruited over half the number of participants they need to reach their goal sample size (American College of Radiology, 2023). However, data from the TMIST and other longitudinal studies investigating the use of DBT is not expected to be available for many more years. As more data is collected about DBT, its impact on performance, longitudinal results, with mammography, and using CAD systems with DBT, more concrete determinations will be able to be made about the best practices for DBT in the breast cancer screening process.

The Present Study

To explore the effect of CAD systems on visual search performance, usability perceptions, and trust when evaluating DBT imaging, we investigated the three common types of CAD aids (binary, aCAD, and iCAD) as well as a novel CAD system that utilized an overall BI-RADS rating. The BI-RADS CAD system presented participants with a suggested classification for the image set from a modified BI-RADS rating scale that included 0 (need more information; similar to an error message), 1 (no target present), 2 (neutral or equal likelihood of target present), and 3 (target present). Similar modified BI-RADS scales are often used in studies investigating how researchers make BI-RADS determinations for breast scan images (Banerjee et al., 2019; Boumaraf et al., 2020; Cortez et al., 2021; Geras et al., 2018; Narváez et al., 2017).

Further contributions to the body of research regarding the usability of CAD

systems came from the study's comparison of four types of CAD interfaces: binary CAD, BI-RADS CAD, aCAD, and iCAD. While past research has examined one to three CAD implementations per study, such comparative research is rare and tends to rely solely on SDT analyses rather than how specific elements of the CAD interfaces may influence diagnostic performance and reader speed (Cunningham et al., 2017; Drew et al., 2020; Du-Crow et al., 2020; Hupse et al., 2013; Kunar, 2022; Kunar et al., 2017; Samulski et al., 2010). To that end, the present study examined SDT, target absent response times, trust and usability outcomes while maintaining target consistency among the four CAD implementations and a control condition with no CAD. Reliability was held consistent among the CAD conditions. Reader performance was assessed in both high and low target prevalence blocks to mimic past research conditions.

Hypotheses

H1: Participants will have higher sensitivity (d' and hit rate) and have faster target absent response times using a CAD interface than without one.

H2: Participants will have higher sensitivity (d' and hit rate) and have faster target absent response times using a CAD interface than without one in the low prevalence condition.

H3: Participants will have the fewest false alarms and have the fastest target absent response times with iCAD aid.

H4: Activation of the iCAD overlay will be more common in the low prevalence block than the high prevalence block.

H5: Participants in the binary CAD condition will have less trust in their automation aid than participants in other CAD conditions when controlling for propensity to trust automation in general.

H6: Participants in the binary CAD condition will have lower usability scores than participants in the other CAD conditions.

Exploratory Hypotheses

H7: There will be significant positive correlations between usability scores and trust in the specific system for each CAD condition.

H8: There will be no significant differences in response criterion between CAD conditions.

CHAPTER TWO

METHODS

Participants

A total of 50 participants were recruited from Clemson University's student population through the Sona system. A repeated measures ANOVA power analysis was conducted in G*Power (Faul et al., 2007) to determine the necessary sample size to detect an effect of CAD condition on diagnostic performance measures. An effect size of $\eta_p^2 = .24$ was estimated based on results from studies by Cunningham et al. (2017) and Kunar (2022) that found similar effects when comparing performance between CAD types. Thus, an ANOVA power analysis with the following parameters, a Cohen's *f* of .56, power of 0.95, an alpha probability of 0.05, and 5 groups (control, binary CAD, BI-RADS CAD, aCAD, and iCAD) was conducted. Based on this analysis, 50 participants (10 per group) should be sufficient to detect any differences in performance due to CAD condition.

Participants were compensated with course credit. While undergraduates with minimal to no experience with radiology (i.e., non-professionals) provided the sample population in this study, previous research has indicated that such non-professional participants have comparable outcomes to professional participants and their results can be used to gain insight into the performance of experienced radiology imaging readers (Adamo et al., 2018; Du-Crow et al., 2020; Fleck et al., 2010; Samulski et al., 2010). All participants had self-reported normal or corrected-to-normal vision (20/32 or better

corrected vision on a Snellen eye chart) and normal cognitive function. The present study complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at Clemson University. Informed consent was obtained from each participant.

The final sample of 50 undergraduate participants from Clemson University had an average age of 18.82 years old, 82% of participants identified as female, 16% as male, and 2% as genderqueer, nonbinary, or genderfluid, 70% of participants identified as White, 16% as Multiracial or Multiethnic, 6% as Hispanic, Latino/a/é, or Spanish, 6% as Black or African American, and 2% as Asian.

Stimuli and Procedures

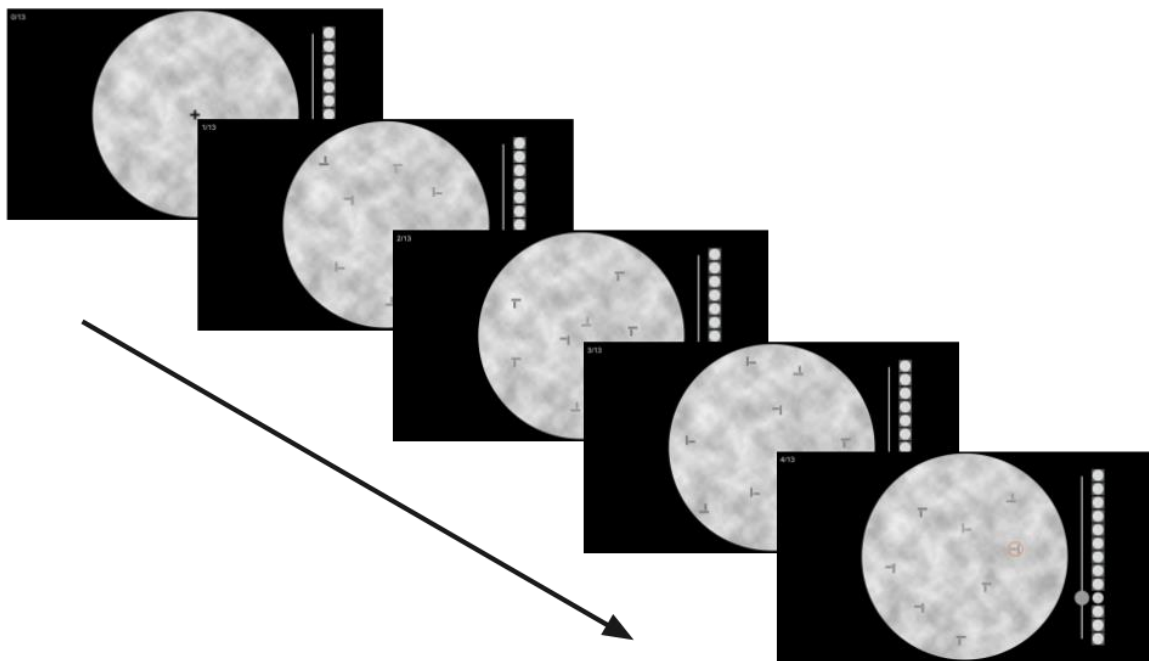
The experiment was programmed and run using PsychoPy (Peirce et al., 2019, 2023). Participants were seated approximately 44.45 cm away from the center of a Dell P2219H, 21.5-in. monitor with a resolution of $1,920 \times 1,080$ pixels and 60-Hz refresh rate. Experiment screens were presented at the same resolution. Participants were instructed to maintain contact between their chin and the provided chinrest throughout the experiment. While the chinrests were set to the same height across participants, adjustable chairs were provided to participants so that they could raise or lower the chair to accommodate different heights while maintaining consistency of head alignment with the monitor.

Search displays combined features of traditional radiology research search displays and CAD displays used by professional readers. Features common across

conditions included a circular primary display, a slider with preview images that visually indicated where participants were in the image set, a layer counter that indicated which layer of the image set the participant was viewing, and a black background (see Figure 2.1). Each image set was composed of 13 layers to approximate industry practice (Sechopoulos & Gheti, 2009; Vedantham et al., 2015). Target-present image sets contained 99 L-shaped distractors and one T-shaped target randomly distributed throughout the 13 layers. Target-absent image sets contained 100 L-shaped distractors randomly distributed throughout the 13 layers. Adobe Photoshop was used to create the cloud background of the primary display as well as the T-shaped targets and L-shaped

Figure 2.1

Navigation Through Image Layers



Note: Depiction of participants moving through layers to find a target in the binary CAD condition.

distractors. The cloud background used a 63-96% white range to mimic radiology imaging (Cain & Mitroff, 2012). The T-shaped targets had a visual angle of $1.3^\circ \times 1.3^\circ$ at the widest points and were composed of rectangles with a width of 0.3° and a centered crossbar. The L-shaped distractors had the same dimensions as the target items but with crossbars offset by 0.3° . Targets and distractors had a slight separation between the perpendicular rectangles and were colored in a range of 47-63% white (Adamo et al., 2018; Cain & Mitroff, 2012). Items had four possible orientations (a rotation of 0° , 90° , 180° , or 270° along the y-axis) for every discrete value in the color range. Item positions, percent white, and rotation were randomized using an algorithm programmed in Python to generate layer, row, and column values in a $13 \times 13 \times 13$ grid within the primary display with five pixels of jitter on the x- and y-axes.

Participants used a slider presented to the right of the primary display to move through the 13 layers. The slider marker, a dark grey circle, could be moved up and down the slider. Thirteen preview images were presented to the right of the slider to provide a visual indication of where the participant was in the image set. Participants were able to move between layers by either moving the slider marker along the slider or clicking on the preview image of the layer they wanted to view. As the participant navigated the layers, the background of the preview image for the layer the participant was on changed from light grey to black. To provide additional layer orientation, the layer number was presented in the top left corner of the search display. Preview images were created in Adobe Photoshop using a blank cloud-filled primary display and black background and

were consistent across layers and conditions to avoid influencing search behavior.

There were four practice trials and 80 experimental trials (84 trials total). Practice was one block with 50% (2) target-present trials. The order of the trials in the practice block remained static across participants and conditions. The experimental trials were separated into two blocks, a high prevalence block and a low prevalence block, with 40 trials in each block. The high prevalence block had a target present in 50% (20) of the image sets. The low prevalence block had a target present in 10% (4) of the image sets. Block order was counterbalanced for the experimental blocks. Trial order within each experimental block remained static across participants and conditions.

Participants were instructed to indicate, via button press, if a target was present (“1”) or absent (“0”) in each image set. Participants were directed to focus on a black fixation cross (1 s) at the beginning of each trial. For the practice trials, participants received feedback on their answers in the form of a green fixation cross if they were correct or a red fixation cross if they were incorrect. Participants did not receive feedback on the experimental trials. During the practice trials, study personnel ensured that participants were aware of how to navigate through the layers using the slider or practice images, knew what the target looked like, and understood how to use the CAD interface, as applicable. Participants were provided with two optional breaks, one after the practice block and one after the first experimental block.

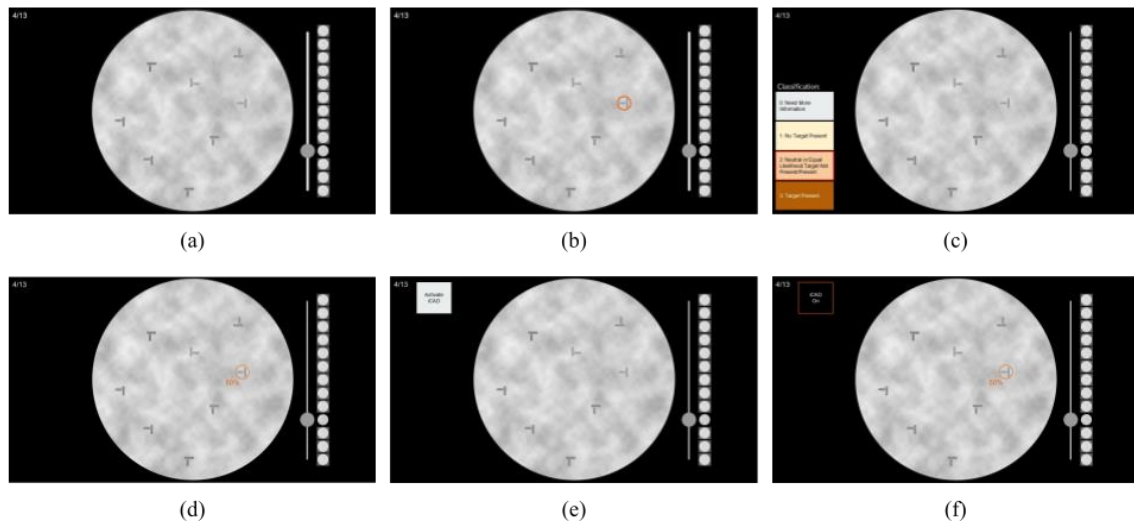
CAD Implementations

Participants were randomly assigned to one of five user interface (UI) conditions: control,

binary CAD, BI-RADS CAD, aCAD, or iCAD. While the UIs for each condition are described in detail below, Figure 2.2 provides side-by-side comparisons of the UIs when there was a target present on the fourth layer. To ensure that the CAD implementations were foundationally equivalent and mimicked real-world CAD systems, all four CAD types were programmed to have a reliability of 80% (Kohli & Saurabh, 2018; Kyono et al., 2018; Parasuraman et al., 1993). To align with the liberal bias of CAD algorithms used in industry (Kunar 2022; Lehman et al., 2015), reliability was implemented so that CAD interfaces were considered unreliable when they incorrectly identified L-shaped distractors as possible targets. CAD aids were also considered unreliable when they indicated an error or did not appropriately produce a CAD cue. While the reliability

Figure 2.2

User Interfaces of Each Condition



Note: The user interfaces for each condition: (a) control; (b) binary CAD; (c) BI-RADS CAD; (d) aCAD; (e) iCAD with the CAD overlay off; and (f) iCAD with the CAD overlay on.

programming for each CAD condition will be expanded on below, Appendix A provides additional information for the reliability programming differences between prevalence conditions for each CAD UI. For all CAD conditions, participants were advised that the CAD system may not be perfectly reliable and that it would be up to the participant to make the final classification of each image set.

The control condition was limited to the base UI of the black background, primary display, layer counter, and slider with preview images, but no additional CAD interface components (see Figure 2.2a). Participants in the control condition should not have been influenced by the programmed reliability of the CAD systems because they did not receive any information from a CAD interface.

The binary CAD interface had the base UI components and aided the participants by circling potential targets (see Figure 2.2b). For this condition, the system was considered reliable when it circled a T-shaped target on target-present trials or did not provide a circled target on target-absent trials and considered unreliable when it circled an L-shaped distractor on target-absent trials. A maximum of one item was circled in each image set.

The BI-RADS CAD interface included the base UI components and provided the participants with a suggested overall rating on a modified BI-RADS scale for each image set (see Figure 2.2c). Participants received additional instruction that the system rated the image set as “0” (need more information; similar to an error message), “1” (no target present), “2” (neutral or equal likelihood of target present), or “3” (target present). The proposed rating was indicated by a red box around the classification the system was

recommending. The rating scale and CAD indicator appeared on all layers for each image set. For this condition, the system was considered reliable when it suggested Category 3 or Category 2 on target-present trials or recommended Category 1 or Category 2 on target-absent trials and unreliable when it suggested Category 0 on target-absent trials (i.e., the system needed more information).

The aCAD interface had the base UI components and aided the participants by both circling potential targets and providing an estimate of how likely the circled item is a target (see Figure 2.2d). For this condition, the system was considered to have been reliable when it circled and provided at least a 50% probability rating for T-shaped targets on target-present trials; circled and provided a 50% probability rating for L-shaped distractors on target-absent trials; or circled and provided a 1-2% probability rating for L-shaped distractors on target-absent trials. The aCAD system was considered to have provided unreliable cues when it did not provide a CAD cue on target-absent trials. A maximum of one item was circled and given a probability rating in each image set.

The iCAD interface had the base UI components with the option for participants to activate an aCAD overlay (see Figure 2.2e) that aided the participants by both circling potential targets and providing an estimate of how likely the circled item is to be a target (see Figure 2.2f). The reliability programming for this condition was the same as for the aCAD condition.

Surveys

Before beginning the experiment and after verifying that they passed the exclusion criteria, participants were directed to complete a pre-study Qualtrics questionnaire. To assess participants' trust in automation in general, they completed the 6-item Propensity to Trust scale (Merritt et al., 2013). For this scale, participants were asked to consider their feelings about automation in general, then rate each statement on a 5-point scale from "Strongly disagree" to "Strongly agree." Items included statements such as "I usually trust machines until there is a reason not to" and "In general, I would rely on machines to assist me." A modified version of this scale replacing "machines" with "automation" was used for this study to better reflect the use of an automated program rather than an automated machine (see Appendix B). The example items above became, "I usually trust automation until there is a reason not to" and "In general, I would rely on automation to assist me." Upon completion of the Propensity to Trust scale, participants were led to the experiment station and instructed to use the chin rest throughout the duration of the study. Study personnel provided additional guidance as needed for the practice trials to ensure that participants understood how to perform the search task (e.g., use the slider to move through the layers; make a determination about target presence for the entire image set rather than each slide; search for a perfect T-shaped target). After completing the four practice trials, participants had the option to take a break, then were instructed to complete the experimental trials.

After completing the experimental trials, participants were redirected to a Qualtrics survey. Demographics information was collected first, including age and gender

identity. To gather information about how participants perceived the usability of each system for further comparisons between the conditions, participants completed the 10-item System Usability Scale (SUS; see Appendix C; Brooke, 1996). For the SUS, participants were asked to consider the UI they had used and rate items on a 5-point scale from “Strongly disagree” to “Strongly agree.” Items included statements such as “I thought the system was easy to use” and “I thought there was too much inconsistency in this system.”

Participants were then asked to consider their trust in the specific UI they used in the study as they completed the 12-item Trust Between People and Automation scale (see Appendix D; Jian et al., 2000). For this scale, participants were asked to select the number from a 7-point scale (1 = “Not at all,” 7 = “Extremely”) that best corresponded to their feelings or impressions about the UI they had used to view and assess the mock radiology images. Items included statements such as “I am suspicious of the system’s intent, action, or outputs” and “The system is reliable.”

Finally, participants were presented with a qualitative survey that asked the participants for their thoughts on the UI they used and what, if any, changes they would make to it (see Appendix E). Items included questions such as “What did you like about the UI?” and “How would you improve the UI?”. Upon completion of the qualitative survey, participants were debriefed, thanked for their participation, and approved for their Sona credits.

CHAPTER THREE

RESULTS

The effectiveness of the CAD implementations was primarily assessed through SDT analyses. Hits were considered when a participant accurately detected a target was present. Correct rejections were considered when a participant accurately determined there was no target present. Misses were considered when a target was present, but the participant did not detect it. False alarms were considered when a target was not present, but the participant determined a target was detected. Hits and false alarms were used to calculate sensitivity (d'), the primary SDT measure of interest when assessing reader performance with and without the aid of a CAD system (Alberdi et al., 2009; Chae et al., 2019; Conant et al., 2019; Fenton et al., 2011; Kunar, 2022; Lehman et al., 2015; Salim et al., 2020; Shoshan et al., 2022). Note, that given only a minority of the trials had a target present, participants' response times were analyzed solely for target-absent trials, in line with previous research exploring visual searches in pseudo-3D images (Adamo et al., 2018).

Differences in Search Performance by Condition and Prevalence

Several 5 x 2 mixed design ANOVAs were performed to evaluate the effects of automation condition (control [no CAD], binary CAD, BI-RADS CAD, aCAD, and iCAD) and prevalence level (low, 10%, and high, 50%) on hit rate, false alarm rate, sensitivity (d'), and target absent response times. The majority of the assumptions for the mixed design ANOVAs were not met: several outliers were found across some of the

automation conditions and prevalence blocks for hit rate and false alarm rate; Shapiro-Wilk tests found that the normality assumption was violated for some of the automation conditions and prevalence levels for all variables; and Levene tests revealed we did not have homogeneity of variances across the automation conditions and prevalence blocks for hit rate and the low prevalence block for sensitivity (see Appendix F). Therefore, robust mixed design ANOVA (Mair & Wilcox, 2023) analyses were used instead.

The robust mixed design ANOVA was performed using the R package *WRS2*, developed by Mair & Wilcox (2023). This package includes several functions for computing robust statistical analyses when the normality and homoscedasticity assumptions are violated, which can raise serious concerns if classical means-based inferential methods are used (e.g., ANOVAs). The function from this package used for computing a between-within subjects ANOVA on the trimmed means is *bwtrim*. The *bwtrim* function uses a between-within subjects ANOVA on the 20% trimmed means. Such percentage-based trims have been suggested to be a good solution to address violations of the normality and homoscedasticity assumptions and produce robust test statistics (Mair & Wilcox, 2023; Field & Wilcox, 2017).

Prior to conducting the mixed design ANOVAs, planned contrasts were performed to explore if participants across the CAD conditions (binary CAD, BI-RADS CAD, aCAD, and iCAD) differed significantly in our outcomes of interest than participants in the control condition.

Control vs. CAD Aids

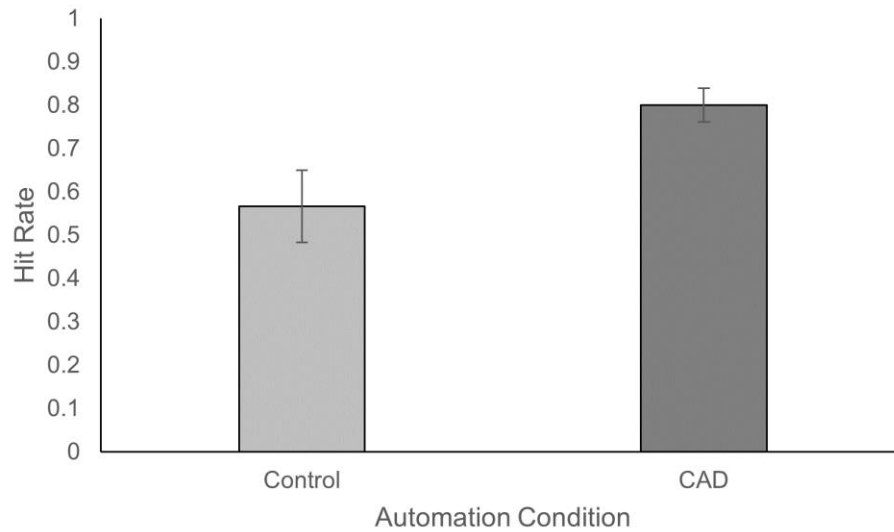
Planned contrasts were conducted to check if there were differences in our measures of interest (hit rate, false alarm rate, sensitivity [d'], and target absent response time) when comparing participants in the control condition to those in the CAD conditions (binary CAD, BI-RADS CAD, aCAD, and iCAD). The results from the planned contrasts indicated that hit rate was significantly higher for the participants who used a CAD system ($M = 0.80$, $SD = 0.25$) than for participants in the control ($M = 0.57$, $SD = 0.26$), $t(45) = -3.22$, $R^2 = .19$, $p < .01$ (see Figure 3.1). However, false alarm rate, sensitivity, and target absent response time were not significantly different for the CAD conditions compared to the control condition, $t(45) = 0.55$, $p = .59$; $t(45) = -1.93$, $p = .06$; and $t(45) = 1.57$, $p = .12$, respectively (see Table 3.1).

Additional planned contrasts were performed to investigate if participants had higher sensitivity (d'), hit rate, and faster target absence response times when using a CAD aid than without one in the low prevalence block. Hit rate in the low prevalence block appeared to be significantly higher for the participants who used a CAD system ($M = 0.75$, $SD = 0.33$) than for participants in the control condition with no CAD assistance ($M = 0.53$, $SD = 0.40$), $t(45) = -2.34$, $R^2 = .45$, $p = .02$ (see Figure 3.2). In contrast, neither sensitivity nor target absent response time were significantly different for the CAD conditions compared to the control condition for the low prevalence block, $t(45) = -1.81$, $p = .08$, and $t(45) = 1.15$, $p = .26$, respectively (see Table 3.2).

Following the planned contrasts, several robust 5 x 2 mixed design ANOVAs were performed to further compare performance between conditions.

Figure 3.1

Comparing Hit Rate for Control Condition vs. CAD Conditions



Note: “CAD” includes binary CAD, BI-RADS CAD, aCAD, and iCAD (error bars show standard errors).

Table 3.1

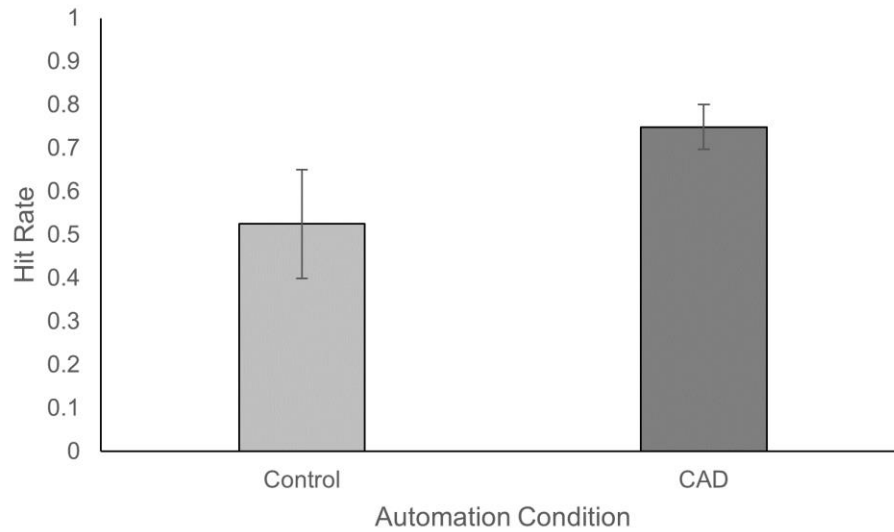
Descriptive Statistics for Control Condition vs. CAD Conditions

Variable	Condition	<i>n</i>	<i>M</i>	<i>SD</i>
Hit Rate	Control	10	.57	.26
	CAD	40	.80	.25
False Alarm Rate	Control	10	.12	.23
	CAD	40	.09	.14
Sensitivity (d')	Control	10	2.37	1.30
	CAD	40	3.25	1.56
Target Absent	Control	10	56.26	13.56
Response Time	CAD	40	43.73	20.14

Note: “CAD” includes binary CAD, BI-RADS CAD, aCAD, and iCAD.

Figure 3.2

Comparing Low Prevalence Block Hit Rate for Control Condition vs. CAD Conditions



Note: “CAD” includes binary CAD, BI-RADS CAD, aCAD, and iCAD (error bars show standard errors).

Table 3.2

Descriptive Statistics for Control Condition vs. CAD Conditions, Low Prevalence

Variable	Condition	<i>n</i>	<i>M</i>	<i>SD</i>
Hit Rate	Control	10	.53	.40
	CAD	40	.75	.33
False Alarm Rate	Control	10	.14	.31
	CAD	40	.09	.15
Sensitivity (<i>d'</i>)	Control	10	2.58	2.22
	CAD	40	3.52	1.95
Target Absent	Control	10	56.82	15.12
Response Time	CAD	40	45.68	22.70

Note: “CAD” includes binary CAD, BI-RADS CAD, aCAD, and iCAD.

Hit Rate

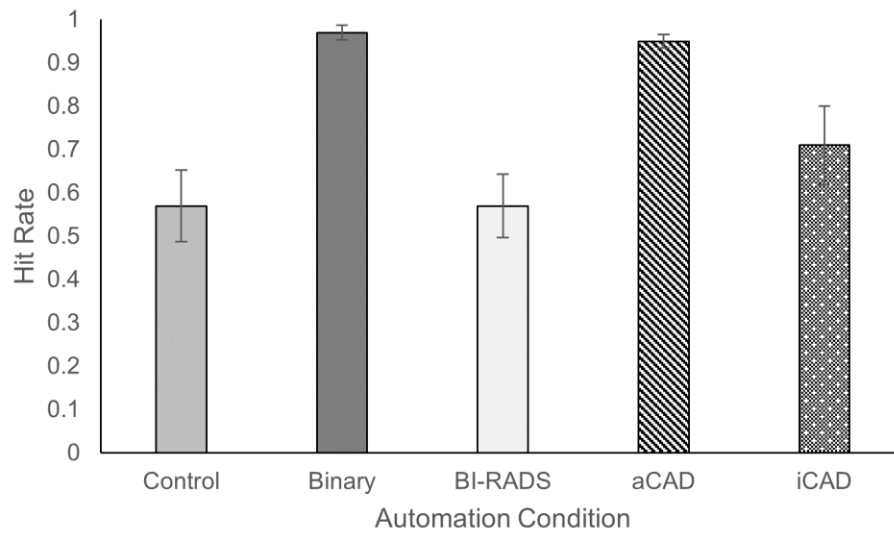
A robust 5 x 2 mixed design ANOVA suggested that there was not a significant interaction effect between automation condition and prevalence level on hit rate, $F(4, 45) = 2.49, p = .10$, or a significant main effect of prevalence level on hit rate, $F(1, 45) = 1.92, p = .19$. However, there was a significant main effect of CAD type on hit rate, $F(4, 45) = 34.99, p < .001, \eta^2 = .42$ (see Figure 3.3). A robust method of post hoc comparisons (Mair & Wilcox, 2023) for differences in hit rate between automation conditions indicated that there was a significant difference between the overall averages for between-subjects group pairwise comparisons, $p < .001$.

To further explore how the CAD types compared regarding hit rate, given that there was no significant interaction found in our robust mixed design ANOVA, a robust method of post hoc comparisons on trimmed means for a one-way ANOVA using a linear contrast expression (Mair & Wilcox, 2023) was performed. The function *lincon* from the R package *WRS2* was used for this analysis. The results indicated that hit rates were significantly better for participants in the binary CAD ($M = 0.97, SD = 0.06$) condition than those in the control ($M = 0.57, SD = 0.26$) or BI-RADS CAD ($M = 0.57, SD = 0.23$) conditions, $\hat{\Psi} = -0.45, p < .01$, and $\hat{\Psi} = 0.48, p < .001$, respectively. Hit rates also appeared to be significantly better for participants in the analog CAD ($M = 0.95, SD = 0.05$) condition than those in the control or BI-RADS CAD conditions, $\hat{\Psi} = -0.43, p < .01$, and $\hat{\Psi} = -0.46, p < .001$, respectively. There were no significant differences in hit rate between binary CAD and analog CAD ($\hat{\Psi} = 0.02, p = .48$), control and BI-RADS CAD

($\hat{\Psi} = 0.03, p = .84$), or iCAD and any of the other automation conditions ($\hat{\Psi}$'s < 0.22 or $\hat{\Psi}$'s $> 0.21, p$'s $> .05$; see Table 3.3).

Figure 3.3

Hit Rate by Automation Condition



Note: Hit rates for each automation condition are shown (error bars show standard errors).

Table 3.3

Descriptive Statistics for Hit Rate by Automation Condition

Condition	<i>n</i>	<i>M</i>	<i>SD</i>
Control	10	.57	.26
Binary CAD	10	.97	.06
BI-RADS CAD	10	.57	.23
aCAD	10	.95	.05
iCAD	10	.71	.29

False Alarm Rate

A robust 5 x 2 mixed design ANOVA suggested that there was not a significant interaction effect between automation condition and prevalence level on false alarm rate, $F(4, 45) = 0.76, p = .57$. There were also no significant main effects of CAD type or prevalence level on false alarm rate, $F(4, 45) = 0.45, p = .77$, and $F(1, 45) = 0.85, p = .37$, respectively (see Table 3.4).

Table 3.4

Descriptive Statistics for False Alarm Rate by Automation Condition

Condition	<i>n</i>	<i>M</i>	<i>SD</i>
Control	10	.12	.23
Binary CAD	10	.04	.03
BI-RADS CAD	10	.07	.10
aCAD	10	.11	.14
iCAD	10	.13	.22

Sensitivity (d')

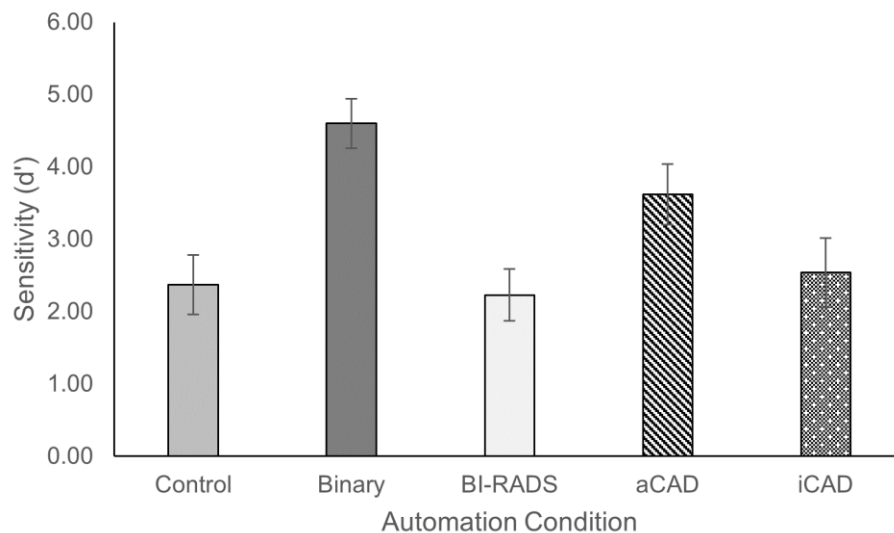
A robust 5 x 2 mixed design ANOVA suggested that there was not a significant interaction effect between automation condition and prevalence level on sensitivity (d'), $F(4, 45) = 1.51, p = .26$, or a significant main effect of prevalence level on sensitivity, $F(1, 45) = 2.22, p = .15$. However, there was a significant main effect of CAD type on sensitivity, $F(4, 45) = 8.48, p < .01, \eta^2 = .33$ (see Figure 3.4). A robust method of post

hoc comparisons (Mair & Wilcox, 2023) for differences in sensitivity between automation condition indicated that there was a significant difference between the overall averages for between-subjects group pairwise comparisons, $p < .001$.

To further explore how the CAD types compared regarding sensitivity, given that there was no significant interaction found in our robust mixed design ANOVA, a robust method of post hoc comparisons for a one-way ANOVA using a linear contrast expression (Mair & Wilcox, 2023) was performed. The results indicated that sensitivity was significantly better for participants in the binary CAD ($M = 4.60, SD = 1.09$) condition than those in the control ($M = 2.37, SD = 1.30$) or BI-RADS CAD ($M = 2.23, SD = 1.14$) conditions, $\hat{\Psi} = -2.63, p < .01$, and $\hat{\Psi} = 2.78, p < .0001$, respectively.

Figure 3.4

Sensitivity (d') by Automation Condition



Note: Sensitivity (d') for each automation condition is shown (error bars show standard errors).

Sensitivity also appeared to be significantly better for participants in the analog CAD ($M = 3.62$, $SD = 1.34$) condition than those in the control or BI-RADS CAD conditions, $\hat{\Psi} = -1.83$, $p = .02$, and $\hat{\Psi} = -1.98$, $p < .01$, respectively. There were no significant differences in sensitivity between binary CAD and analog CAD ($\hat{\Psi} = 0.80$, $p = .29$), control and BI-RADS CAD ($\hat{\Psi} = 0.15$, $p = .79$), or iCAD and any of the other automation conditions ($\hat{\Psi} < -0.55$ or $\hat{\Psi} > 1.27$, p 's $> .05$; see Table 3.4).

Table 3.5

Descriptive Statistics for Sensitivity (d') by Automation Condition

Condition	n	M	SD
Control	10	2.37	1.30
Binary CAD	10	4.60	1.09
BI-RADS CAD	10	2.23	1.14
aCAD	10	3.62	1.34
iCAD	10	2.54	1.52

Target Absent Response Time

A robust 5 x 2 mixed design ANOVA suggested that there was not a significant interaction effect between automation condition and prevalence level on target absent response times, $F(4, 45) = 0.88$, $p = .51$. There were also no significant main effects of CAD type or prevalence level on target absent response times, $F(4, 45) = 2.01$, $p = .16$, and $F(1, 45) = 3.24$, $p = .09$, respectively (see Table 3.6).

Table 3.6*Descriptive Statistics for Target Absent Response Times by Automation Condition*

Condition	<i>n</i>	<i>M</i>	<i>SD</i>
Control	10	56.26	13.56
Binary CAD	10	48.15	19.72
BI-RADS CAD	10	47.51	21.34
aCAD	10	35.96	19.54
iCAD	10	43.30	20.65

Activation of iCAD

A one-way within-subjects ANOVA was performed to evaluate if activation of the iCAD overlay was influenced by whether the participant was in the low prevalence block ($M = 0.492$, $SD = 0.501$) and the high prevalence block ($M = 0.635$, $SD = 0.482$). Our results indicated that the difference in iCAD activation was not statistically significant, $F(1, 9) = 2.45$, $p = .15$. These results suggest that target prevalence does not impact how much participants activate the iCAD overlay.

Trust in the Specific CAD System

Although an ANCOVA was initially proposed to explore if trust in the specific system (i.e., specific trust) differed depending on automation condition when controlling for propensity to trust automation in general (i.e., general trust), an initial calculation of Pearson's correlation coefficient between specific trust and general trust indicated that they were not significantly correlated, $r = -0.16$, $n = 40$, $p = .33$. With the lack of

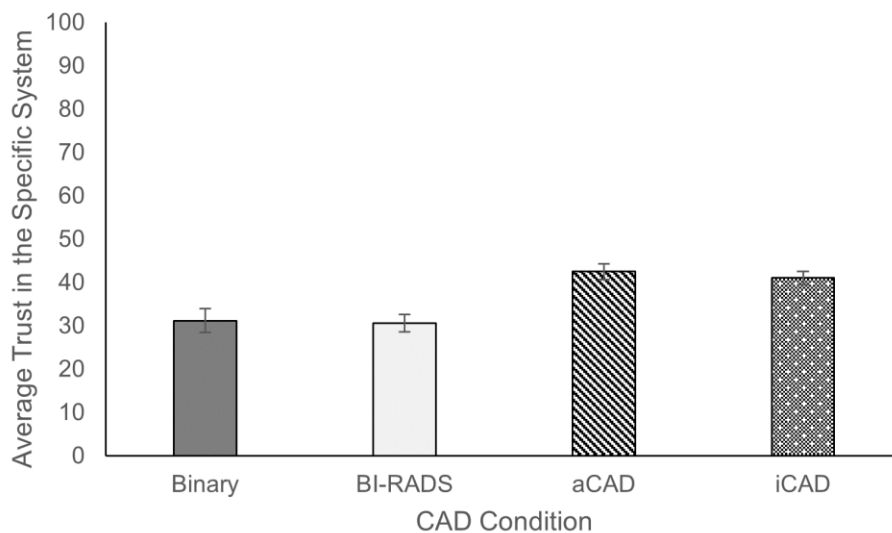
correlation between the intended covariate and dependent variable, this analysis was reconfigured to a one-way between-subjects ANOVA.

The one-way between-subjects ANOVA was performed to evaluate the effect of the automation system (binary CAD, BI-RADS CAD, aCAD, and iCAD) on specific trust in the system. There was a significant difference between specific trust depending on CAD aid, $F(3, 36) = 9.26, p < .001, \eta^2 = .44$. The large effect size indicated that 43.56% of the variance in specific trust could be explained by the type of CAD aid.

To further explore this relationship, post hoc analyses were conducted using Tukey's HSD. The results indicated that participants in the aCAD ($M = 42.57, SD = 5.76$) and iCAD ($M = 41.04, SD = 4.75$) conditions had significantly greater trust in their

Figure 3.5

Average Trust in the Specific System by CAD Condition



Note: Average specific trust scores for each CAD condition are shown (error bars show standard errors).

Table 3.7*Descriptive Statistics for Average Specific Trust by CAD Condition*

Condition	<i>n</i>	<i>M</i>	<i>SD</i>
Binary CAD	10	31.18	8.70
BI-RADS CAD	10	30.69	6.36
aCAD	10	42.57	5.76
iCAD	10	41.04	4.75

specific system than participants in the binary ($M = 31.18$, $SD = 8.70$) and BI-RADS ($M = 30.69$, $SD = 6.36$) conditions ($p < .01$ for all noted comparisons; see Figure 3.5). In contrast, specific trust was not significantly different between participants in the binary and BI-RADS conditions ($p = .998$) or in the aCAD and iCAD conditions ($p = .950$; see Table 3.7).

Perceptions of the Usability of CAD Systems

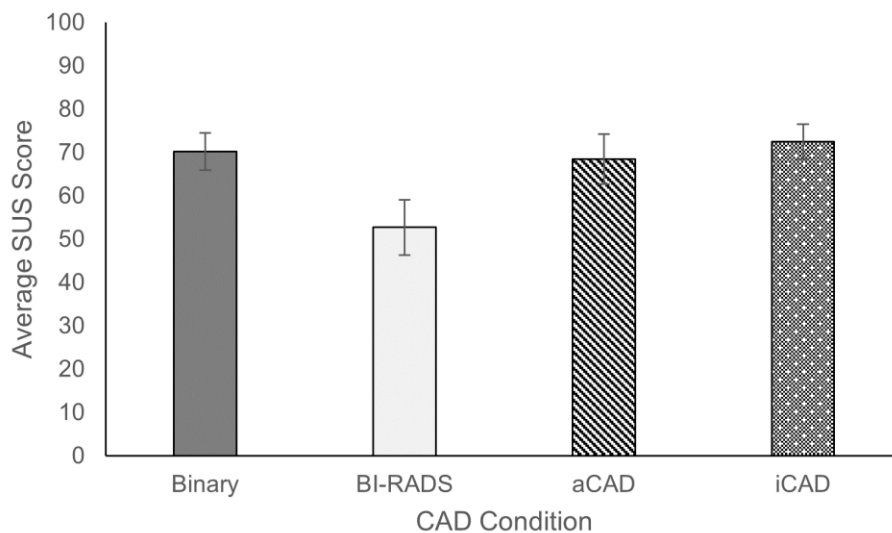
A one-way between-subjects ANOVA was performed to evaluate if SUS scores (i.e., perceived usability) differed depending on automated aid (binary CAD, BI-RADS CAD, aCAD, and iCAD) condition. There was a significant difference between perceived usability for the CAD systems, $F(3, 36) = 2.95$, $p = .045$, $\eta^2 = .20$. The large effect size indicated that 19.75% of the variance in SUS scores could be explained by the type of CAD aid.

With this significant result, planned contrasts based on findings from the literature review and to study our novel BI-RADS CAD aid were conducted to explore if the SUS scores for the binary CAD condition differed from the other CAD systems (BI-RADS

CAD, aCAD, and iCAD); if SUS scores for the BI-RADS CAD condition differed from the other CAD systems (binary CAD, aCAD, and iCAD); and if SUS scores for the aCAD system differed from the iCAD system. The results from the planned contrasts indicated that SUS scores were significantly lower for the BI-RADS CAD system ($M = 52.75$, $SD = 20.15$) than for the other CAD systems (binary CAD [$M = 70.25$, $SD = 13.67$], aCAD [$M = 68.50$, $SD = 18.30$], and iCAD [$M = 72.50$, $SD = 12.86$]), $t(36) = -2.77$, $R^2 = .18$, $p < .01$ (see Figure 3.6). However, SUS scores were not significantly different for the binary CAD system compared to the other CAD implementations, $t(36) = -0.04$, $p = .97$. SUS scores were also not significantly different between the aCAD and iCAD systems, $t(36) = -0.54$, $p = .59$ (see Table 3.8).

Figure 3.6

Average SUS Score by CAD Condition



Note: Average SUS scores for each CAD condition are shown (error bars show standard errors).

Table 3.8*Descriptive Statistics for Average SUS Score by CAD Condition*

Condition	<i>n</i>	<i>M</i>	<i>SD</i>
Binary CAD	10	70.25	13.67
BI-RADS CAD	10	52.75	20.15
aCAD	10	68.50	18.30
iCAD	10	72.50	12.86

Exploratory Analyses*Investigation of CAD UIs: Trust and Usability*

For the exploratory analysis to investigate if participants' trust in the specific automated system related to the perceived usability of that system, Pearson's correlation coefficients were calculated for the SUS and Trust Between People and Automation scores for each CAD condition (binary CAD, BI-RADS CAD, aCAD, and iCAD). None

Table 3.9*Correlations Between SUS and Specific Trust Scores by CAD Condition*

CAD Condition	SUS and Trust Pearson's Correlation	P-Value
Binary CAD	.19	.60
BI-RADS CAD	.58	.08
aCAD	.29	.42
iCAD	.59	.07

of the correlations were significant (p 's > .05; see Table 3.9).

Additionally, a one-way between-subjects ANOVA was performed to check if there were any significant differences in Propensity to Trust scores between CAD conditions (binary CAD, BI-RADS CAD, aCAD, and iCAD). The results indicated that participant's general trust in automation did not differ between the automation conditions, $F(3, 36) = 0.76, p = .52$.

Qualitative Usability Feedback

Responses to the qualitative usability survey regarding participants' impressions of the UI they used were distilled into main ideas for general use perceptions and suggestions for improvement (see Appendix G). Participant feedback was further processed by coding for positive comments (e.g., "It was easy to use"), negative comments (e.g., "My brain couldn't get past what it said sometimes and it felt like it compromised my searching for the target"), and neutral comments (e.g., "It circled the ts that were there but it also circled ls so yes and no"). Comments relating to the task itself instead of the UI (e.g., "Tedious") were also considered as "neutral" comments as the participant was not considered to have felt positively or negatively about the UI itself. Proportions of positive to negative and neutral comments were compiled across the CAD conditions to provide additional context for the qualitative assessment.

After gathering the main themes and coding participant responses for positive, negative, and neutral comments, findings from the qualitative survey reflected the results from the SUS survey. The aCAD ($M = .75, SD = .00$), iCAD ($M = .75, SD = .12$), and

binary CAD ($M = .68$, $SD = .17$) implementations received the highest proportion of positive responses while the mock BI-RADS CAD system ($M = .63$, $SD = .18$) received the lowest proportion of positive comments. Respondents commented that the aCAD, iCAD, and binary CAD aids were easy to use and helped them complete the task more efficiently but did not like having to use the mouse instead of the keyboard to move through the layers, that the circles were off-center for some items (see Appendix H), or the low percentages in the aCAD and iCAD implementations.

It should be noted that upon review of trials with circled items, three circles (4%) were off-center in the binary CAD condition and two (3%) were off-center in the aCAD and iCAD conditions (grouped together because they had the same circle programming). Additionally, average accuracy was calculated for trials with circled items to investigate a measure of reader performance. While average accuracy was below .70 for four trials in the binary CAD condition and four trials in the aCAD and iCAD conditions, only two of the eight identified trials had circles that obscured portions of the enclosed items. The two identified trials with off-center circles were in the aCAD and iCAD conditions, one circling an L-shaped distractor with an indication of 50% likelihood of being a target and the other circling a T-shaped target with an indication of 98% likelihood of being a target. While these trials with off-center circles were also in the binary CAD condition, all but one participant in this condition correctly determined target presence in the image set. Considering this information, off-center circles were unlikely to have meaningfully impacted participant performance.

For the BI-RADS CAD system, participants noted that the system was easy to use

and was a good guide for when the target would be present, but disagreed about whether the CAD aid increased their confidence or made them second-guess themselves.

Unreliability of the system was a common negative comment across all CAD conditions. As an interesting note, the BI-RADS CAD system received more neutral than negative feedback ($M_{neg} = .13$, $SD_{neg} = .13$; $M_{neu} = .25$, $SD_{neu} = .26$) while the aCAD ($M_{neg} = .18$, $SD_{neg} = .12$; $M_{neu} = .08$, $SD_{neu} = .12$), iCAD ($M_{neg} = .18$, $SD_{neg} = .12$; $M_{neu} = .08$, $SD_{neu} = .12$), and binary CAD ($M_{neg} = .18$, $SD_{neg} = .12$; $M_{neu} = .15$, $SD_{neu} = .13$) implementations received more negative than neutral feedback.

Suggestions for CAD aid improvements were primarily to improve the systems' reliability and accuracy. Other suggestions for improvement included changing the colors, providing a better explanation for the percentages, and centering the circles over the items so that it would be easier for participants to determine if the crossbar was perfectly centered.

SDT: Response Criterion

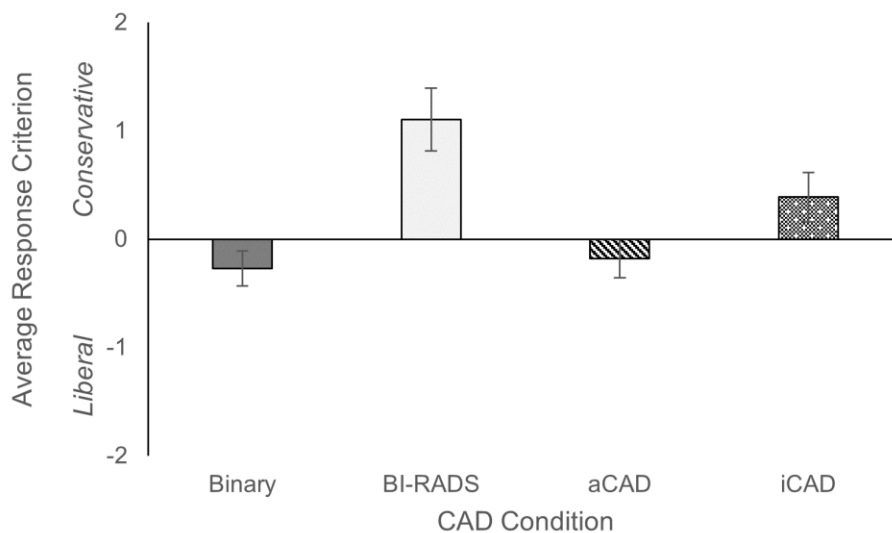
Response criterion was assessed from hits and false alarms to determine how likely a participant was to say that a target is detected. A participant was considered to have a more liberal response criterion if they had a greater number of hits and false alarms (i.e., were more likely to say the target was present than not; a criterion of less than zero), and a more conservative response criterion if they had a greater number of misses and correct rejections (i.e., more likely to say the target was absent than not; a criterion of more than zero). Note that a participant would be considered to be unbiased if they had a criterion of

zero. A one-way between-subjects ANOVA was performed to evaluate if response criterion differed between the four CAD conditions (binary CAD, BI-RADS CAD, aCAD, and iCAD). The results suggested that response criterion changed as a function of automation condition, $F(3, 36) = 8.29, p < .001, \eta^2 = .41$. The effect size indicated that 40.85% of the variance in response criterion could be explained by the type of CAD aid.

To further explore this relationship, post hoc analyses were conducted using Tukey's HSD. The results indicated that participants in the BI-RADS condition ($M = 1.11, SD = 0.92$) had a significantly more conservative response criterion than participants in the binary CAD condition ($M = -0.27, SD = 0.51; p < .001$) and those in the aCAD condition ($M = -0.18, SD = 0.56; p < .01$; see Figure 3.7). A one-sample t-test

Figure 3.7

Average Response Criterion by CAD Condition



Note: Average response criterion for each CAD condition is shown (error bars show standard errors).

found that the criterion shift for participants in the BI-RADS condition was both large and significantly different from a response criterion of zero, $t(9) = 3.78, p < .01, d = 1.20$. While the bar plot indicated that the criterion shift for participants in the iCAD condition may have also been significantly different than zero, a one-sample t-test found that it was not, $t(9) = 1.71, p = .12$.

CHAPTER FOUR

DISCUSSION

Differences in Search Performance by Condition and Prevalence

The primary contributions of this study are the development of a novel CAD system with a BI-RADS implementation and an investigation of how four types of CAD systems (binary CAD, the novel BI-RADS CAD, aCAD, and iCAD) compare on search performance for DBT imaging. Although participants in the present study did not appear to have been impacted by the low prevalence effect (i.e., there were no significant differences in performance between the high and low prevalence blocks for any of the automation conditions in regard to hit rate, false alarm rate, sensitivity [d'], or target absence response time), the use of a CAD aid tended to improve or not harm hit rate and sensitivity. Specifically, participants who used the binary CAD or aCAD systems had higher hit rates and sensitivities than participants in the control or BI-RADS CAD conditions. These results partially support our initial hypothesis (H1) that participants would have higher sensitivity and faster target absent response times when using a CAD interface than without using one. However, further analyses suggested that participants did not have higher sensitivity and faster target absent response times when using a CAD interface than without one in the low prevalence block, failing to support our second hypothesis (H2). There also was no difference on any measure of interest between participants in the iCAD condition and those in the other automation conditions, failing to support our third hypothesis (H3).

Of particular interest in our analyses was the performance of participants in the BI-RADS CAD condition. With attention turning to the possibility of adding BI-RADS ratings to CAD systems, it is important to determine if such information would hinder or aid radiologists in their assessment of breast imaging. If the presentation of a classification has a negative impact on radiologists' diagnostic performance, it would warrant significantly more caution when implementing such ratings or suggest not to implement them at all. The present study indicated that the BI-RADS CAD implementation was not significantly different from at least one of the other CAD conditions (iCAD) on any measure, nor did it lead to participants performing worse than they did in the control condition. This suggests that the addition of a recommended BI-RADS classification for an image set at least does not measurably harm reader performance, supporting continued investigation into the best way to present this information to the reader. The addition of a global BI-RADS rating to CAD systems could significantly improve standardization and reduce intra- and inter-observer variability across radiologists in the interpretation of breast cancer imaging, leading to improved patient outcomes and supporting the goal of the BI-RADS implementation in radiological assessments (Burnside et al., 2009; Magny et al., 2023; Qian et al., 2015). Still, without strong quantifiable support that participants using the BI-RADS CAD system may perform as good as or better than when using other CAD systems, these results are tepid at best; the BI-RADS CAD system requires additional testing to determine if there are superior ways to present the information to assist radiologists in achieving the desired improvements of standardization and reliability of ratings.

With the continued use and improvement of CAD aids in radiology, it is imperative that the different CAD UIs are tested and directly compared with each other to ensure future CAD systems maximize their helpfulness to radiologists. What little past research has been done comparing CAD aids has typically only used two types of CAD implementations (e.g., binary CAD vs. aCAD [Cunningham et al., 2016] or binary CAD vs. iCAD [Drew et al., 2020; Kunar, 2022]) and has not always used direct comparison within the same study but rather comparison to previous research (e.g., Samulski et al., 2010). Even less has been done to investigate how the use of CAD impacts visual search tasks with DBT imaging, which is, in itself, a relatively new area of study. The present study has already contributed to this body of research by comparing four types of CAD aids, three of which are already in use throughout the United States, as well as a novel BI-RADS CAD implementation. While we found that participants performed better with the binary CAD and aCAD systems than with the iCAD system, conflicting with past research (e.g., Drew et al., 2020; Kunar, 2022; Samulski et al., 2010), the present study was the first to explore how different types of CAD aids, including a novel BI-RADS CAD, are used with DBT rather than mammographic images, which may explain some of this discrepancy.

Similarly, while readers in the novel BI-RADS CAD condition did not perform better than those in other CAD conditions, they did not perform worse than those in the control condition, indicating that providing a global classification did not harm reader performance. This is a promising first step for testing future CAD implementations with a global classification. It should be noted that in real world implementations, it is likely that

a global classification such as a suggested BI-RADS rating would be integrated with an aCAD or iCAD system, as these systems have already been proven to benefit radiologist performance and are increasing in popularity (Conant et al., 2019; iCAD, Inc., 2023; Tan et al., 2017; Tan et al., 2015; Qian et al., 2015). Given the neutral impact of our BI-RADS CAD system on reader performance and the recent calls suggesting a BI-RADS classification may help with improving inter- and intra-rater reliability as well as important diagnostic outcomes (e.g., sensitivity), there remains reason to continue exploring how to best integrate BI-RADS ratings with CAD systems and what impact different implementations have on reader performance.

The Lack of a Prevalence Effect

Regarding the benefits of using CAD aids when assessing breast cancer screening imaging, recent research has suggested that the assistance of such automated systems may help to mitigate the prevalence effect, though this largely depends on the CAD implementation (e.g., binary CAD vs. iCAD or the addition of a global classification) and accuracy of the system (particularly regarding false positives; Drew et al., 2020; Kunar, 2022; Tan et al., 2015). It is notable that participants in the present study did not appear to be impacted by the prevalence effect despite many previous studies demonstrating that its influence is very real even when the low prevalence condition in lab studies is designed with a 10% prevalence rather than the 0.5% prevalence level of breast screening images with signs of breast cancer (e.g., Hout et al., 2015; Taylor et al., 2022; Wolfe et al., 2005).

This may be explained by the “scanner” behavior of participants, likely an artifact of the design of the mock DBT image sets. In the little research that has been done comparing how radiologists search 2D images (e.g., mammography) versus pseudo-3D images (e.g., DBT), two dominant search strategies have emerged: “scanning” and “drilling” (Drew et al., 2013; Wen et al., 2016). “Scanners” tend to search each 2D slice in the pseudo-3D image set before moving on to the next one; “drillers” tend to keep their gaze focused on one area of the display as they move through the layers of the pseudo-3D image set before moving on to another area of the display to repeat the same “drilling” behavior. Unlike the Drew et al. (2013), Wen et al. (2016), or Adamo et al. (2018) studies, the present study did not use actual tomography images sources from a medical imaging database or ensure that items were spread across a certain number of slices to provide a sense of the items having depth (e.g., having an increase in clarity of the item as the participants “drilled” closer to it). Additionally, the primary focus of the present study was on the impact and comparison of CAD implementations across search outcomes, so ensuring there was a similar sense of depth in the image sets was not a priority in the creation of the stimuli.

Together, these factors likely encouraged participants to engage in “scanning” behavior, turning each layer into a 2D image search. With this in mind, the target prevalence in our study might be interpreted differently: each block could theoretically be perceived to have 520 2D images, of which 20 would have a target present in the high prevalence block (3.80% prevalence) and four would have a target present in the low prevalence block (0.77% prevalence). From this perspective, both blocks would have

relatively low target prevalence. In other words, there would be essentially no high prevalence block to contrast with because participants would be experiencing the low prevalence effect in both blocks. It should also be noted that some previous research has suggested that the prevalence effect, if it exists in detection tasks, is so small as to not meaningfully affect outcomes or may otherwise be not as important as other factors for search tasks with low prevalence targets (Fleck & Mitroff, 2007). While the lack of a prevalence effect in the present study does not necessarily have a significant impact on the validity of the design of the mock CAD systems, it provides an intriguing addition to the small but growing body of research regarding visual search in pseudo-3D medical imaging.

Activation of iCAD

The defining component of iCAD systems is that the CAD overlay is only presented to participants if they activate it, such as by pressing a button or clicking in an area of interest. Previous research has suggested that this activation by the user is what allows readers to get the most benefit of the CAD system, as they will primarily use it to validate their findings (similar to a second reader, which was the original intent was for the use of CAD) and they will not be distracted by the false positive markings when they do not activate the system (Drew et al., 2020; Du-Crow et al., 2020; Kunar, 2022; Nishikawa & Bae, 2018; Nishikawa & Gur, 2014). However, the lack of activation of the CAD overlay may result in viewers missing the targets that were marked correctly by the iCAD system (Kunar, 2022). In one of the only other studies that investigated the use of

the CAD overlay in an iCAD system, Kunar (2022) found that participants only activated the iCAD aid in 34% of trials. Evaluating how often and when participants activate the iCAD system may provide insight into how to improve its UI to encourage its optimal use in practice.

To that end, the present study explored how prevalence levels might impact iCAD activation. Although participants differed in how often they activated the iCAD overlay, 49.2% in the low prevalence block and 63.5% in the high prevalence block, this was not a significant difference (56.4% activation across all trials). While all calculated iCAD activation percentages were higher than the 34% activation found by Kunar (2022), Kunar's first experiment used modified mammogram images taken from the Digital Database for Screening Mammography (DDSM) as stimuli for 1,000 experimental trials and focused on the use of CAD in low prevalence conditions (10% target-present trials) compared to the present study's use of mock DBT images in which participants searched for a T-shaped target among L-shaped distractors in both high (50% target present) and low (10% target present) prevalence conditions across 80 experimental trials. Such differences in stimuli, number of trials, and prevalence could account for some of the discrepancies in iCAD activation.

Although the present study did not find a significant difference in how often participants activated the iCAD overlay between high and low prevalence blocks, there are some points of interest regarding patterns of iCAD use to note. Some participants never activated the iCAD overlay while other participants activated the CAD overlay for all or almost all of the trials (see Appendix I). Participants who never activated the CAD

overlay may have not felt the need to use the extra assistance it could have provided due to the simplicity of the search task or because they had confidence in their ability to complete the task without the use of the CAD aid. Participants who always activated the CAD overlay may have wanted extra assistance to make the task easier and faster or because they were not as confident in their ability to find the T-shaped target in the image sets. Despite the lack of significance in the results, examining the patterns of activation of the iCAD overlay at the participant level supports observations in existing literature that there are differences in how radiologists use CAD aids when assessing breast screening imaging (i.e., disuse and misuse of the CAD systems; Du-Crow et al., 2019; Jorritsma et al., 2015; Kunar et al., 2017; Nishikawa & Bae, 2018; Nishikawa & Gur, 2014).

Perceived Usability and Trust in the Specific CAD Systems

There has been a dearth of research looking at how the design of the CAD system UIs might impact user perceptions of the system. Key aspects under consideration are typically the system's usability and how much users trust the CAD aid, two components that have been found to influence each other and impact the adoption and use of CAD systems in radiology (Filice & Ratwani, 2020; Hoff & Bashir, 2015; Jorritsma et al., 2015). The present study found that while the binary, aCAD, and iCAD implementations were perceived to be more usable than the BI-RADS CAD system, the aCAD and iCAD systems were perceived as more trustworthy than both the binary and BI-RADS CAD aids. While these results failed to support our hypothesis (H6) that the binary CAD implementation would have lower SUS scores than the other CAD UIs, they provided

partial support for one of our other hypothesis (H5), that the binary CAD system would be perceived as less trustworthy than the other CAD aids when controlling for propensity to trust automation in general. Although these findings do not fully support our hypotheses, they provide a certain amount of support for findings from previous research that supplying more information and having higher perceived reliability may improve the perceived usability of and trust in a system.

Perceived Reliability

Previous research has suggested that trust in an automated aid may be influenced by how reliable the CAD system is perceived to be by the participant. If an automation aid's reliability is perceived to be below a particular threshold (studies suggest 70%), they will be less likely to trust and use it (Parasuraman & Riley, 1997; Wickens & Dixon, 2007). To mimic current CAD reliability and design, the CAD aids were programmed such that participants in the binary CAD condition had the highest proportion of objectively wrong cues (see Appendix A). Participants likely perceived the binary CAD system as unreliable when it circled a distractor, incorrectly indicating it may have been a target. To account for the greater variability in information provided to participants in the BI-RADS CAD, aCAD, or iCAD conditions, these systems were programmed to be considered "unreliable" if the system classified the image set as "Category 0" (i.e., the equivalent of an error message) or did not provide a CAD cue when the target was absent. It should be noted that this difference in programming was an attempt to mimic the real-world reliability of CAD systems and may have artificially influenced trust such

that participants using the binary CAD aid may have had less trust in the system because it may have appeared to be less reliable.

Regarding the perceived trustworthiness of the BI-RADS CAD system, its similarity in programming to the aCAD and iCAD systems suggests that participants' lesser trust in the BI-RADS CAD may have come from its implementation of the global ratings. While the aCAD and iCAD systems used direct and specific cues (i.e., the circling of an object and providing a "target likelihood" score), the BI-RADS CAD system used more indirect guidance (i.e., a suggested classification of the image set). Additionally, some participants said they were confused by the "0" rating of the BI-RADS CAD ("how could it have known the target was present before but then not know if it is present another time?"). With these potential disadvantages when compared to the aCAD and iCAD aids, it is reasonable that participants had less trust in the binary and BI-RADS CAD systems while also not differing in trust between these two aids.

The influence of perceived reliability on perceived trustworthiness of automated aids has been suggested to impact the perceived usability of CAD systems (Filice & Ratwani, 2020; Jorritsma et al., 2015; Nishikawa & Bae, 2018). Further, International Organization for Standardization (ISO) usability standards such as ISO 9241-11:2018 (2023) define usability as how much a system aids users to achieve their goals effectively and efficiently. The binary CAD implementation was not considered as trustworthy as the aCAD and iCAD aids, likely impacted by its presentation of a greater number of false positives to users compared to these CAD systems. It may be surprising, then, that this lack of trust in the system would not correspond to lower SUS scores for the binary CAD

aid. However, recall that participants in the binary CAD and aCAD conditions performed better on measures of hit rate and sensitivity (d') than those in the iCAD and BI-RADS CAD conditions while also not differing significantly in terms of false alarm rate or target absent response time. Despite the perceived deficit in trustworthiness of the binary CAD system, users were still able to achieve their goals successfully and efficiently. The better performance of participants in the binary CAD condition likely contributed to their higher SUS scores and boosted the binary CAD implementation to be perceived as more usable than the BI-RADS CAD system, moving it closer to the aCAD and iCAD systems.

Providing More Information

The lower levels of trust in the binary CAD and BI-RADS CAD systems might also be attributed to the information provided by these systems in comparison to the aCAD and iCAD ones. Research has suggested that providing additional information (e.g., the probability ratings) may make the system more useful to the radiologist and the greater perceived usability may lead to greater trust in these systems (Drew et al., 2020; Filice & Ratwani, 2020; Gao et al., 2019; Jorritsma et al., 2015; Nishikawa & Bae, 2018; Nishikawa & Gur, 2014). While the binary CAD system indicated potential targets with a circle, it was programmed to sometimes circle distractors. Although the aCAD and iCAD systems also circled distractors, they indicated these items had a low likelihood of being a target. This additional information likely helped participants make their determination.

While the BI-RADS CAD aid provided a global classification rating which some participants indicated helped them make a determination (“it helped me redirect my

focus”; “even if it wasn’t 100% accurate or had 100% certainty it still guided me to a conclusion”), it did not have the benefit of the more direct and specific visual cues of the circles and likelihood ratios that were part of the aCAD and iCAD implementations. This broader sort of assistance may have promise for the future in combination with other CAD implementations but did not seem to provide enough information on its own to make a difference in performance or inspire trust.

It is important to note that in a real-world implementation, the BI-RADS rating would likely be integrated into an existing binary CAD, aCAD, or iCAD system as a way to provide even more information so that radiologists can make a BI-RADS determination for each image set (as is current standard practice) rather than a binary one (i.e., saying if a target is present or absent as in the present study). Such a rating on the BI-RADS scale is crucial for guiding patient care and can be challenging to determine even with the use of current CAD aids (i.e., binary CAD, aCAD, or iCAD systems; Berg et al., 2000; Boumaraf et al., 2020; Burnside et al., 2009; Geras et al., 2018; Lazarus et al., 2006; Magny et al., 2023; Melnikow et al., 2016; Obenauer et al., 2005; Pijnappel et al., 2004). While the BI-RADS classification may not have been seen as providing more information in the context of the task for this study, and thus was perceived to be less trustworthy and usable than the aCAD and iCAD systems, it may be considered beneficial information to have in real-world applications when radiologists must determine BI-RADS ratings for DBT image sets to guide patient care.

The Relationship Between Trust and Usability

We have noted throughout this section that previous studies have suggested a connection between users' trust in an automated aid system and how they perceive the usability of that system. Exploratory analyses were performed to verify that trust in specific systems is related to perceived usability across CAD conditions. However, despite an appearance that there may have been a correlation between perceived trustworthiness and usability of a system for at least the aCAD and iCAD systems, the present study found no relationship between participants' trust in their specific CAD system (i.e., Trust Between People and Automation scores) and their perceived usability of that CAD system (i.e., SUS scores). This conflicts with previous research that has suggested a relationship between trust in and usability of automated aids.

The lack of significant correlations between trustworthiness and usability of the CAD systems may have been due to having a small sample size (10 participants for each CAD condition). Although our participant recruitment aligned with that suggested by our power analysis, the effect of the CAD implementations on trust and usability may simply have been too small to be detected within our sample. We can also consider that some previous research has suggested that if a person gets to choose whether they use the system (such as for the iCAD aid or if they ignored the suggested classifications in the corner of the BI-RADS CAD UI), users may have their trust in the system influenced by factors other than usability (e.g., the environment or other contextual information; Acemyan & Kortum, 2012). Additionally, Nielsen (1993) set out that usability encompasses five aspects of a system: learnability, efficiency, memorability, errors, and

satisfaction. While trust and usability were likely impacted by the programmed errors of the CAD aids, participants expressed that the systems were easy to learn, use and understand as well as helped with finding the target faster (see Appendix G). Such responses may have been from the simplicity of the task (finding a “T” amongst “Ls”) or the interfaces. These positive qualities not explicitly related to how much a participant might trust the CAD system may have been factors that influenced the usability ratings, perhaps enough to nullify any underlying relationship between trust and usability that might have existed.

SDT: Response Criterion

The present study found that participants’ response criterion was influenced by the CAD condition they were assigned to, such that participants in the BI-RADS CAD condition were more conservative (i.e., more likely to say the target was absent than not) than participants in the binary CAD or aCAD conditions but did not differ from participants in the iCAD condition. This is mostly contrary to our hypothesis (H8) that response criterion would not be significantly different between the CAD conditions. It is also in contrast to previous research that has indicated that response criterion becomes more conservative with CAD use, particularly when using binary CAD implementations (Drew et al., 2020).

Participants’ response criterion may have been influenced by the design of the CAD systems, specifically by the more direct visual cues in the binary CAD and aCAD systems. If a participant heavily relied on the presence or absence of a circle when

making their determination, given the programming of the study to favor false positives like real-world CAD systems, then they may have been more likely to say that the target was present even when there was no T-shaped target in the image set. In contrast, while participants in the iCAD condition had the option to activate the CAD overlay to provide them with the same information as in the aCAD system, some participants chose not to activate the overlay. By not activating the CAD overlay, participants in the iCAD condition essentially viewed the same display as those in the BI-RADS condition (i.e., there were no circles to indicate potential targets). Without the direct and specific visual indicator to draw their attention to potential targets, participants who did not activate the iCAD overlay could be expected to perform similarly to those in the BI-RADS condition.

Although our results were largely contrary to our initial hypothesis for this exploratory analysis, they provide important insights into how the design of the automation may influence readers' response criterion, which, in the case of breast cancer screenings, can mean the difference between catching signs of cancer in the early stages when it is easier and less expensive to treat rather than later when the consequences can be more severe.

CHAPTER FIVE

LIMITATIONS AND FUTURE DIRECTIONS

One of the primary contributions of the present study was the creation of a novel mock BI-RADS CAD system to begin addressing the recent interest in exploring the integration of the BI-RADS scale with CAD aids to improve standardization in the implementation of BI-RADS, decrease inter- and intra-rater variability, and improve accuracy (Berg et al., 2000; Boumaraf et al., 2020; Geras et al., 2018; Lazarus et al., 2006; Melnikow et al., 2016; Obenauer et al., 2005; Pijnappel et al., 2004; Tan et al., 2017; Tan et al., 2015; Qian et al., 2015). The present study was a first step in testing a BI-RADS CAD implementation and took a broad view into exploring if a suggested global classification for an image set could provide similar outcomes to the CAD systems that use more specific cues (i.e., binary CAD, aCAD, and iCAD). Although participant performance in the BI-RADS condition did not largely differ than those in the control or iCAD conditions, variations of BI-RADS CAD systems may find markedly different results. Additionally, with one of the main goals of a BI-RADS CAD implementation being to improve standardization and decrease inter- and intra-reader variability, future studies should investigate if changing the task to a multi-choice determination (i.e., mimicking how radiologists use the BI-RADS scale to guide patient care rather than simply determining the presence or absence of signs of cancer) that matches the BI-RADS scale used in the BI-RADS CAD design would make the integration of BI-RADS with CAD systems more useful and thereby may produce an improvement in performance that was not seen in the present study.

As noted previously, the present study tested a stand-alone version of a BI-RADS CAD implementation when it is more likely that the BI-RADS scale will be integrated with existing and future aCAD or iCAD systems, thus providing both a global assessment rating of the image set (the suggested BI-RADS classification) and more specific, direct visual cues (e.g., circling potential targets and providing additional information about the circled item, such as a probability rating). Future studies might consider combining the suggested BI-RADS classification with an aCAD or iCAD implementation to get a more holistic picture of how the implementation of these systems might impact the performance of radiologists rather than isolating the global rating aspect.

Another major contribution of the present study was its exploration of visual search performance with pseudo-3D images (i.e., DBT). To improve this aspect of the present study, future research could add a sense of depth to the items across multiple slides in each image set, change how participants move through the layers (e.g., allow scrolling with the mouse wheel), or use actual DBT images from one of the many breast imaging databases to make the pseudo-3D effect feel more realistic. By adding a sense of depth to the images, participants may search the image sets differently, which could also impact their use of the CAD systems.

Given how little research has been done in comparing variations in CAD implementations with DBT imaging, there are many opportunities to extend the present study to continue adding to this body of literature. Further investigations into the presentation of global classifications for image sets could assist with the refinement of the BI-RADS CAD implementation and newer CAD systems that have started extending

the likelihood scores of analog CAD aids to case-level likelihood scores to help radiologists assess the complexity of cases and prioritize their workload (Conant et al., 2019; iCAD, Inc., 2023; Tan et al., 2017; Tan et al., 2015; Qian et al., 2015). Future studies may also explore multiple-target searches to better compare to real-world DBT assessments. For example, BI-RADS ratings are determined based on an assessment of multiple factors such as type of abnormality (e.g., mass, calcification, asymmetry, or architectural distortion) and characteristics of an identified abnormality (e.g., for masses, their size, shape, and margins; American Cancer Society, 2022; Conant et al., 2019; Magny et al., 2022).

Similarly, radiologists would typically have to provide localization information regarding identified abnormalities or areas of concern in real-world tasks. While the present study required a simple case-level determination of target presence, adding a localization component (i.e., requiring the participant to identify where they believe the target to be) would make the task more analogous to real-world assessments, mimic previous research, and provide a better measure of whether participants had correctly identified a target or mis-identified a distractor (Conant et al., 2019; Drew et al., 2013; Kunar, 2022; Wen et al., 2016). Localization could also be implemented in the iCAD system via its activation mechanism: future iterations could require participants to click on an area of suspicion to activate the CAD overlay for that specific area rather than a button that turns it on for the entire image set (Drew et al., 2020; Du-Crow et al., 2020; Hupse et al., 2013; Samulski et al., 2010).

To explore more broad visual search applications, the addition of eye tracking

could add to the literature about scanning and drilling behaviors in pseudo-3D imaging, particularly as studies in this area have typically been done without the use of CAD aids (Drew et al., 2013; Wen et al., 2016). Changes to visual search instructions (e.g., time pressure, providing an expected target prevalence, or explaining different search strategies) have also been shown to influence search behavior; testing changes in instructions could provide insight into how CAD systems might be designed to guide radiologists in their assessment of DBT images (Clark et al., 2014; Cox et al., 2021; Wen et al., 2016). Finally, it could be expected that university students may not be as motivated to find their target(s) as radiologists are to find signs of breast cancer. Future studies could incorporate methods to try to increase participant motivation and effort, such as via a scoring system, to encourage more comparable performance to actual radiologists (Drew et al., 2020; Fleck et al., 2010; Miranda & Palmer, 2013).

CHAPTER SIX

CONCLUSION

The present study explored the effect of four types of CAD systems (binary CAD, aCAD, iCAD, and a novel mock BI-RADS CAD system) on visual search performance, usability perceptions, and trust when evaluating mock DBT imaging. Our results indicated that participants had better hit rate and sensitivity when using the binary CAD and aCAD aids compared to a control condition and the novel BI-RADS CAD condition. Participants in the iCAD condition did not differ significantly from the control or other CAD conditions on any of the measures of interest (hit rate, false alarm rate, sensitivity, or target absent response time), though this could be due to differences in patterns of activation of the iCAD overlay, given that some participants activated the CAD overlay on virtually all trials while other participants rarely, if ever, activated the CAD overlay. Regardless of CAD type or lack of CAD aid, there was no evidence of a prevalence effect on any of the outcomes of interest. This may have been because participants searched each layer as though it was a 2D image, effectively making both blocks low prevalence blocks. Such an effect may also have contributed to a lack of differentiation in activation of the iCAD overlay between the high and low prevalence blocks.

Though trust in a system has previously been shown to be related to its perceived usability, the results of the present study did not fully support this. Participants in the aCAD and iCAD conditions had more trust in their specific automated systems than did participants in the binary CAD and BI-RADS CAD conditions, likely due to the addition of the proportion of likelihood information in the aCAD and iCAD conditions. However,

when it came to perceived usability, participants in the aCAD, iCAD, and binary CAD conditions had higher SUS scores than participants in the BI-RADS CAD condition. This may have been influenced by the greater specificity and directness in the visual CAD indicators of the binary CAD, aCAD, and iCAD systems as opposed to the vaguer global classification rating for each image set provided by the BI-RADS CAD system. Despite the agreement between higher trust and usability ratings for the aCAD and iCAD systems, exploratory analyses of correlations between trust in the specific system and SUS score were not related for any of the tested CAD aids. Such a lack of correlation may have been due to the task being too easy or if the CAD systems were all adequate for the task, but none particularly excelled.

Future research can build on the UIs developed in the present study, particularly for the novel mock BI-RADS CAD aid, to explore variations in presentation of the CAD indicators, the BI-RADS rating scale, the mock DBT images, and other aspects of the display to investigate what and how these changes may impact visual search performance with different types of CAD systems and pseudo-3D imaging. The future of breast cancer screening imaging assessments lies in the use of aCAD and iCAD systems that integrate suggested BI-RADS ratings to further improve radiologists' assessments of DBT imaging; it is imperative that more research is done to better understand how radiologists can take advantage of these CAD systems to produce better outcomes for patients as DBT imaging becomes the primary screening tool for breast cancer.

REFERENCES

- Acemyan, C. Z., & Kortum, P. (2012). The relationship between trust and usability in systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1), 1842-1846. <https://doi.org/10.1177/1071181312561371>
- Adamo, S. H., Ericson, J.M., Nah, J. C., Brem, R., & Mitroff, S. R. (2018). Mammography to tomosynthesis: Examining the differences between two-dimensional and segmented-three-dimensional visual search. *Cognitive Research: Principles and Implications*, 3, article 17. <https://doi.org/10.1186/s41235-018-0103-x>
- Aizenman, A., Drew, T., Ehinger, K. A., Georgian-Smith, D., & Wolfe, J. M. (2017). Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: an eye tracking study. *Journal of Medical Imaging*, 4(4), 045501. <https://doi.org/10.1117/1.JMI.4.4.045501>
- Alberdi, E., Povyakalo, A. A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology*, 11(8), 909-918. <https://doi.org/10.1016/j.acra.2004.05.012>
- Alberdi, E., Povyakalo, A. A., Strigini, L., & Ayton, P. (2009). Computer aided detection: Risks and benefits for radiologists' decisions. In E. Samei & E. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 320 - 332). Cambridge University Press.
- Alagoz, O., Lowry, K. P., Kurian, A. W., Mandelblatt, J. S., Ergun, M. A., Huang, H.,

- Lee, S. J., Schechter, C. B., Tosteson, A. N. A., Miglioretti, D. L., Trentham-Dietz, A., Nyante, S. J., Kerlikowske, K., Sprague, B. L., Stout, N. K., & the CISNET Breast Working Group (2021). Impact of the COVID-19 pandemic on breast cancer mortality in the US: Estimates from collaborative simulation modeling. *Journal of the National Cancer Institute*, *113*(11), 1484–1494.
<https://doi.org/10.1093/jnci/djab097>
- American Cancer Society. (2022, January 14). *What does the doctor look for on a mammogram?* <https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/mammograms/what-does-the-doctor-look-for-on-a-mammogram.html>
- American College of Radiology. (2023). *Tomosynthesis Mammographic Imaging Screening Trial (TMIST)*. <https://www.acr.org/Research/Clinical-Research/TMIST>
- Balleyguier, C., Arfi-Rouche, J., Levy, L., Toubiana, P. R., Cohen-Scali, F., Toledano, A. Y., & Boyer, B. (2017). Improving digital breast tomosynthesis reading time: A pilot multi-reader, multi-case study using concurrent computer-aided detection (CAD). *European Journal of Radiology*, *97*, 83-89.
<https://doi.org/10.1016/j.ejrad.2017.10.014>
- Banerjee, I., Bozkurt, S., Alkim, E., Sagreiya, H., Kurian, A. W., & Rubin, D. L. (2019). Automatic inference of BI-RADS final assessment categories from narrative mammography report findings. *Journal of Biomedical Informatics*, *92*, 103137.
<https://doi.org/10.1016/j.jbi.2019.103137>

- Barazi, H., & Gunduru, M. (2023). Mammography BI RADS grading. In *StatPearls*. StatPearls Publishing.
- Benedikt, R. A., Boatsman, J. E., Swann, C. A., Kirkpatrick, A. D., & Toledano, A. Y. (2018). Concurrent computer-aided detection improves reading time of digital breast tomosynthesis and maintains interpretation performance in a multireader multicase study. *American Journal of Roentgenology*, *210*(3), 685-694.
<https://doi.org/10.1016/j.ejrad.2017.10.014>
- Berg, W. A., Campassi, C., Langenberg, P., & Sexton, M. J. (2000). Breast Imaging Reporting and Data System inter- and intraobserver variability in feature analysis and final assessment. *American Journal of Roentgenology*, *174*(6), 1769-1777.
<https://doi.org/10.2214/ajr.174.6.1741769>
- Bernardi, D., Ciatto, S., Pellegrini, M., Anesi, V., Burlon, S., Cauli, E., Depaoli, M., Larentis, L., Malesani, V., Targa, L., Baldo, P., & Houssami, N. (2012). Application of breast tomosynthesis in screening: Incremental effect on mammography acquisition and reading time. *The British Journal of Radiology*, *85*(1020), e1174–e1178. <https://doi.org/10.1259/bjr/19385909>
- Berry, D. A., Cronin, K. A., Plevritis, S. K., Fryback, D. G., Clarke, L., Zelen, M., Mandelblatt, J. S., Yakovlev, A. Y., Habbema, J. D., Feuer, E. J., & Cancer Intervention and Surveillance Modeling Network (CISNET) Collaborators. (2005). Effect of screening and adjuvant therapy on mortality from breast cancer. *The New England Journal of Medicine*, *353*(17), 1784–1792.
<https://doi.org/10.1056/NEJMoa050518>

- 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. <https://doi.org/10.1016/j.actpsy.2014.08.005>
- Blinder, V. S., & Gany, F. M. (2020). Impact of cancer on employment. *Journal of Clinical Oncology, 38*(4), 302–309. <https://doi.org/10.1200/JCO.19.01856>
- Boumaraf, S., Liu, X., Ferkous, C., & Ma, X. (2020). A New Computer-Aided Diagnosis System with Modified Genetic Feature Selection for BI-RADS Classification of Breast Masses in Mammograms. *BioMed Research International, 2020*, article 7695207. <https://doi.org/10.1155/2020/7695207>
- Bozkurt, S., Gimenez, F., Burnside, E. S., Gulkesen, K. H., & Rubin, D. L. (2016). Using automatically extracted information from mammography reports for decision-support. *Journal of Biomedical Informatics, 62*, 224-231. <https://doi.org/10.1016/j.jbi.2016.07.001>
- Bozkurt, S., Lipson, J. A., Senol, U., & Rubin, D. L. (2015). Automatic abstraction of imaging observations with their characteristics from mammography reports. *Journal of the American Medical Informatics Association, 22*(e1), e81–e92. <https://doi.org/10.1136/amiajnl-2014-003009>
- Burnside, E. S., Sickles, E. A., Bassett, L. W., Rubin, D. L., Lee, C. H., Ikeda, D. M., Mendelson, E. B., Wilcox, P. A., Butler, P. F., & D'Orsi, C. J. (2009). The ACR BI-RADS experience: Learning from history. *Journal of the American College of Radiology, 6*(12), 851–860. <https://doi.org/10.1016/j.jacr.2009.07.023>
- Cain, M. S., & Mitroff, S. R. (2012). Memory for found targets interferes with

- subsequent performance in multiple-target visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 39(5), 1398-1408.
<https://doi.org/10.1037/a0030726>
- Castro, S. M., Tseytlin, E., Medvedeva, O., Mitchell, K., Visweswaran, S., Bekhuis, T., & Jacobson, R. S. (2017). Automated annotation and classification of BI-RADS assessment from radiology reports. *Journal of Biomedical Informatics*, 69, 177-187. <https://doi.org/10.1016/j.jbi.2017.04.011>
- Chae, E. Y., Kim, H. H., Jeong, Jw., Chae, S., Lee, S., & Choi, Y. (2019). Decrease in interpretation time for both novice and experienced readers using a concurrent computer-aided detection system for digital breast tomosynthesis. *European Radiology*, 29, 2518–2525. <https://doi.org/10.1007/s00330-018-5886-0>
- Chan, H. P., Samala, R. K., & Hadjiiski, L. M. (2020). CAD and AI for breast cancer—recent development and challenges. *The British Journal of Radiology*, 93(1108), 20190580. <https://doi.org/10.1259/bjr.20190580>
- Chen, W., & Howe, P. D. (2016). Comparing breast screening protocols: Inserting catch trials does not improve sensitivity over double screening. *PloS one*, 11(10), e0163928. <https://doi.org/10.1371/journal.pone.0163928>
- Chong, A., Weinstein, S. P., McDonald, E. S., & Conant, E. F. (2019). Digital breast tomosynthesis: Concepts and clinical practice. *Radiology*, 292(1), 1-14.
<https://doi.org/10.1001/jamainternmed.2019.1058>
- Clark, K., Cain, M. S., Adcock, R. A., & Mitroff, S. R. (2014). Context matters: The structure of task goals affects accuracy in multiple-target visual search. *Applied*

- Ergonomics*, 45(3), 528–533. <https://doi.org/10.1016/j.apergo.2013.07.008>
- Cole, E. B., Zhang, Z., Marques, H. S., Edward Hendrick, R., Yaffe, M. J., & Pisano, E. D. (2014). Impact of computer-aided detection systems on radiologist accuracy with digital mammography. *American journal of Roentgenology*, 203(4), 909–916. <https://doi.org/10.2214/AJR.12.10187>
- Coleman, C. (2017). Early detection and screening for breast cancer. *Seminars in Oncology Nursing*, 33(2), 141-155. <https://doi.org/10.1016/j.soncn.2017.02.009>
- Conant, E. F., Toledano, A. Y., Periaswamy, S., Fotin, S. V., Go, J., Boatsman, J. E., & Hoffmeister, J. W. (2019). Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiology: Artificial Intelligence*, 1(4), 13. <https://doi.org/10.1186/s41235-022-00361-1>
- Cortez, D., Nascimento, J.C., & Santiago, C. (2021). *Mammogram Classification and Segmentation Through Deep Learning* [Master's thesis, Técnico Lisboa]. Semantic Scholar.
- Cox, P. H., Kravitz, D. J., & Mitroff, S. R. (2021). Great expectations: Minor differences in initial instructions have a major impact on visual search in the absence of feedback. *Cognitive Research: Principles and Implications*, 6(1), 19. <https://doi.org/10.1186/s41235-021-00286-1>
- Cunningham, C.A., Drew, T. & Wolfe, J.M. (2017). Analog computer-aided detection (CAD) information can be more effective than binary marks. *Attention, Perception, & Psychophysics*, 79, 679–690. <https://doi.org/10.3758/s13414-016-1250-0>

Dang, P. A., Freer, P. E., Humphrey, K. L., Halpern, E. F., & Rafferty, E. A. (2014). Addition of tomosynthesis to conventional digital mammography: Effect on image interpretation time of screening examinations. *Radiology*, *270*(1), 49–56. <https://doi.org/10.1148/radiol.13130765>

Division of Cancer Prevention and Control. (2023, July 25). *Breast cancer in men*. Centers for Disease Control and Prevention. <https://www.cdc.gov/cancer/breast/men/index.htm>

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>

Drew, T., Evans, K., Võ, M. L., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in radiology: What can you see in a single glance and how might this guide visual search in medical images? *Radiographics*, *33*(1), 263–274. <https://doi.org/10.1148/rg.331125023>

Drew, T., & Reback, I. (2017). Low target prevalence exacerbates problems with computer-aided detection (CAD) during visual search. *Journal of Vision*, *17*(10), 1125. <https://doi.org/10.1167/17.10.1125>

Drew, T., Vo, M. L., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision*, *13*(10), 3. <https://doi.org/10.1167/13.10.3>

Du-Crow, E., Astley, S. M., & Hulleman, J. (2019). Is there a safety-net effect with computer-aided detection? *Journal of Medical Imaging (Bellingham, Wash.)*,

7(2), 022405. <https://doi.org/10.1117/1.JMI.7.2.022405>

- Du-Crow, E., Astley, S. M., & Hulleman, J. (2020). Suspicious minds: effect of using a lesion likelihood score on reader behaviour with interactive mammographic CAD. In *15th International Workshop on Breast Imaging (IWBI2020)* (115130Y ed., Vol. Proc. SPIE 11513). SPIE. <https://doi.org/10.1117/12.2556472>
- Duffy, S.W., Tabár, L., Yen, A.M.-F., Dean, P.B., Smith, R.A., Jonsson, H., Törnberg, S., Chen, S.L.-S., Chiu, S.Y.-H., Fann, J.C.-Y., Ku, M.M.-S., Wu, W.Y.-Y., Hsu, C.-Y., Chen, Y.-C., Svane, G., Azavedo, E., Grundström, H., Sundén, P., Leifland, K., Frodis, E., Ramos, J., Epstein, B., Åkerlund, A., Sundbom, A., Bordás, P., Wallin, H., Starck, L., Björkgren, A., Carlson, S., Fredriksson, I., Ahlgren, J., Öhman, D., Holmberg, L. and Chen, T.H.-H. (2020), Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer*, 126(13), 2971-2979. <https://doi.org/10.1002/cncr.32859>
- Dustler M. (2020). Evaluating AI in breast cancer screening: A complex task. *The Lancet Digital Health*, 2(3), e106–e107. [https://doi.org/10.1016/S2589-7500\(20\)30019-4](https://doi.org/10.1016/S2589-7500(20)30019-4)
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697–718. [https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7)
- 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. <https://doi.org/10.1371/journal.pone.0064366>

- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.
<https://doi.org/10.3758/BF03193146>
- Fazal, M. I., Patel, M. E., Tye, J., & Gupta, Y. (2018). *The past, present and future role of artificial intelligence in imaging*. *European Journal of Radiology*, *105*, 246–250. <https://doi.org/10.1016/j.ejrad.2018.06.020>
- Fenton, J. J., Abraham, L., Taplin, S. H., Geller, B. M., Carney, P. A., D'Orsi, C., Elmore, J. G., Barlow, W. E., & Breast Cancer Surveillance Consortium. (2011). Effectiveness of computer-aided detection in community mammography practice. *Journal of the National Cancer Institute*, *103*(15), 1152–1161.
<https://doi.org/10.1093/jnci/djr206>
- Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19-38. <https://doi.org/10.1016/j.brat.2017.05.013>
- Filice, R. W., Mongan, J., & Kohli, M. D. (2020). Evaluating artificial intelligence systems to guide purchasing decisions. *Journal of the American College of Radiology*, *17*(11), 1405–1409. <https://doi.org/10.1016/j.jacr.2020.09.045>
- Filice, R. W., & Ratwani, R. M. (2020). The case for user-centered artificial intelligence in radiology. *Radiology: Artificial Intelligence*, *2*(3), e190095.
<https://doi.org/10.1148/ryai.2020190095>
- Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search.

- Psychological Science*, 18(11), 943–947. <https://doi.org/10.1111/j.1467-9280.2007.02006.x>
- 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), 60–71. <https://doi.org/10.1037/a0018629>
- Friedewald, S. M., Rafferty, E. A., Rose, S. L., Durand, M. A., Plecha, D. M., Greenberg, J. S., Hayes, M. K., Copit, D. S., Carlson, K. L., Cink, T. M., Barke, L. D., Greer, L. N., Miller, D. P., & Conant, E. F. (2014). Breast cancer screening using tomosynthesis in combination with digital mammography. *JAMA*, 311(24), 2499–2507. <https://doi.org/10.1001/jama.2014.6095>
- Gandomkar, Z., & Mello-Thoms, C. (2019). Visual search in breast imaging. *The British Journal of Radiology*, 92(1102), article 20190057. <https://doi.org/10.1259/bjr.20190057>
- Gao, Y., Geras, K. J., Lewin, A. A., & Moy, L. (2019). New frontiers: An update on computer-aided diagnosis for breast imaging in the age of artificial intelligence. *American Journal of Roentgenology*, 212(2), 300–307. <https://doi.org/10.2214/AJR.18.20392>
- Geras, K. J., Wolfson, S., Shen, Y., Wu, N., Kim, S. G., Kim, E., Heacock, L., Parikh, U., Moy, L., & Cho, K. (2018). High-resolution breast cancer screening with multi-view deep convolutional neural networks. *ArXiv*. <https://doi.org/10.48550/arXiv.1703.07047>
- Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A.,

- Jemal, A., & Siegel, R. L. (2022). Breast cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(6), 524-541. <https://doi.org/10.3322/caac.21754>
- Godwin, H. J., Menneer, T., Cave, K. R., Thaibsyah, M., & Donnelly, N. (2015). The effects of increasing target prevalence on information processing during visual search. *Psychonomic Bulletin and Review*, 22(2), 469–475. <https://doi.org/10.3758/s13423-014-0686-2>
- Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Peninsula Publishing. (Original work published 1966)
- Gur, D. (2007). Tomosynthesis: Potential clinical role in breast imaging. *American Journal of Roentgenology*, 189(3), 614–615. <https://doi.org/10.2214/AJR.07.2588>
- Gur, D., Sumkin, J. H., Rockette, H. E., Ganott, M., Hakim, C., Hardesty, L., Poller, W. R., Shah, R., & Wallace, L. (2004). Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *Journal of the National Cancer Institute*, 96(3), 185–190. <https://doi.org/10.1093/jnci/djh067>
- Haygood, T. M., Wang, J., Atkinson, E. N., Lane, D., Stephens, T. W., Patel, P., & Whitman, G. J. (2009). Timed efficiency of interpretation of digital and film-screen screening mammograms. *American Journal of Roentgenology*, 192(1), 216–220. <https://doi.org/10.2214/AJR.07.3608>
- Henriksen, E. L., Carlsen, J. F., Vejborg, I. M., Nielsen, M. B., & Lauridsen, C. A. (2019). The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. *Acta Radiologica*,

- 60(1), 13–18. <https://doi.org/10.1177/0284185118770917>
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407-434.
<https://doi.org/10.1177/0018720814547570>
- Houssami, N., Lockie, D., Giles, M., Noguchi, N., Marr, G., & Marinovich, M. L. (2023). Two-year follow-up of participants in the BreastScreen Victoria pilot trial of tomosynthesis versus mammography: Breast density-stratified screening outcomes. *The British Journal of Radiology*, 96(1148), 20230081.
<https://doi.org/10.1259/bjr.20230081>
- Hout, M. C., Walenchok, S. C., Goldinger, S. D., & Wolfe, J. M. (2015). Failures of perception in the low-prevalence effect: Evidence from active and passive visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 41(4), 977–994. <https://doi.org/10.1037/xhp0000053>
- Hupse, R., Samulski, M., Lobbes, M. B., Mann, R. M., Mus, R., den Heeten, G. J., Beijerinck, D., Pijnappel, R. M., Boetes, C., & Karssemeijer, N. (2013). Computer-aided detection of masses at mammography: Interactive decision support versus prompts. *Radiology*, 266(1), 123–129.
<https://doi.org/10.1148/radiol.12120218>
- iCAD, Inc. (2020, January 14). *ProFound AI™ for digital breast tomosynthesis technology from iCAD, Inc.* [Video]. YouTube.
<https://www.youtube.com/watch?v=oYQ09x2HRTs>
- iCAD, Inc. (2023). *ProFound AI® for digital breast tomosynthesis* [Brochure].

- https://www.icadmed.com/wp-content/uploads/2023/11/DMM252_ProFound_AI_Rev_1.pdf
- Indian Radiologist. (2021, June 3). *BIRADS LEXICON / MADHAVI CHANDRA / BREAST IMAGING MASTERCLASS 20201* [Video]. YouTube.
- <https://www.youtube.com/watch?v=Ixf73GAKPnU>
- International Organization for Standardization. (2023). *Ergonomics of human-system interaction—Part 11: Usability: Definitions and concepts* (ISO Standard No. 9241-11:2018). <https://www.iso.org/standard/63500.html>
- Jairam, M. P., & Ha, R. (2022). A review of artificial intelligence in mammography. *Clinical Imaging*, 88, 36–44. <https://doi.org/10.1016/j.clinimag.2022.05.005>
- Jhangiani, N., Philip, M. & Jatoi, I. (2023). Breast cancer screening guidelines: Discrepancies raise concerns about validity. *Breast Cancer*. Advance online publication. <https://doi.org/10.1007/s12282-023-01493-y>
- Jorritsma, W., Cnossen, F., & van Ooijen, P. M. (2014). Merits of usability testing for PACS selection. *International Journal of Medical Informatics*, 83(1), 27–36. <https://doi.org/10.1016/j.ijmedinf.2013.10.003>
- Jorritsma, W., Cnossen, F., & van Ooijen, P. M. (2015). Improving the radiologist-CAD interaction: Designing for appropriate trust. *Clinical Radiology*, 70(2), 115–122. <https://doi.org/10.1016/j.crad.2014.09.017>
- Katzen, J., & Dodelzon, K. (2018). A review of computer aided detection in mammography. *Clinical Imaging*, 52, 305-309. <https://doi.org/10.1016/j.clinimag.2018.08.014>

- Keen, J. D., Keen, J. M., & Keen, J. E. (2018). Utilization of computer-aided detection for digital screening mammography in the United States, 2008 to 2016. *Journal of the American College of Radiology*, 15(1 Pt A), 44-48.
<https://doi.org/10.1016/j.jacr.2017.08.033>
- Kerlikowske, K., Su, Y.-R., Sprague, B. L., Tosteson, A. N. A., Buist, D. S. M., Onega, T., Henderson, L. M., Alsheik, N., Bissell, M. C. S., O'Meara, E. S., Lee, C. I., & Miglioretti, D. L. (2022). Association of screening with digital breast tomosynthesis vs digital mammography with risk of interval invasive and advanced breast cancer. *JAMA*, 327(22), 2220-2230.
<https://doi.org/10.1001/jama.2022.7672>
- Kim, H. E., Kim, H. H., Han, B. K., Kim, K. H., Han, K., Nam, H., Lee, E. H., & Kim, E. K. (2020). Changes in cancer detection and false-positive recall in mammography using artificial intelligence: A retrospective, multireader study. *The Lancet Digital Health*, 2(3), e138–e148. [https://doi.org/10.1016/S2589-7500\(20\)30003-0](https://doi.org/10.1016/S2589-7500(20)30003-0)
- Kohli, A., & Saurabh, J. (2018). Why CAD failed in mammography. *Journal of the American College of Radiology*, 15(3), 535-537.
<https://doi.org/10.1016/j.jacr.2017.12.029>
- Kunar, M. A. (2022). The optimal use of computer aided detection to find low prevalence cancers. *Cognitive Research: Principles and Implications*, 7, 13.
<https://doi.org/10.1186/s41235-022-00361-1>
- 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), 369–385. <https://doi.org/10.1037/xap0000132>
- Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search. *Radiology*, 116(3), 527-532. <https://doi.org/10.1148/116.3.527>
- Kyono, T., Gilbert, F. J., & van der Schaar, M. (2018). MAMMO: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. *ArXiv*. <https://doi.org/10.48550/arXiv.1811.02661>
- Lam Shin Cheung, J., Ali, A., Abdalla, M., & Fine, B. (2023). U"AI" testing: User interface and usability testing of a chest x-ray AI tool in a simulated real-world workflow. *Canadian Association of Radiologists Journal*, 74(2), 314–325. <https://doi.org/10.1177/08465371221131200>
- Lazarus, E., Mainiero, M. B., Schepps, B., Koelliker, S. L., & Livingston, L. S. (2006). BI-RADS lexicon for US and mammography: Interobserver variability and positive predictive value. *Radiology*, 239(2), 385-391. <https://doi.org/10.1148/radiol.2392042127>
- Lee J. D. (2008). Review of a pivotal human factors article: "Humans and automation: Use, misuse, disuse, abuse." *Human Factors*, 50(3), 404–410. <https://doi.org/10.1518/001872008X288547>
- Lee, J. H., Kim, K. H., Lee, E. H., Ahn, J. S., Ryu, J. K., Park, Y. M., Shin, G. W., Kim, Y. J., & Choi, H. Y. (2022). Improving the performance of radiologists using artificial intelligence-based detection support software for mammography: A multi-reader study. *Korean Journal of Radiology*, 23(5), 505–516.

<https://doi.org/10.3348/kjr.2021.0476>

- Lee, S. E., Kim, G. R., Yoon, J. H., Han, K., Son, W. J., Shin, H. J., & Moon, H. J. (2023). Artificial intelligence assistance for women who had spot compression view: Reducing recall rates for digital mammography. *Acta Radiologica*, *64*(5), 1808–1815. <https://doi.org/10.1177/02841851221140556>
- Lehman, C. D., Arao, R. F., Sprague, B. L., Lee, J. M., Buist, D. S., Kerlikowske, K., Henderson, L. M., Onega, T., Tosteson, A. N., Rauscher, G. H., & Miglioretti, D. L. (2017). National performance benchmarks for modern screening digital mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology*, *283*(1), 49–58. <https://doi.org/10.1148/radiol.2016161174>
- Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine*, *175*(11), 1828–1837. <https://doi.org/10.1001/jamainternmed.2015.5231>
- Lekadir, K., Feragen, A., Fofanah, A. J., Frangi, A. F., Buyx, A., Emelie, A., Lara, A., Porras, A. R., Chan, A.-W., Navarro, A., Glocker, B., Botwe, B. O., Khanal, B., Beger, B., Wu, C. C., Cintas, C., Langlotz, C. P., Rueckert, D., Mzurikwao, D., ... Starmans, M. P. A. (2023). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *ArXiv*. <https://doi.org/10.48550/arXiv.2309.12325>
- Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., Aussó, S., Cerdá Alberich, L., Marias, K., Tsiknakis, M., Colantonio, S., Papanikolaou, N.,

- Salahuddin, Z., Woodruff, H. C., Lambin, P., & Martí-Bonmatí, L. (2021). FUTURE-AI: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. *ArXiv*.
<https://doi.org/10.48550/arXiv.2109.09658>
- Liberman, L., & Menell, J. H. (2002). Breast imaging reporting and data system (BI-RADS). *Radiologic Clinics*, 40(3), 409-430. [https://doi.org/10.1016/S0033-8389\(01\)00017-3](https://doi.org/10.1016/S0033-8389(01)00017-3)
- Lowry, K. P., Trentham-Dietz, A., Schechter, C. B., Alagoz, O., Barlow, W. E., Burnside, E. S., Conant, E. F., Hampton, J. M., Huang, H., Kerlikowske, K., Lee, S. J., Miglioretti, D. L., Sprague, B. L., Tosteson, A. N. A., Yaffe, M. J., & Stout, N. K. (2020). Long-term outcomes and cost-effectiveness of breast cancer screening with digital breast tomosynthesis in the United States. *Journal of the National Cancer Institute*, 112(6), 582–589. <https://doi.org/10.1093/jnci/djz184>
- Magny, S. J., Shikhman, R., & Keppke, A. L. (2023). Breast Imaging Reporting and Data System. In *StatPearls*. StatPearls Publishing.
- Mahfouz, M. (2016, February 7). *Mammography BIRADS lexicon - Prof Dr. Rasha Kamal (In Arabic)* [Video]. YouTube.
https://www.youtube.com/watch?v=tM_yqIJ1oeQ
- Mair, P., & Wilcox, R. (2023). *Robust statistical methods using WRS2*. The Comprehensive R Archive Network. <https://cran.r-project.org/web/packages/WRS2/vignettes/WRS2.pdf>
- Mann, R. M., Hooley, R., Barr, R. G., & Moy, L. (2020). Novel approaches to screening

- for breast cancer. *Radiology*, 297(2), 266-285.
<https://doi.org/10.1148/radiol.2020200172>
- Marinovich, M. L., Hunter, K. E., Macaskill, P., & Houssami, N. (2018). Breast cancer screening using tomosynthesis or mammography: A meta-analysis of cancer detection and recall. *Journal of the National Cancer Institute*, 110(9), 942–949.
<https://doi.org/10.1093/jnci/djy121>
- Masud, R., Al-Rei, M., & Lokker, C. (2019). Computer-aided detection for breast cancer screening in clinical settings: Scoping review. *JMIR Medical Informatics*, 7(3), e12660. <https://doi.org/10.2196/12660>
- Melnikow, J., Fenton, J. J., Whitlock, E. P., Miglioretti, D. L., Weyrich, M. S., Thompson, J. H., & Shah, K. (2016). Supplemental screening for breast cancer in women with dense breasts: A systematic review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 164, 268-278.
<https://doi.org/10.7326/M15-1789>
- Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, 55(3), 520–534.
<https://doi.org/10.1177/0018720812465081>
- Miranda, A. T., & Palmer, E. M. (2014). Intrinsic motivation and attentional capture from gamelike features in a visual search task. *Behavior Research Methods*, 46(1), 159–172. <https://doi.org/10.3758/s13428-013-0357-7>
- Mushlin, A. I., Kouides, R. W., & Shapiro, D. E. (1998). Estimating the accuracy of

- screening mammography: a meta-analysis. *American Journal of Preventive Medicine*, 14(2), 143–153. [https://doi.org/10.1016/s0749-3797\(97\)00019-6](https://doi.org/10.1016/s0749-3797(97)00019-6)
- Myers, E. R., Moorman, P., Gierisch, J. M., Havrilesky, L. J., Grimm, L. J., Ghatge, S., Davidson, B., Montgomery, R. C., Crowley, M. J., McCrory, D. C., Kendrick, A., & Sanders, G. D. (2015). Benefits and harms of breast cancer screening: A systematic review. *JAMA*, 314(15), 1615–1634. <https://doi.org/10.1001/jama.2015.13183>
- Narayan, A. K., Lee, C. I., & Lehman, C. D. (2020). Screening for breast cancer. *Medical Clinics of North America*, 104(6), 1007-1021. <https://doi.org/10.1016/j.mcna.2020.08.003>
- Narváez, F., Díaz, G., Poveda, C., & Romero, E. (2017). An automatic BI-RADS description of mammographic masses by fusing multiresolution features. *Expert Systems with Applications*, 74, 82-95. <https://doi.org/10.1016/j.eswa.2016.11.031>
- Nelson, H. D., Fu, R., Cantor, A., Pappas, M., Daeges, M., & Humphrey, L. (2016). Effectiveness of breast cancer screening: Systematic review and meta-analysis to update the 2009 U.S. Preventive Services Task Force recommendation. *Annals of Internal Medicine*, 164(4), 244–255. <https://doi.org/10.7326/M15-0969>
- Nielsen, J. (1993). *Usability engineering*. Elsevier Inc. <https://doi.org/10.1016/C2009-0-21512-1>
- Nishikawa, R. M., & Bae, K. T. (2018). Importance of better human-computer interaction in the era of deep learning: Mammography computer-aided diagnosis as a use case. *Journal of the American College of Radiology*, 15(1 Pt A), 49-52.

- <https://doi.org/10.1016/j.jacr.2017.08.027>
- Nishikawa, R. M., & Gur, D. (2014). CADe for early detection of breast cancer-current status and why we need to continue to explore new approaches. *Academic Radiology*, *21*(10), 1320–1321. <https://doi.org/10.1016/j.acra.2014.05.018>
- Obenauer, S., Hermann, K. P., & Grabbe, E. (2005). Applications and literature review of the BI-RADS classification. *European Radiology*, *15*, 1027-1036. <https://doi.org/10.1007/s00330-004-2593-9>
- Oeffinger, K. C., Fontham, E. T., Etzioni, R., Herzig, A., Michaelson, J. S., Shih, Y. C., Walter, L. C., Church, T. R., Flowers, C. R., LaMonte, S. J., Wolf, A. M., DeSantis, C., Lortet-Tieulent, J., Andrews, K., Manassaram-Baptiste, D., Saslow, D., Smith, R. A., Brawley, O. W., Wender, R., & American Cancer Society. (2015). Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. *JAMA*, *314*(15), 1599–1614. <https://doi.org/10.1001/jama.2015.12783>
- Pace, L. E., & Keating, N. L. (2014). A systematic assessment of benefits and risks to guide breast cancer screening decisions. *JAMA*, *311*(13), 1327-1335. <https://doi.org/10.1001/jama.2014.1398>
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced “complacency.” *The International Journal of Aviation Psychology*, *3*(1), 1-23. https://doi.org/10.1207/s15327108ijap0301_1
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

- <https://doi.org/10.1518/001872097778543886>
- 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*, *51*(1), 195–203.
- <https://doi.org/10.3758/s13428-018-01193-y>
- Peirce, J. W., MacAskill, M. R., Höchenberger, R., Bridges, D., Gray, J. R., Simpson, S., Lindeløv, J., Sogo, H., Halchenko, Y., Kastman, E., Hogman, W., & Ilixa Ltd. (2023). *PsychoPy* (Version 2023.2.1) [Software]. Open Science Tools, Ltd.
- <https://psychopy.org/index.html>
- Pijnappel, R. M., Peeters, P. H., Hendriks, J. H., & Mali, W. P. (2004). Reproducibility of mammographic classifications for non-palpable suspect lesions with microcalcifications. *The British Journal of Radiology*, *77*(916), 312–314.
- <https://doi.org/10.1259/bjr/84593467>
- Pisano, E. D., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J. K., Acharyya, S., Conant, E. F., Fajardo, L. L., Bassett, L., D'Orsi, C., Jong, R., Rebner, M., & Digital Mammographic Imaging Screening Trial (DMIST) Investigators Group (2005). Diagnostic performance of digital versus film mammography for breast-cancer screening. *The New England Journal of Medicine*, *353*(17), 1773–1783.
- <https://doi.org/10.1056/NEJMoa052911>
- Qian, W., Sun, W., & Zheng, B. (2015). Improving the efficacy of mammography screening: the potential and challenge of developing new computer-aided detection approaches. *Expert Review of Medical Devices*, *12*(5), 497–499.

- <https://doi.org/10.1586/17434440.2015.1068115>
- Richman, I. B., Hoag, J. R., Xu, X., Forman, H. P., Hooley, R., Busch, S. H., & Gross, C. P. (2019). Adoption of digital breast tomosynthesis in clinical practice. *JAMA Internal Medicine*, *179*(9), 1292–1295.
- <https://doi.org/10.1001/jamainternmed.2019.1058>
- Rocha García, A. M., & Mera Fernández, D. (2019). Breast tomosynthesis: State of the art. *Radiología*, *61*(4), 274-285. <https://doi.org/10.1016/j.rxeng.2019.03.008>
- RSNA. (2022, May 2). *Advances in artificial intelligence in breast health (2022)* [Video]. YouTube. <https://www.youtube.com/watch?v=H-B-1qENcdc>
- Ryser, M. D., Lange, J., Inoue, L. Y. T., O'Meara, E. S., Gard, C., Miglioretti, D. L., Bulliard, J. L., Brouwer, A. F., Hwang, E. S., & Etzioni, R. B. (2022). Estimation of breast cancer overdiagnosis in a U.S. breast screening cohort. *Annals of Internal Medicine*, *175*(4), 471–478. <https://doi.org/10.7326/M21-3577>
- Salim, M., Dembrower, K., Eklund, M., Lindholm, P., & Strand, F. (2020). Range of radiologist performance in a population-based screening cohort of 1 million digital mammography examinations. *Radiology*, *297*(1), 33–39.
- <https://doi.org/10.1148/radiol.2020192212>
- 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. <https://doi.org/10.1007/s00330-010-1821-8>
- Schünemann, H. J., Lerda, D., Quinn, C., Follmann, M., Alonso-Coello, P., Rossi, P. G.,

- Lebeau, A., Nyström, L., Broeders, M., Ioannidou-Mouzaka, L., Duffy, S. W., Borisch, B., Fitzpatrick, P., Hofvind, S., Castells, X., Giordano, L., Canelo-Aybar, C., Warman, S., Mansel, R., ... European Commission Initiative on Breast Cancer (ECIBC) Contributor Group. (2020). Breast cancer screening and diagnosis: A synopsis of the European Breast Guidelines. *Annals of Internal Medicine*, 172(1), 46–56. <https://doi.org/10.7326/M19-2125>
- Sechopoulos, I., & Ghetti, C. (2009). Optimization of the acquisition geometry in digital tomosynthesis of the breast. *Medical Physics*, 36(4), 1199-1207. <https://doi.org/10.1118/1.3090889>
- Shoshan, Y., Bakalo, R., Gilboa-Solomon, F., Ratner, V., Barkan, E., Ozery-Flato, M., Amit, M., Khapun, D., Ambinder, E. B., Oluyemi, E. T., Panigrahi, B., DiCarlo, P. A., Rosen-Zvi, M., & Mullen, L. A. (2022). Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis. *Radiology*, 303(1), 69–77. <https://doi.org/10.1148/radiol.211105>
- Sickles, E. A., D’Orsi, C. J., Bassett, L. W., et al. ACR BI-RADS® Mammography. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. Reston, VA, American College of Radiology; 2013.
- Siegel, R. L., Miller, K. D., Sandeep Wagle, N., & Jemal, A. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1), 17-48. <https://doi.org/10.3322/caac.21763>
- Sippo, D. A., Warden, G. I., Andriole, K. P., Lacson, R., Ikuta, I., Birdwell, R. L., & Khorasani, R. (2013). Automated extraction of BI-RADS final assessment

- categories from radiology reports with natural language processing. *Journal of Digital Imaging*, 26, 989-994. <https://doi.org/10.1007/s10278-013-9616-5>
- Siu, A. L., & U.S. Preventive Services Task Force. (2016). Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, 164(4), 279–296. <https://doi.org/10.7326/M15-2886>
- Sun, L., Legood, R., dos-Santos-Silva, I., Gaiha, S. M., & Sadique, Z. (2018) Global treatment costs of breast cancer by stage: A systematic review. *PLoS ONE*, 13(11), e0207993. <https://doi.org/10.1371/journal.pone.0207993>
- Tan, M., Aghaei, F., Wang, Y., & Zheng, B. (2017). Developing a new case based computer-aided detection scheme and an adaptive cueing method to improve performance in detecting mammographic lesions. *Physics in Medicine and Biology*, 62(2), 358–376. <https://doi.org/10.1088/1361-6560/aa5081>
- Tan, M., Qian, W., Pu, J., Liu, H., & Zheng, B. (2015). A new approach to develop computer-aided detection schemes of digital mammograms. *Physics in Medicine and Biology*, 60(11), 4413–4427. <https://doi.org/10.1088/0031-9155/60/11/4413>
- Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6), 401–409. <https://doi.org/10.1037/h0058700>
- Taylor, J. E. T., Hilchey, M. D., Weidler, B. J., & Pratt, J. (2022). Eliminating the low-prevalence effect in visual search with a remarkably simple strategy. *Psychological Science*, 33(5), 716–724. <https://doi.org/10.1177/09567976211048485>
- 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. <https://doi.org/10.1016/j.ejca.2008.02.016>
- Trieu, P. D. Y., Lewis, S. J., Li, T., Ho, K., Tapia, K. A., & Brennan, P. C. (2020). Reader characteristics and mammogram features associated with breast imaging reporting scores. *The British Journal of Radiology*, 93(1114), 20200363. <https://doi.org/10.1259/bjr.20200363>
- US Food and Drug Administration. Summary of Safety and Effectiveness Data. R2 Technologies (P970058) 1998. P970058. <http://www.fda.gov/ohrms/dockets/98fr/123098b.txt>
- Vedantham, S., Karellas, A., Vijayaraghavan, G. R., & Kopans, D. B. (2015). Digital breast tomosynthesis: State of the art. *Radiology*, 277(3), 663-684. <https://doi.org/10.1148/radiol.2015141303>
- Wen, G., Aizenman, A., Drew, T., Wolfe, J. M., Haygood, T. M., & Markey, M. K. (2016). Computational assessment of visual search strategies in volumetric medical images. *Journal of Medical Imaging*, 3(1), 015501. <https://doi.org/10.1117/1.JMI.3.1.015501>
- Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, 8(3), 201–212. <https://doi.org/10.1080/14639220500370105>
- Wolfe, J., Horowitz, T. & Kenner, N. (2005). Rare items often missed in visual searches. *Nature*, 435, 439–440. <https://doi.org/10.1038/435439a>

Zuckerman, S. P., Conant, E. F., Keller, B. M., Maidment, A. D., Barufaldi, B.,

Weinstein, S. P., Synnestvedt, M., & McDonald, E. S. (2016). Implementation of synthesized two-dimensional mammography in a population-based digital breast tomosynthesis screening program. *Radiology*, *281*(3), 730–736.

<https://doi.org/10.1148/radiol.2016160366>

APPENDICES

Appendix A

Reliability-inspired Programming by Prevalence and CAD Conditions

Table A.1 shows how the four CAD systems presented information to the participant based on whether a target was present or absent, the prevalence condition of the experimental block, and inspired by past research using an overall reliability of 80%.

Table A.1

Reliability-inspired Programming by Prevalence and CAD Conditions

	Low Prevalence			High Prevalence		
	Total Trials	# Trials Present	# Trials Absent	Total Trials	# Trials Present	# Trials Absent
Binary CAD						
No circle	28	0	28	12	0	12
Circle	12	4	8	28	20	8
BI-RADS CAD						
Category 0	8	0	8	8	0	8
Category 1	26	0	26	7	0	7
Category 2	4	2	2	10	5	5
Category 3	2	2	0	15	15	0
aCAD and iCAD						
No % or circle	8	0	8	8	0	8
1-2% + circle	26	0	26	7	0	7
50% + circle	4	2	2	10	5	5
96-99% + circle	2	2	0	15	15	0

Appendix B

Propensity to Trust

Merritt et al., 2013

Instructions: Thinking about your feelings about automation in general, please rate the following items from *Strongly Agree* to *Strongly Disagree*:

1. I usually trust automation until there is a reason not to.
2. For the most part, I distrust automation.
3. In general, I would rely on automation to assist me.
4. My tendency to trust automation is high.
5. It is easy for me to trust automation to do its job.
6. I am likely to trust automation even when I have little knowledge about it.

Appendix C

System Usability Scale (SUS)

Brooke, 1996

Instructions: Thinking about the user interface you just used to view the images, please rate the following items from *Strongly Agree* to *Strongly Disagree*:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Appendix D

Trust Between People and Automation

Jian et al., 2000

Instructions: Below is a list of statements for evaluating trust between people and automation. Please select the number that best corresponds to your feelings or impressions about the user interface you just used to view and rate the images (1 = *Not at all*, 7 = *Extremely*):

1. The system is deceptive.
2. The system behaves in an underhanded manner.
3. I am suspicious of the system's intent, action, or outputs.
4. I am wary of the system.
5. The system's actions will have a harmful or injurious outcome.
6. I am confident in the system.
7. The system provides security.
8. The system has integrity.
9. The system is dependable.
10. The system is reliable.
11. I can trust the system.
12. I am familiar with the system.

Appendix E

Qualitative Usability Questionnaire

Instructions: Thinking about the user interface you just used to view the images, please answer the following questions:

1. Did you have a positive experience using the user interface (UI)?
2. What did you like about the UI?
3. What did you not like about the UI?
4. Did you feel like the UI aided you in your task?
5. How would you improve the UI?

Appendix F

Assumptions for the 5x2 Mixed Design ANOVAs

Tables F.1 and F.2 show the results of tests for the assumptions of the 5x2 mixed design ANOVAs that were used to explore the first three hypotheses (i.e., differences in search performance by condition and prevalence). The normality assumption (see Table F.1) was tested using the Shapiro-Wilk test and the homoscedasticity assumption (see Table F.2) was tested using Levene's test. It should be noted that means and standard deviations were the same for hit rate for all participants in the binary CAD condition for the low prevalence block.

Table F.1

Normality Assumption for the 5x2 Mixed Design ANOVAs

Variable	Condition	Prevalence		Shapiro-Wilk	
		Block	<i>n</i>	Statistic	P-Value
Hit Rate	Control	Low	10	0.88	.14
		High	10	0.88	.14
	Binary CAD	Low	10	--*	--*
		High	10	0.62	< .001
	BI-RADS CAD	Low	10	0.90	.24
		High	10	0.85	.06
	aCAD	Low	10	0.51	< .001
		High	10	0.86	.07
	iCAD	Low	10	0.85	.06
		High	10	0.84	.04

Variable	Condition	Prevalence		Shapiro-Wilk	
		Block	<i>n</i>	Statistic	P-Value
False Alarm Rate	Control	Low	10	0.54	< .001
		High	10	0.71	.001
	Binary CAD	Low	10	0.86	.07
		High	10	0.78	.008
	BI-RADS CAD	Low	10	0.74	.002
		High	10	0.51	< .001
	aCAD	Low	10	0.78	.007
		High	10	0.75	.004
	iCAD	Low	10	0.69	< .001
		High	10	0.48	< .001
Sensitivity (d')	Control	Low	10	0.94	.56
		High	10	0.85	.06
	Binary CAD	Low	10	0.78	.01
		High	10	0.90	.20
	BI-RADS CAD	Low	10	0.96	.75
		High	10	0.85	.06
	aCAD	Low	10	0.94	.58
		High	10	0.92	.33
	iCAD	Low	10	0.89	.18
		High	10	0.95	.67
Target Absent Response Time	Control	Low	10	0.90	.23
		High	10	0.97	.85
	Binary CAD	Low	10	0.86	.08
		High	10	0.91	.29
	BI-RADS CAD	Low	10	0.85	.06
		High	10	0.83	.03

Variable	Condition	Prevalence		Shapiro-Wilk	
		Block	<i>n</i>	Statistic	P-Value
	aCAD	Low	10	0.85	.05
		High	10	0.87	.09
	iCAD	Low	10	0.92	.34
		High	10	0.93	.49

Note: *All means and standard deviations were the same for the binary CAD participants in the low prevalence block. Instead, a Q-Q plot was examined and indicated normality.

Table F.2

Homoscedasticity Assumption for the 5x2 Mixed Design ANOVAs

Variable	Prevalence		Levene	
	Block	<i>n</i>	Statistic*	P-Value
Hit Rate	Low	50	8.80	< .001
	High	50	3.65	.01
False Alarm Rate	Low	50	0.73	.58
	High	50	0.12	.97
Sensitivity (d')	Low	50	4.21	.006
	High	50	0.77	.55
Target Absent Response Time	Low	50	0.39	.82
	High	50	0.56	.70

Note: *Levene's test was used to check the homogeneity of variance of the between-subjects factor, automation, at each level of the within-subjects factor, prevalence block.

Appendix G

Responses to the Qualitative Usability Questionnaire

Table G.1 shows themes of positive and negative participant responses to the qualitative usability survey for the four CAD systems as well as participants' suggestions for improvements to the CAD aids. Bolded themes were mentioned by multiple participants or otherwise emphasized in the responses.

Table G.1

Themes from Responses to the Qualitative Survey by CAD Condition

Did you have a positive experience using the user interface (UI)?		
Automation	Positive	Negative
Binary CAD	Easy to use, made things go by quicker, concept was easy to understand, wouldn't rely on it	Couldn't see the full letter, system was sometimes incorrect
BI-RADS CAD	Helped participants identify the target, easy to use	--
aCAD	Easy to navigate, user-friendly	Wouldn't rely on it
iCAD	Good aid	Wouldn't rely on it
What did you like about the UI?/What did you not like about the UI?		
Automation	Positive	Negative
Binary CAD	Very thorough instructions; intuitive slider ; straightforward; easy to understand ; simple ; helped find the T faster	Would have preferred button vs slider; sometimes circle was off-center ; unreliable ; when nothing was circled, had to go through all slides

What did you like about the UI?/What did you not like about the UI?		
Automation	Positive	Negative
BI-RADS CAD	Easy to use ; straightforward; good instructions; increased confidence ; simple ; guide for when target was there ; user-friendly; easy to navigate	Hard to focus on so many shapes; unreliable; scrolling ; made participants second guess themselves
aCAD	Easy to use ; the percentages were helpful ; red color to indicate a T; easy to learn	The slider; circles would be off-center sometimes; unreliable ; having to use the mouse instead of the keyboard to go through layers ; slow and nonresponsive sometimes
iCAD	Easy to use ; user-friendly; clear; focuses your attention; helped complete the task more efficiently ; the slider	The colors made it hard to see; unreliable; made participants second-guess themselves; scrolling; finicky and responded slow; the low percentages
Did you feel like the UI aided you in your task?		
Automation	Positive	Negative
Binary CAD	Helped most of the time; helped a little ; drew your eye to the right place most of the time	Unreliable , which wasted time
BI-RADS CAD	Helped with making a final conclusion/check ; redirected focus; somewhat helpful	Occasionally wrong

aCAD	Easier to find the Ts	--
iCAD	Sometimes inaccurate, but still helpful	--
Automation	How would you improve the UI?	
Binary CAD	Change the slider to a button press; more accurate/reliable ; center the circle; use a different color instead of red (e.g., green); different letters not just a T	
BI-RADS CAD	More reliable/accurate ; more contrast in the images; add ability to rotate the images; say "unsure" instead of "need more information" (makes more sense); button to go to next slide instead of scrolling	
aCAD	Better explain what the percentages mean; more accurate/reliable; center the circles so it is easier to tell if the T is aligned	
iCAD	More engaging; more accurate/reliable ; add colors/highlight, showing which slide on the side has the T, don't ID items with lower than 70%	

Appendix H

Circle Trial Information

Table H.1 below shows further information about the trials where participants may have been impacted by off-center circles. Tables H.2 and H.3 show screenshots of the circled items in trials with off-center circles. Note that the aCAD and iCAD systems used the same programming for circle presentation, so they are grouped together.

Table H.1

Circle Trial Information

Prevalence Block	Off-Center Circle Trials
Binary CAD	
Low	1 (0.03)
High	2 (0.05)
Total	3 (0.04)
aCAD and iCAD	
Low	0 (0.00)
High	2 (0.10)
Total	2 (0.03)

Table H.2

Screengrabs of Off-Center Circled Items in the Binary CAD Condition




Trial 4, Low Block	
Trial 24, High Block	
Trial 33, High Block	

Table H.3

Screengrabs of Off-Center Circled Items in the aCAD and iCAD Conditions

Trial 24, High Block	
Trial 33, High Block	

Table H.4 below shows further information about trials with circles where participants had particularly low accuracy. Trials where participants in the selected conditions had an average accuracy of less than .70 in at least one block order condition are detailed below. Tables H.5 and H.6 show screengrabs of the circled items identified in Table H.4. Note that the aCAD and iCAD systems used the same programming for circle presentation, so they are grouped together.

Table H.4*Circle Trial Information for Low-accuracy Trials*

Trial #	Target Presence	Clear View or Partial Coverage	Prevalence Block	Prevalence Order	
				High-Low	Low-High
Binary CAD					
2	Absent	Clear	Low	.80	.60
13	Absent	Clear	High	.60	1.00
17	Present	Clear	High	.40	.60
17	Absent	Clear	Low	.60	.60
aCAD and iCAD					
24	Present	Partial	High	1.00	.42
30	Absent	Clear	Low	.50	.58
33	Present	Partial	High	.88	.67
38	Present	Clear	High	.75	.58

Note: Proportions are of participants who correctly determined target presence grouped by the order of prevalence blocks.

Table H.5*Screengrabs of Circles for Identified Trials in the Binary CAD Condition*









Trial 2, Low Block Clear	
Trial 13, High Block Clear	
Trial 17, Low Block Clear	
Trial 17, High Block Clear	

Table H.6

Screengrabs of Circles for Identified Trials in the aCAD and iCAD Conditions

Trial 24, High Block Partial	
Trial 30, Low Block Clear	
Trial 33, High Block Partial	
Trial 38, High Block Clear	

Appendix I

Descriptive Statistics for Total iCAD Activation

Table I.1 shows the means and standard deviations for total iCAD overlay activation for each participant in the iCAD condition.

Table I.1

Descriptive Statistics for Total Activation of iCAD Overlay

Participant	<i>M</i>	<i>SD</i>
6	0.000	0.000
9	0.613	0.490
15	0.000	0.000
18	0.925	0.265
20	0.013	0.112
25	1.000	0.000
35	0.988	0.112
43	0.625	0.487
51	0.938	0.244
57	0.538	0.502