COMPUTER AIDED DETECTION IN MAMMOGRAPHY

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF BIOLOGY, MEDICINE AND HEALTH

2022

Ethan Du-Crow

School of Health Sciences
Division of Informatics, Imaging and Data Sciences

Contents

List of Tables 8

List of Figures 10

Abbreviations 16

Abstract 19

Declaration 20

Copyright 21

Acknowledgements 23

List of Publications and Abstracts 24

1 Introduction 25

1.1 Clinical motivation 25

1.2 Aims 28

1.3 Thesis overview 29

2 Breast Cancer and Mammographic Screening 32

2.1 Introduction 32

2.2 Breast anatomy and types of cancer 35

2.3 Breast screening 39

2.4 Motivation for CAD 43

2.5 Summary 44

3 Computer Aided Detection **46**

3.1 Introduction 46

3.2 CAD methodology 47

3.2.1 General CAD workflow 47

3.2.2 Microcalcification cluster detection 50

3.2.3 Mass and architectural distortion detection 55

3.2.4 Recent advances 59

3.3 Analysis of performance 61

3.4 Modes of use and prompting techniques 66

3.5 Usage and effectiveness 71

3.6 Human-CAD Interaction 74

3.7 Summary 76

4 Visual Search **78**

4.1 Introduction 78

4.2 Visual search theory 81

4.3 Factors affecting visual search 92

4.4 Summary 102

5 Eye Tracking **105**

5.1 Introduction 105

5.2 Eye tracking methodology 106

5.3 Eye tracking devices 113

5.3.1 EyeLink1000 113

5.3.2 Pupil Labs 117

5.4 Visual search studies 125

5.4.1 Studies in mammography 125

5.4.2 Studies with CAD 136

5.4.3 Studies without eye tracking 141

5.5 Summary 143

6 Safety-net effect with CAD **146**

6.1 Introduction 146

6.2 Methods 148

6.2.1 Stimuli 148

6.2.2 Experimental setup 152

6.2.3 Experimental procedure 153

6.2.4 Analysis 161

6.3 Results 163

6.3.1 Eye tracking data 163

6.3.2 Trial time and coverage 164

6.3.3 Targets and prompts fixated 165

6.3.4 Observer performance 167

6.3.5 Target detectability 170

6.4 Discussion 171

6.5 Summary 175

7 Interactive CAD **176**

7.1 Introduction 176

7.2 Methods 179

7.2.1 Stimuli 180

7.2.2 Experimental setup and procedure 182

7.2.3 Analysis 191
7.3 Results 192
7.3.1 Observer performance 192
7.3.2 Effect of prompts 194
7.3.3 Effect of image score 200
7.4 Discussion 204
7.5 Summary 210

8 Dual-screen eye tracking **212**
8.1 Introduction 212
8.1.1 Eye tracking in mammography 212
8.1.2 Eye tracking with dual screens 214
8.1.3 Pupil Labs glasses 215
8.2 Methods 218
8.2.1 Materials 218
8.2.2 Calibration procedure 219
8.2.3 Calculating the useful field-of-view 221
8.2.4 Stimuli 222
8.2.5 Experimental procedure 223
8.2.6 Data analysis 225
8.2.7 Participants 228
8.3 Results 229
8.3.1 Target detectability 229
8.3.2 Calibration decay 231
8.4 Discussion 238
8.5 Summary 242

9 Interactions in CAD 245

9.1 Introduction 245

9.2 Experiment design 246

9.2.1 Stimuli 246

9.2.2 Training 248

9.2.3 Experimental procedure 252

9.2.4 CAD Prompts 254

9.2.5 Analysis 255

9.3 Results 255

9.3.1 Trial times 256

9.3.2 Observer performance 257

9.3.3 Effect of prompting 259

9.3.4 Lab versus online 263

9.4 Discussion 265

9.5 Summary 268

10 From lab to clinic 269

10.1 Introduction 269

10.2 Methods 271

10.2.1 Stimuli 271

10.2.2 CAD prompts 273

10.2.3 Experimental setup 277

10.2.4 Training 278

10.2.5 Experimental procedure 280

10.2.6 Analysis 282

10.3 Discussion 283

10.4 Summary 286

11 Conclusions and Future Work 288

11.1 Summary 288

11.2 Contributions to Knowledge 294

11.3 Future Work 295

11.3.1 Unprompted targets 295

11.3.2 Multiple targets 297

11.3.3 Reducing the impact of false prompts 297

11.3.4 Reader experience 299

11.3.5 Alternative modalities 299

Word Count: 65,092

List of Tables

2.1 Overview of breast screening programmes worldwide 34

3.1 CAD sensitivity and false prompt rate measured across multiple studies 54
3.2 Definitions of true and false positives and negatives in mammography 62

5.1 Specifications for the EyeLink1000 tracker 116
5.2 Specifications for the Pupil Core headmounted tracker 125

6.1 Comparison of single reading with CAD to single reading alone and double reading, in terms of sensitivity and recall rate 147
6.2 Results from Kolmogorov–Smirnov test for splitting calcification clusters into image sets 150

7.1 Summary of results from Experiment 1 showing trial time, image coverage, sensitivity, and false positives per image 193
7.2 Summary of results from Experiment 2 showing trial time, image coverage, sensitivity, and false positives per image 193
7.3 Summary of observer behaviour on queries for each experiment categorised by type of prompt (true prompt, false prompt, and no prompt available) 195
7.4 Proportion of fixated targets that were not marked but were queried . . 200

8.1 Specifications of commercially available eye trackers. 216

8.2 Results from the head pose tracking technique to measure distance to the screen 222
8.3 Size and brightness factors, and number of each type for the eight target categories 223
8.4 Comparisons of accuracy distributions to reference accuracies 234

9.1 Distribution of prompts in online experiment 254
9.2 Image categories in the online experiment 255
9.3 Time spent on each image in total, pre CAD and post CAD 256

10.1 Characteristics of OPTIMAM images 272
10.2 False prompt distribution across images used in study 274

List of Figures

2.1 Breast anatomy 35

2.2 CC view mammogram 37

2.3 Common abnormalities on mammograms 38

3.1 Flowchart of general stages of a CAD algorithm 48

3.2 Mammogram marked by CAD system 49

3.3 Microcalcification detection process described by McLoughlin et al. (2004) 52

3.4 Visualisation of spicule detection algorithm from Muralidhar et al. (2010) 58

3.5 Comparison between the structure of traditional machine learning and deep learning methods, adapted from Jiménez-Gaona et al. (2020) . . 60

3.6 Example ROC curve 64

3.7 Longitudinal study cancer detection rates 66

3.8 CAD operating points 67

3.9 Image from Cunningham et al. (2016) showing analogue CAD prompt approach 69

3.10 Bright highlighting of abnormal regions 70

4.1 Diagram of the human eye 79

4.2 Search slopes of various efficiencies 82

4.3 Schematic representation of Guided Search 6.0, from Wolfe (2021a) . 85

4.4 Estimated size of the functional visual field in tasks of different difficulty 88
4.5 An example of an image patch in a hybrid foraging search task 91
4.6 Examples of perceptual errors in radiology 93
4.7 Results from Evans et al. (2013) demonstrating the low prevalence effect 94
4.8 Examples of inattentional blindness 99

5.1 Diagram of corneal reflections for various translations and rotations of the eye 107
5.2 Schematic of eye tracking setup 108
5.3 Diagram for calculation of eye gaze direction 109
5.4 Diagram of EyeLink1000 114
5.5 Screenshot of the EyeLink setup screen 115
5.6 Screenshot of the EyeLink record screen 116
5.7 Photograph of Pupil Labs - Pupil Core Headset 117
5.8 Screenshot of Pupil Capture software used for data collection with Pupil Core headset 118
5.9 Screenshot of eye camera image showing algorithm view in Pupil Capture 120
5.10 Schematic of Pupil Labs model for eyeball position estimation 121
5.11 Screenshot of Pupil Player software used for data analysis with Pupil Core headset 123
5.12 Screenshot of head pose tracking in Pupil Player software 124
5.13 Search path of mammographer interpreting abnormal mammogram . . 126
5.14 Cross-recurrence plot for comparing the visual search of two radiologists from Gandomkar et al. (2018) 128
5.15 Porportion of saccades hitting target versus distance from target, form Wolfe et al. (2021b) 131

5.16 Images from Mall et al. (2019*a*) showing how attentional levels were defined for a deep learning prediction model 136
5.17 Visual search patterns with and without CAD prompts from Drew et al. (2012) 137
5.18 Scanpath of a non-expert participant reading a mammogram with subtle prompts versus no prompts, from Hatton et al. (2004) 138
5.19 Heatmaps of visual search of participants in no-CAD, binary CAD, and iCAD conditions, from Drew et al. (2020) 139

6.1 Example of a microcalcification cluster 149
6.2 Example of experiment image with microcalcification cluster inserted. 152
6.3 Photo of experimental setup 153
6.4 Training screen to introduce participants to microcalcification clusters 154
6.5 Example of interactive training image with CAD 156
6.6 Example of experiment image in no-CAD condition 158
6.7 Example of experiment image in CAD condition 160
6.8 Example of typical eye tracking data for single participant 164
6.9 Mean trial time and image coverage 165
6.10 Mean target fixated percentage and dwell time for each condition . . . 166
6.11 Percentage of prompts acted on in the CAD condition 167
6.12 Observer sensitivity for all images, and prompted and unprompted targets 168
6.13 Mean false positive per image rate of observers for all trials, and for prompt absent and false prompt present trials 169
6.14 Percentage of participants that detected each target in the no-CAD and CAD conditions 171

7.1 Clinical example of an interactive CAD system 177
7.2 Distribution of ScreenPoint image scores 178

7.3 Target detection percentages for image sets A′ and B′ 182

7.4 Example of experiment training introducing interactive CAD 184

7.5 Example of experiment training with interactive CAD example showing false prompts 185

7.6 Example of experiment training introducing overall image score . . . 186

7.7 Example of experiment image in the CAD+Score condition 188

7.8 Distribution used for assigning prompt confidence values 190

7.9 Number of images with each image score 191

7.10 Distribution of the mean number of regions queried per image in each experiment 194

7.11 Dwelltime on regions before and after participants queried them . . . 196

7.12 Percentage of targets queried versus detection rate 197

7.13 True and false prompts acted on versus prompt confidence values for Experiment 1 198

7.14 True and false prompts acted on versus prompt confidence values for CAD condition in Experiment 2 198

7.15 True and false prompts acted on versus prompt confidence values for CAD+Score condition in Experiment 2 199

7.16 Trial time and false positive percentage change between the CAD and CAD+Score conditions for the range of image scores 201

7.17 Proportion of prompts marked by participants in Experiment 2 for a given combination of prompt confidence and image score 202

7.18 Percentage of prompts acted on versus prompt confidence for individual prompts 203

7.19 Percentage of prompts acted on versus image score 203

8.1 Photograph of experimental setup 219

8.2 Calibration target and mean validation validation accuracy versus calibration target presentation duration 220
8.3 Example of experiment image 223
8.4 Schematic of the positions of targets in the calibration procedure . . . 225
8.5 Flowchart of the data analysis in Pupil Player software 227
8.6 Flowchart of analysis of data exported from Pupil Player 228
8.7 Target detection rates for different mass categories 230
8.8 Target Detection rate for individual masses versus their size in pixels . 231
8.9 Validation accuracy across participants as a function of the validation number 232
8.10 Validation precision across participants as a function of the validation number 233
8.11 Calibration decay for sequential validations and relative to validation 1 236
8.12 Mean validation accuracy across participants as a function of validation number for all validation targets, corner targets only, and all except the corner targets 237
8.13 Mean validation precision across participants as a function of validation number for all validation targets, corner targets only, and all except the corner targets 238

9.1 Target sensitivities from lab experiments 247
9.2 Examples of mass targets for training in online experiment 248
9.3 Examples of calc targets for training in online experiment 249
9.4 Online experiment CAD training screen 250
9.5 Online experiment training trial with feedback 251
9.6 Online experiment trial example 253
9.7 Trend of trial time versus image number 257

9.8 Boxplot of observer sensitivities 258
9.9 Boxplot of false positives per image for participant clicks 259
9.10 Median number of markers and removed by participants post-CAD . . 260
9.11 Observer sensitivity for marked targets for each image category 261
9.12 False positive response rate for each image category 263
9.13 Lab versus online target sensitivity for participants 264

10.1 Visualisation of the top-hat filters that were applied to images for highlighting calc regions 276
10.2 Mammogram with filters applied to highlight calc regions 276
10.3 Photography of experiment screen 277
10.4 Experiment screenshots demonstrating zooming 278
10.5 Screenshot of experiment training 279

Abbreviations

ABUS	Automated breast ultrasound
AET	Attentional Engagement Theory
AFROC	Alternative free-response receiver operating characteristic
ANN	Artificial Neural Network
AUC	Area under the receiver operating characteristics curve
CAD	Computer aided detection
CADe	Computer aided detection
CADx	Computer aided diagnosis
CAP	Computer assisted perception
CAVS	Computer-assisted visual search
CC	Cranial-Caudal
CNN	Convolutional Neural Network
CR	Corneal reflection
CT	Computerised Tomography
DBT	Digital breast tomosynthesis
DCIS	Ductal carcinoma in situ
EOG	Electro-oculography
FDA	US Food and Drug Administration

FFDM	Full-field digital mammography
FIT	Feature Integration Theory
FN	False negative
FOV	Field of view
FP	False positive
FPF	False positive fraction
FrACT	Freiburg Vision Test
FROC	Free-response receiver operating characteristic
FVF	Functional visual field
GS	Guided Search
ICSN	International Cancer Screening Network
IR	Infrared
JAFROC	Jackknife free-response receiver operating characteristic
LCIS	Lobular carcinoma in situ
MLO	Mediolateral-oblique
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
NHSBSP	National Health Service Breast Screening Programme
NPV	Negative predictive value

OMI-DB	OPTIMAM Mammograohy Image Database
pAUC	Partial area under the receiver operating characteristics curve
PERFORMS	PERsonal perFORmance in Mammographic Screening
PPV	Positive predictive value
ROC	Receiver operating characteristic
ROI	Region of interest
RT	Reaction time
SVM	Support Vector Machine
TN	True negative
TP	True positive
UFOV	Useful field of view
VAS	Visual analogue scale
VWM	Visual working memory

Abstract

Breast screening programmes, in which mammograms are examined for signs of cancer, have been implemented in many countries. In the UK, all mammograms are reviewed by two expert medical readers. Because abnormalities are variable, subtle, and infrequent, this task is difficult and prone to human error. Computer aided detection (CAD) systems aim to improve the performance of expert readers by indicating potentially abnormal regions that may otherwise have been missed. CAD can improve performance of some readers, but often at the cost of an increase in the false positive rate due to the high number of prompts on normal regions.

This thesis explores the role of CAD in mammography through a series of visual search experiments, using simulated images and targets analogous to mammography screening. First, CAD was evaluated as a second reader, where the image is first viewed unaided and then once again with CAD. This initial unaided search was found to be truncated in terms of review time and the amount of the image viewed ($p<0.001$).

Subsequently, an interactive CAD system was investigated where prompts are only displayed when readers actively query regions. Scores accompanied the prompts to denote the likelihood that they marked a target, and they had a much greater impact on whether the prompt would be marked by the observer ($p<0.001$) than an image score indicating the likelihood that a target was present somewhere in the image ($p=0.72$).

CAD systems often use different prompts to indicate the type of abnormality. A further study was conducted with two target types and multiple different prompts on single images. Readers' ability to detect targets was unaffected by false prompts, whether or not they indicated the same target as the true prompt ($p\geq0.30$).

A methodology is outlined for tracking eye movements across a clinical radiology workstation using eye tracking glasses. An observer experiment showed that recalibrating the glasses every 5 minutes would maintain a reasonable level of accuracy and precision in future studies.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property

and/or Reproductions described in it may take place is available in the University IP Policy (see `http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487`), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see `http://www.manchester.ac.uk/library/aboutus/regulations`) and in The University's policy on presentation of Theses

Acknowledgements

I would like to express my gratitude to the following people who have supported me throughout my doctoral programme:

My supervisors Prof. Sue Astley and Dr. Johan Hulleman for their endless support, advice, and encouragement. Without your guidance, this thesis would not have been possible.

Dr. Lucy Warren at the National Co-ordinating Centre for the Physics of Mammography for providing me with the microcalcification clusters that featured so heavily in this thesis.

Joleen Kirsty Eden and Rita Borgen, consultant radiographers at Burnley General Hospital, for their time, expertise, and effort in annotating mammograms. Also, for their helpful suggestions on the methodology of our study.

Dr. Elaine Harkness for her valuable advice on the statistical approaches used in this thesis.

Dr. Steven Squires for running hundreds of mammograms through a density algorithm.

My colleagues Georgia, Anna Maria, and Luke for their friendship. You made a dim Stopford Building a great place to work.

My family and friends for their love and support throughout my PhD. Thank you to my parents for always looking out for me.

Last but certainly not least, my partner Bridie. Thank you for your love and your calming presence. Your support and positive encouragement kept me going until the end.

List of Publications and Abstracts

- Du-Crow, E., Warren, L., Astley, S. M. and Hulleman, J. (2019), Is there a safety-net effect with computer-aided detection (CAD)?, in 'Proc. SPIE 10952, Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment', 109520J.

- Du-Crow, E., Astley, S. M. and Hulleman, J. (2019), 'Is there a safety-net effect with computer-aided detection?', *Journal of Medical Imaging* 7(02), 1.

- Du-Crow, E., Astley, S. M. and Hulleman, J. (2020), Suspicious minds: effect of using a lesion likelihood score on reader behaviour with interactive mammographic CAD, in C. V. Ongeval, N. Marshall and H. Bosmans, eds, '15th International Workshop on Breast Imaging (IWBI2020)', SPIE.

- Du-Crow, E., Warren, L. M., Astley, S. M. and Hulleman, J. 'Is there a safety-net effect with computer-aided detection (CAD)?', accepted for oral presentation at SPIE Medical Imaging 2019, San Diego.

- Du-Crow, E., Astley, S. M. and Hulleman, J., 'Suspicious minds: effect of using a lesion likelihood score on reader behaviour with interactive mammographic CAD', accepted for oral presentation at International Workshop on Breast Imaging 2020, Leuven.

Chapter 1

Introduction

1.1 Clinical motivation

Female breast cancer is the most common cancer worldwide (Sung et al., 2021). Large-scale screening programmes, using x-ray mammography to detect signs of cancer, have been implemented in many countries around the world in order to reduce mortality rates. In the UK, as part of the NHS Breast Screening Programme (NHSBSP), mammograms are read by two medical experts who decide whether further investigations are required. Despite this, some cancers are missed at screening. Furthermore, there is a critical shortage of radiologists in the UK (The Royal College of Radiologists, 2020). This results in high workloads and increased wait times for patients. A potential tool for reducing workload and improving performance is computer aided detection (CAD).

CAD is a technology designed to aid expert medical readers in the detection of cancer and disease in medical images by prompting readers to evaluate potentially abnormal regions. While CAD has not yet been formally adopted in the UK, it is widely used in the US. CAD has been shown to improve breast screening performance of a single radiologist in terms of sensitivity but at the cost of a significant increase in recalls (Nishikawa et al., 2012), and has not yet been shown to be a better option than, or even equivalent to double reading (Gilbert et al., 2008b; Azavedo et al., 2012). The increase in recalls is a consequence of the high number of false positive prompts (around 0.5 per image (The et al., 2009; Cole et al., 2012)) from CAD systems, one of the biggest issues with CAD. The interaction between medical experts and CAD systems is also not well-understood. Many studies have implemented fixed operating points (i.e. the balance between sensitivity and specificity of the prompting algorithm) but the interaction between the prompts of multiple algorithms has not been investigated.

To gain a better understanding of human-CAD interactions, experiments were carried out using non-expert observers and synthetic mammogram-like images with small and subtle targets. The effects of the type of prompt used, the interaction between multiple algorithms each with a different operating point, distraction caused by false positive prompts marking non-target areas, and how prompts affect overall search performance were investigated. In addition to detection and error rates, eye tracking was used to further analyse visual search behaviour, particularly the effects of false prompts on how the image is evaluated by the observers.

The typical workflow for reading with CAD is for the medical expert to first interpret the mammogram unaided and then to interpret the image again with the aid of CAD, referred to as second reader CAD. It was explored whether the anticipation of a second search aided by CAD leads to a less thorough initial search compared to observers not using CAD (described as the ‘safety-net’ effect (Astley and Gilbert, 2004)), and the implications of this in relation to previous CAD studies.

An alternative approach to second reader CAD is interactive CAD, where prompts are only displayed when the reader requests them at a particular region in the image, provided there is a prompt available at that region. Interactive CAD software often provides a score with the prompt to denote the likelihood of the region being malignant and a global image score to denote the likelihood of a lesion being present somewhere in the image. It was investigated how observers' search and behaviour are affected by interactive CAD prompts and how the prompt-level and image-level scores influence decision making.

Since mammograms contain multiple types of abnormalities, separate CAD algorithms must be deployed to prompt each type. Typically, the shape or symbol used for prompts will be unique to each abnormality. On many images therefore, there will be multiple prompts denoting different abnormalities. It was measured how the presence of a prompt of one target type impacts how observers behave on the prompt of another, and how they react to multiple prompts denoting the same target type.

Experiments with medical experts in mammography require the eye tracking system to be able to track across a wide field-of-view with two monitors to replicate the clinical setting of standard breast screening. A methodology was developed to track eye movements across a dual-screen radiology setup using wearable eye tracking glasses. Using this methodology, a study was planned to investigate the impact of CAD prompts on the search behaviour of medical experts.

1.2 Aims

The principal aim of the research described in this thesis is to increase understanding of human-CAD interaction through visual search experiments, identifying the strengths of CAD and where improvements need to be made.

Specific aims are:

1. Analyse the impact of a second viewing with CAD on an initial unaided viewing of an image, in terms of search thoroughness and observer sensitivity.
2. Investigate the impact of interactive CAD prompts on observer behaviour and visual search, and how local and global confidence scores affect how observers react to prompts.
3. Investigate the interaction between multiple CAD prompts of different target types, and how the presence of one prompt type affects how observers react to another.
4. Develop a methodology for dual-screen eye-tracking for a clinical radiology workstation.
5. Examine how CAD prompts affect the visual search behaviour of expert readers and the distraction caused by false positive prompts on a full clinical radiology setup with CAD as a second reader.

1.3 Thesis overview

This section provides a brief overview of the remaining chapters in this thesis. Chapters 2 to 5 review the relevant literature and Chapters 6 to 10 detail observer experiments.

Chapter 2 provides an overview of breast cancer and breast screening in the UK, including the arguments for and against screening. This chapter outlines the motivation for CAD, highlighting the need for improved cancer detection rates and a reduced workload for radiologists.

Chapter 3 introduces CAD, describing the history of algorithm development for breast screening, the general methodology of CAD algorithms for mammography, and its usage around the world. The efficacy of CAD is presented, along with the different modes of operation. Literature on human-CAD interaction is reviewed and the motivation for gaining a better understanding of how observers are influenced by CAD is presented.

Chapter 4 describes the key features of eye movements and the various fundamental models of visual search. This chapter concludes with a literature review of the factors that affect visual search in mammography and other imaging modalities.

Chapter 5 presents the various methods for tracking eye movements and a description of the eye tracking devices used in this thesis. This is followed by a literature review of visual search studies in mammography and with CAD, and alternative methods to improve observer performance without eye tracking.

Chapter 6 describes an observer study investigating the safety-net effect with CAD. Eye tracking was used to compare search behaviour between conditions with and without CAD.

Chapter 7 focuses on interactive CAD and presents two observer studies that investigated how lesion likelihood scores and overall image scores influenced the way observers reacted to prompts. Eye tracking was used once again to quantify how long participants attended to prompts and target regions, in addition to the amount of the image that was viewed in the CAD and no-CAD conditions.

Chapter 8 describes the methodology used for tracking eye movements across a dual-screen clinical radiology setup. Furthermore, this chapter presents the results of an observer study used to measure the accuracy of the system and provides recommendations for running experiments using the setup.

Chapter 9 describes an online observer study that evaluated the impact of multiple prompts of different target types on observer performance and the potential for distraction due to the presence of false prompts. This chapter also presents comparisons between the performance of participants completing studies online versus in a laboratory setting.

Chapter 10 presents the methodology for a planned observer study with expert readers using a dual-screen eye tracking setup. Due to the COVID-19 pandemic, no data was collected for this study. The possible outcomes of the study in relation to the literature are also discussed.

Chapter 11 concludes the thesis, and discusses the main contributions and the possible direction of future work.

Chapter 2

Breast Cancer and Mammographic Screening

2.1 Introduction

Breast cancer is the most common cancer in the UK. Between 2015 and 2017, an average of 55,200 new cases were registered each year, accounting for 15% of all cancer cases (Cancer Research UK, 2021*a*). In England, breast cancer incidence increased from 165.5 to 176.4 cases per 100,000 females between 2008 and 2015, and down to 166.7 in 2017 (Office for National Statistics, 2019). The WHO estimated that 2.26 million new breast cancer cases were diagnosed worldwide in 2020; in less developed regions it is the most frequent cause of cancer death for women, and in more developed regions it is the second most frequent (WHO, 2020). Early detection of breast cancer is important for a good prognosis; for example, in the UK, 97.9% of women diagnosed at the early stage survive for five years or more, compared with 26.2% at the late stage (Cancer Research UK, 2021*a*). Breast cancer screening programmes have been implemented in a number of countries with the aim of reducing mortality rates and enabling less intensive treatments.

An overview of breast screening programmes currently implemented in 26 countries is given in Table 2.1. The majority of the screening programmes were started in the 1980s and 90s, and cover similar age groups between 40 and 75 years old. The participation rate varies between countries, which is likely to be most affected by access to affordable screening. The effectiveness of these programmes is the subject of much debate (see Section 2.3). These programmes typically use mammography, which are low-dose x-rays of the compressed breast. During acquisition of a mammogram, the breast is compressed between two plates and a single projection will be taken of the flattened breast. Two views are typically acquired, the craniocaudal (from the top down) and the mediolateral oblique (from the side at an angle).

It is important to note the different categories of expert readers who interpret mammograms and other breast images in the UK. These are: radiologists (doctors with one of their specialisms in diagnosis and treatments of breast diseases), advanced practitioner radiographers, and breast clinicians. In the NHSBSP, all mammograms are read by at least one radiologist (Jenkins et al., 2013). From here onwards, all trained clinical image interpreters will be referred to as ‘readers’ unless otherwise stated. Quality assurance measures are used to maintain the professional standards of all readers in screening programmes. Readers in the NHSBSP are required to review a minimum of 5,000 cases per year and undergo performance audits (The Royal College of Radiologists, 2013). Also, readers in the UK partake in the PERFORMS (PERsonal perFORmance in Mammographic Screening) scheme (Gale, 2010), which is a self-assessment and training programme.

Table 2.1: Overview of breast cancer screening programmes in 26 countries in the International Cancer Screening Network (ICSN) (National Cancer Institute, 2012*a*).

Region/Country	Programme Type[1]	Year Programme Began	Detection Methods in Routine Use[2]	Age Groups Covered	Number of Women Screened (2010)	Participation Rate (2010)
Australia	NS	1991	MM, DM	40-75+	data not available	data not available
Canada	NS	1988	MM, DM, CBE	50-69	196,187	47.3%
China	NS	2009	MM, CBE, U	40-59	1,200,000	data not available
Denmark	S	1991	DM	50-69	275,000	73.0%
Finland	N	1987	DM	50-64	data not available	85.0%
France	N	1989	MM, DM, CBE	50-74	2,343,980	52.3%
Iceland	N	1987	DM, CBE	40-69	20,517	60.0%
Israel	N	1997	MM, DM	50-74	220,000	72.0%
Italy	NS	2002	MM, DM	50-69	1,340,311	60.5%
Japan	NS	1977	MM, DM, CBE	40-75+	2,492,868	19.0%
Korea	N	1999	MM, DM	40-75+	2,602,928	39.3%
Luxembourg	N	1992	DM	50-69	14,586	64.0%
Netherlands	N	1989	MM, DM	50-74	961,766	80.7%
New Zealand	N	1998	MM, DM	45-69	211,922	67.5%
Norway	N	1996	DM	50-69	199,818	76.0%
Poland	N	2006	MM, DM	50-69	985,364	39.0%
Portugal (Central Region)	S	1990	DM	45-69	100,348	63.0%
Portugal (Alentejo Region)	S	1997	DM	45-69	7,298	58.4%
Saudi Arabia	S	2007	DM	40-64	6,200	19.0%
Spain (Catalonia)	NS	1995	MM, DM	50-69	527,000	65.0%
Spain (Navarra)	NS	1990	DM	45-69	40,016	87.3%
Sweden	S	1986	MM, DM	40-74	1,414,000	70.0%
Switzerland	NS	1999	MM, DM	50-69	60,700	48.2%
United Kingdom	N	1988	MM, DM	50-69	1,957,124	73.3%
United States	O	1995	MM, DM, CBE	40-75+	416,000	66.5%
Uruguay	O	1990	MM, CBE, U, BSE	40-69	352,000	data not available

[1] Programme types: N (National screening policy with national programme implementation), NS (National screening policy with state/provincial/regional screening programme implementation), S (State/Provincial/Regional screening and programme implementation), O (Other).

[2] Detection methods: MM (screen-film mammography), DM (digital mammography), T (Tomosynthesis/3-D mammography), CBE (clinical breast examination), BSE (breast self-examination), MRI (Magnetic Resonance Imaging), U (ultrasound), CT (Computerised Tomographic Imaging).

2.2 Breast anatomy and types of cancer

The anatomy of the female breast is shown in Figure 2.1. The main components are fibrous tissue, glandular tissue, fat, and neurovascular structures. The fibrous tissue is the connective or supportive tissue, and the glandular tissue is the functional tissue (parenchyma), which includes the ducts and lobules (Jesinger, 2014). Milk is produced by the lobules and is supplied to the nipple through the ductal network. The fibrous and glandular tissue is often referred to collectively as fibroglandular tissue, and the relative amounts of fibroglandular tissue and fat determine the breast density. Figure 2.2 shows a labelled mammogram. The radiodense fibroglandular tissue and the vascular structures appear brighter than the fat.

1	Chest wall
2	Pectoralis muscles
3	Lobules
4	Nipple surface
5	Areola
6	Lactiferous duct
7	Fatty tissue
8	Skin

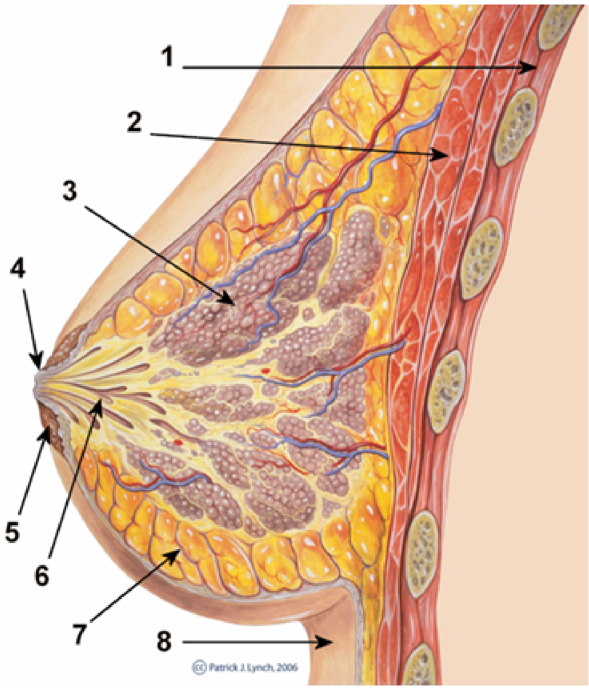


Figure 2.1: Illustration of breast anatomy, adapted from Lynch (2006).

Breast cancer is generally classified as invasive or non-invasive, and usually forms within the lining of the ducts (in 85–90% of cases), but can also form in the lobules in 10–15% of cases (Du et al., 2018). Invasive cancers are those that have spread out of the ducts or lobules and into the surrounding tissue. Non-invasive cancers have not spread from where they originated and are called ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS) when the cancer is in the ducts and lobules, respectively. Many non-invasive cancers still require treatment, since up to 40% of DCIS progress to invasive cancers if left untreated (Cowell et al., 2013). Similarly, women with LCIS have a 30%–40% lifetime risk of developing invasive breast cancer, compared to the average lifetime risk of 12.5% (Breastcancer.org, 2016). Invasive breast cancers may enter the lymph nodes under the arm and travel through the lymphatic system, spreading to other parts of the body. This is metastatic breast cancer and accounts for around 90% of deaths for women with breast cancer (Fouad et al., 2015).

Triple negative breast cancer accounts for 15% of cases and is called triple negative because the cancers do not have receptors for oestrogen, progesterone and HER2 protein, and therefore treatment targeting these is ineffective (Cancer Research UK, 2021*b*). Triple negative breast cancer is the most aggressive subtype of breast cancer and has the poorest prognosis (Bao and Prasad, 2019). There are also rare cancers such as inflammatory breast cancer and Paget's disease (Cancer Research UK, 2021*b*).

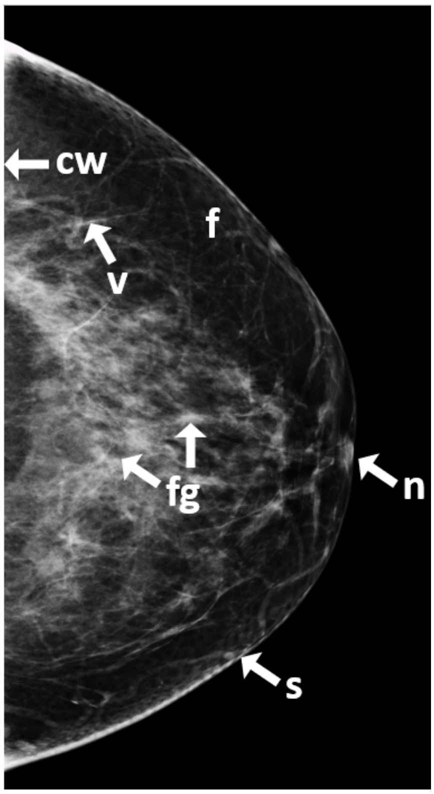


Figure 2.2: Craniocaudal (CC) view mammogram. The main features are labeled as skin (s), chest wall (cw), fat (f), nipple (n), fibroglandular tissue (fg) and vascular structures (v).

There are three main appearances of abnormalities visible on mammograms. These are masses, microcalcifications, and architectural distortions. The mass shown in Figure 2.3(a) is known as a ‘spiculated mass’, it has an irregular outline with lines (spicules) radiating from the central density, and is the strongest sign of malignancy for a mass; Liberman et al. (2002) reported an 80% positive predictive value (PPV) for spiculated margins detected with with MR imaging and had the highest PPV among mammographic features (Mahoney et al., 2012). Microcalcifications, seen in Figure 2.3(b), are small ($<$1 mm in diameter) calcium deposits and can be an early sign of breast cancer. The probability that a microcalcification is a sign of malignancy is determined by their appearance, distribution, and density. Microcalcification clusters that are more tightly packed are more likely to be due to a malignant cancer. Usually, tightly packed is defined as >5 calcifications in 1 cm^3 (Henrot et al., 2014). Architectural distortions, such as the one shown in Figure 2.3(c), are characterised by radiating lines of tissue or distorted tissue, but with no central mass. They are indicative of breast cancer – an estimated 67% of screening mammograms showing architectural distortions represented malignancy (Bahl et al., 2015). In a study investigating areas in need of improvement for readers in the NHSBSP, architectural distortions were rated as the most difficult abnormal feature for readers to detect amongst malignant cases ($p<0.05$) (Scott and Gale, 2006).

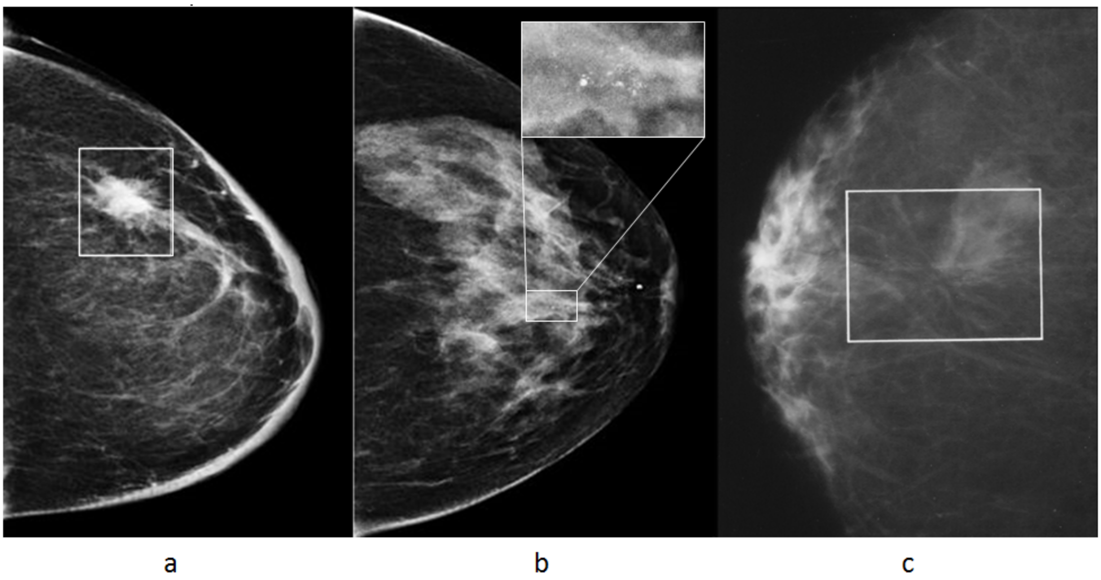


Figure 2.3: Common abnormalities in digital mammograms: mass (a), microcalcification cluster (b), and architectural distortion (c). Images a and b adapted from Al-Ghaib (2015), and c from Baker et al. (2003).

2.3 Breast screening

The NHSBSP offers breast screening to women in the UK every three years between the ages of 50 and 70, with a current trial extending the age range in some regions of the country by three years at both ends. Two-view mammograms are taken for each breast, in the mediolateral-oblique (MLO) and craniocaudal (CC) views. Mammograms are read independently by two expert readers with review in case of disagreement. Around 1 in 25 women will be recalled, and approximately one quarter of those will be diagnosed with breast cancer (Cancer Research UK, 2020). The NHSBSP is also responsible for monitoring women at high risk of developing breast cancer due to factors such as high breast density, significant family history or previous radiotherapy (Jenkins et al., 2013). According to guidelines set by the National Institute for Health and Care Excellence (NICE), women at high risk receive formal assessment and may be eligible for yearly MRI scans (NICE, 2017).

Breast cancer mortality rates decreased in the UK between 1988 and 2018 from 59.1 deaths to 33.3 deaths per 100,000 females (Cancer Research UK, 2021*a*). This decrease in mortality corresponds to the introduction of the NHSBSP (Table 2.1). However, the extent to which this decrease is attributable to mammographic screening, and the benefits of screening in general, is an area of extensive debate.

A case-control study by Massat et al. (2015), nested within the NHSBSP, reported that breast cancer mortality was 39% lower for women who attended screening than those who did not, after correcting for self-selection bias. The Independent UK Panel on Breast Cancer Screening reviewed the benefits and harms of breast screening, and estimated there was a 20% reduction in mortality for women invited to screening associated with the NHSBSP, preventing 1,300 deaths a year (Marmot et al., 2013). They also acknowledged overdiagnosis as a cost of screening – where cancers are found that would not have otherwise been detected or have been life-threatening and subsequently undergo unnecessary treatment. It is not currently possible to distinguish between the detected cancers that should be treated and those that should not. The panel estimated that for each cancer death prevented, around three cases will be overdiagnosed and unnecessarily treated (Marmot et al., 2013).

More recently, a large-scale study was conducted of 549,091 women eligible to attend breast screening in Sweden (Duffy et al., 2020). They demonstrated that there was a 41% reduction in breast cancer mortality risk within 10 years ($p<0.001$) for women attending screening versus those not attending. Furthermore, this reduction in mortality risk was accompanied by a 25% reduction in the rate of advanced breast cancers.

The Cochrane Breast Cancer Group's review of breast screening (Gøtzsche and Jørgensen, 2013) was more critical of screening than the UK panel. The review looked at seven randomised trials with over 600,000 women in total, and concluded that for every 2,000 women invited to screening for 10 years, one life will be saved, 10 healthy women will be overdiagnosed, and 200 women will experience psychological distress from a false alarm. While some studies argue that screening has a significant impact on reductions in breast cancer mortality rates (Heywang-Köbrunner et al., 2011; Marmot et al., 2013; Massat et al., 2015; Duffy et al., 2020), others argue that other factors such as improvements in cancer treatment and 'breast awareness' (awareness of importance of self-examinations) have been more effective (Gøtzsche and Jørgensen, 2013; Jørgensen and Gøtzsche, 2010; Narod et al., 2015). It should be noted that improvements in cancer treatment could not explain the reduction in mortality reported by Duffy et al. (2020), since the analysis period of 10 years would have seen the same treatments applied for both screened and non-screened cohorts. The benefits of screening are clear and mostly agreed upon, but the extent of overdiagnosis associated with screening is likely to continue to be debated.

Mammography, both screen-film and digital, is the most widely used imaging modality for breast cancer screening (see Table 2.1). Whilst it is proven to be effective for cancer detection and is also the most cost-effective screening technique, it has its limitations. A disadvantage of mammography is the use of harmful ionising radiation, but the risk of developing fatal breast cancer from a mammogram is small – between 1.3 and 1.7 cases per 100,000 for women aged 40 at exposure and 1 in 1 million for women aged 80 at exposure (Hendrick, 2010). Also, for women with dense breast tissue, mammography has a reduced sensitivity due to the masking of cancers. Sensitivity of mammography was found to decrease from between 85.7%–88.8% for women with fatty breast tissue to between 62.2%–68.1% for women with extremely dense breast tissue (Freer, 2015). Denser breasts are more prevalent for pre-menopausal women, and since the study by Hendrick (2010) focused on 40 and 80 year olds, the results are not directly applicable to the NHSBSP since both ages fall outside of the screening age group. MRI has been investigated as an alternative to mammography, and may offer improved sensitivity (DeMartini and Lehman, 2008). However, MRI would not be suitable for primary modality of a screening programme due to the high cost, long scan times, unsuitability for larger women, and use of a contrast agent. Research is ongoing into the use of abbreviated MRI for particular groups of women within stratified screening programmes (Comstock et al., 2020). Abbreviated MRI uses a single phase to reduce the cost, scan time, and interpretation time of MRI imaging (Comstock et al., 2020).

Digital breast tomosynthesis (DBT) is an x-ray imaging technique that has also been investigated as an adjunct or alternative modality to mammography. During acquisition of a DBT image, the x-ray tube moves in an arc around the breast, taking multiple 2D projections to create a volumetric image of the breast. Each projection is a fraction of the dose of FFDM, with the total dose from DBT similar to FFDM. A study investigating the effectiveness of DBT in the NHSBSP found that DBT significantly improved specificity compared to mammography and achieved a similar sensitivity (Gilbert et al., 2015). Since DBT images use multiple projections, many slices are generated, and the viewing time for readers is longer than for conventional mammography, which puts further strain on readers' workload. DBT is often used alongside FFDM, resulting in an overall higher dose. Furthermore, microcalcifications are often easier to detect in 2D, whereas masses are easier to detect in 3D because they are disambiguated from overlying tissue, and therefore it is preferable to have both views. For women with dense breasts attending screening in the US and Germany, abbreviated MRI was found to have a significantly higher detection rate for invasive cancers compared to DBT (Comstock et al., 2020).

Another imaging modality that is currently being investigated as an adjunct to mammography is automated breast ultrasound (ABUS), targeted specifically at women with dense breasts. During image acquisition, the scanner is pressed onto the breast and the transducer moves automatically to take multiple image slices, which form a 3D image of the breast. Since the scanning is automated, it eliminates the issue with hand-held ultrasound of operator variability. Screening studies have reported increases in sensitivity between 26.7% and 50.0% using mammography and ABUS combined compared to mammography alone (Kelly et al., 2010; Brem et al., 2015; Wilczek et al., 2016). ABUS also increased specificity by 3.5 percentage points in one study (Kelly et al., 2010), but decreased specificity by 0.7 percentage points (Wilczek et al., 2016) and 13.4 percentage points (Brem et al., 2015) in the other studies. Like DBT, ABUS has the advantage of being a 3D technique and therefore reduces the effect of overlying tissues for mass detection. However, unlike both DBT and mammography, ABUS does not use ionising radiation.

2.4 Motivation for CAD

There is currently a critical shortage of radiologists in the UK. With 8.1 radiologists (including trainees) per 100,000 people, the UK ranks well below the European median of 12.7 per 100,000 (The Royal College of Radiologists, 2020). Between 2014 and 2019 there was a mean increase of 4% per year in the consultant radiologist workforce. However, over the same period, there was an estimated 7% per annum increase in the number of x-rays, CTs and MRIs reported. This disparity between increase in workload and increase in workforce has led to 99% of radiology departments being unable to complete required work within contracted hours in 2019 (The Royal College of Radiologists, 2020). This has had a damaging impact for NHS patients. For example, in the month of December 2019 alone, over 42,000 patients were waiting more than 6 weeks for diagnostic tests and procedures (Office for National Statistics, 2020).

In addition to their workload, expert readers face a difficult challenge when searching for cancers in mammograms. It has been reported that between 20%–30% of cancers present in a screening population are initially missed at screening (Bird et al., 1992; Majid et al., 2003). These missed cancers are either due to features of the mammogram itself or search errors by the reader. Mammographic features responsible for missed cancers include dense breast tissue, improper positioning of the breast during imaging, and poor technique from the mammographer (Bird et al., 1992; Majid et al., 2003). Expert readers are also affected by low prevalence effects; Evans et al. (2013) observed an 18% decrease in cancer miss rate between high (50%) and low (1%) prevalence conditions. Furthermore, there are factors that may cause readers to make errors in how they actually search the images. These include: satisfaction of search (presence of an abnormality, potentially benign, causing the reader to terminate search early and miss another abnormality), search errors (abnormality never fixated), recognition errors (abnormality fixated but not long enough to recognise as an abnormality), and decision errors (abnormality fixated but wrongly dismissed) (Kundel, 2004; Berlin, 2007; Krupinski, 2010).

Computer aided detection (CAD) systems have been developed to detect potential abnormalities in mammograms and other imaging modalities, with the aim of increasing the number of cancers detected with screening. CAD systems may reduce visual search errors made by radiologists by prompting them to look at suspicious areas of images that they may have missed. If CAD systems were proven to improve sensitivity without increasing the number of recalls, single reading with CAD could replace double reading and therefore reduce the workload of radiologists. Furthermore, it is well known that there is a large variation amongst the performance of readers of mammograms (Elmore et al., 1994; Sickles et al., 2002; Elmore et al., 2009), and CAD has the potential to reduce this variance (Jiang et al., 2001).

2.5 Summary

Women aged between 50 and 70 are invited for a mammogram every 3 years as part of the NHSBSP. Subject to the availability of readers, each mammogram is double read, as is standard practice across Europe. Screening programmes have been credited with reductions in mortality rates among the populations of women invited to screening. There is also an associated cost of screening in overdiagnosis, where non-life-threatening cancers are detected and women undergo unnecessary treatments.

Mammograms are the standard choice for breast screening but are known to have reduced sensitivity for dense breasts. There is potential for other imaging modalities such as DBT and ultrasound to be used more frequently in screening for women with dense breasts or using these as an adjunct to mammography.

There are severe problems currently facing the radiology workforce in the UK, with the critical shortage of radiologists expected to increase. There is a need for a solution to reduce workload and improve cancer detection. CAD has the potential to replace the second reader, reduce reader variance, and give a boost to the cancer detection rate. However, the relationship between expert readers and CAD is complex, and this will be discussed in more detail in the next chapter.

Chapter 3

Computer Aided Detection

3.1 Introduction

Systems that aid detection and diagnosis are referred to as computer aided detection (CADe) and computer aided diagnosis (CADx). CADe systems locate and prompt the reader to potentially suspicious regions of an image, such as a mass or microcalcification cluster. CADx systems are used to help the reader decide if abnormal regions are probably benign or malignant, and whether they should be further investigated with a biopsy or additional imaging. In this thesis, the focus will be on detection systems and so 'CAD' will refer to computer aided detection only.

The first methods of computerised analysis of mammograms used optical scanning techniques to produce density images that could then be used for automated analysis of abnormalities (Winsberg et al., 1967). Twenty years later, when screen-film mammograms could be digitised, the first paper describing a system for aiding readers in detecting microcalcificiations in digital mammograms was published (Chan et al., 1987, 1990). The first CAD system to receive approval from the US Food and Drug Administration (FDA) for use as a 'second reader' was ImageCheckerM1000 (R2 Technology, Inc.) in 1998 – a commercial system for the detection of suspicious regions in digitised screen-film mammograms (FDA, 2008). Second reader CAD systems work as follows: the expert reader initially reviews the mammogram unaided, and then again with the use of the CAD output.

The rise of FFDM led to the adaptation of commercial CAD systems to be able to support both digitised screen-film and fully digital mammograms (Yarusso et al., 2000; Li et al., 2006). An estimated 92% of mammograms read in 2016 in the US used CAD systems as a second reader (Keen et al., 2017). More recently, CAD has been used to replace a second reader for cases where the reader agrees with CAD (McKinney et al., 2020), and alternative modes of CAD operation have been explored, where prompts only appear when a region of an image has been queried by the reader (Rodríguez-Ruiz et al., 2019a).

This chapter describes the general methodology of CAD systems and how they are operated, including potential methods that have been explored to improve how readers use them. It will also discuss the usage of CAD around the world and the success of CAD in various observer studies. Finally, this chapter will review studies investigating the relationship between readers and CAD systems, and the ways in which this can be optimised.

3.2 CAD methodology

3.2.1 General CAD workflow

There is a multitude of CAD algorithms, and whilst there are large variations in how they operate and in their implementation, they generally follow the steps outlined in Figure 3.1. First, a digital mammogram is used as the input, either directly from a digital mammography system or from a digitised screen-film mammogram. The image is then processed to enhance anatomical and pathological features specific to the task, and also to remove background (particularly for digitised film images). For example, pre-filtering to enhance the breast edge, which if ill-defined can lead to inaccurate CAD results (Karssemeijer, 1993). This stage also acts to remove noise and any image artefacts that may have appeared during the image acquisition process. Initial segmentation procedures are often required to label image structures such as the pectoral muscle, chest wall, nipple, breast tissue, as well as the image background. Some CAD systems implement a signal detection stage prior to segmentation, where potential lesions are then identified and subsequently segmented from the background and normal mammographic features, and region boundaries are outlined (Nishikawa, 2007).

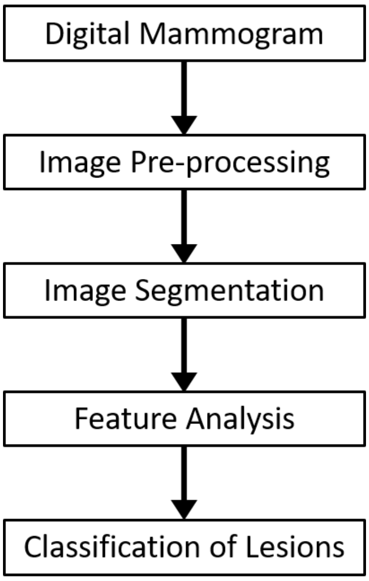


Figure 3.1: General stages of a CAD algorithm.

Features such as shape, size, and texture, are extracted from the segmented regions. Segmentation and feature extraction methods are sensitive to the particular abnormality that the algorithm is designed to detect. These features are then fed into a classifier, such as a support vector machine or neural network, to discriminate between true and false positives, and each feature is assigned a probability of malignancy. Comparisons may be made between different views of the same breast to look for similarities, and the same view of different breasts to look for asymmetries (Hologic Inc., 2014). A threshold is applied to determine which regions should be marked on the image, and those above this threshold, i.e., those with the highest probability, are marked. An example of a typical CAD output is shown in Figure 3.2. Algorithms are first trained on a set of images containing the specific abnormalities of interest, called the training set. Developers are careful not to over-train the algorithms – where the system becomes specific to the training set it cannot be applied effectively to a different set of images.

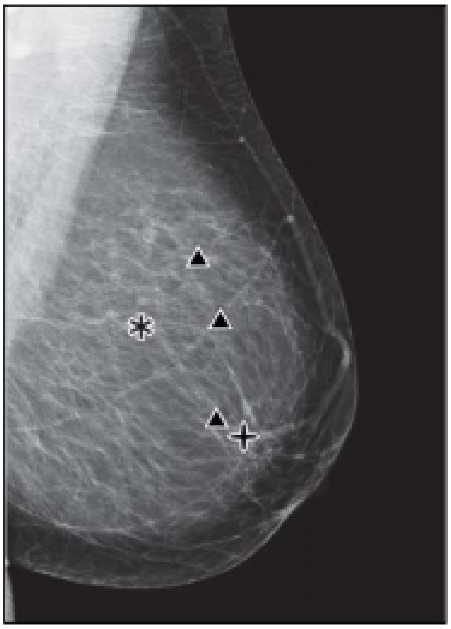


Figure 3.2: Example output of ImageChecker 10.0 on a mammogram Hologic Inc. (2014). The triangle indicates a microcalcification cluster, the asterisk indicates a mass, and the cross indicates a 'malc' – a microcalcification cluster and mass at the same location.

3.2.2 Microcalcification cluster detection

Microcalcifications can be difficult for readers to find in mammograms due to their small size and low contrast, especially in dense tissue. However, it is an easier task for a computer since the properties of microcalcifications are considerably different from normal breast structures (Astley, 2004). Examples of methods for detecting microcalcifications are discussed below, but many more can be found in the literature (Sampat et al., 2005; Nishikawa, 2007; Rizzi et al., 2012; Jing et al., 2015; Yassin et al., 2018). Most methods detect and classify calcifications according to their most important characteristics: their size, shape, density, and distribution (Sampat et al., 2005).

Image enhancement methods are among the simplest approaches, where the contrast between the calcifications and background tissue is increased (Jing et al., 2015). An early method developed by Nishikawa et al. (1995) consists of three stages. Firstly, a difference-image technique uses two filters, one to enhance small structures and another to suppress them. The resulting suppressed image is subtracted from the enhanced image to produce a difference image that highlights small structures and suppresses background tissue. Second, signal extraction is achieved by global thresholding to reduce 98% of pixels to background level, and a series of morphological structuring elements are used to remove signals <3 pixels in size. This stage is completed with local thresholding to suppress signals outside of a chosen factor of the standard deviation of pixel values for that area. Finally, a feature extraction process assesses the properties of the remaining signals to identify potential microcalcification clusters through filtering and grouping of signals.

The process of grouping individual calcifications into clusters is important for classifying malignant clusters and reducing the number of false positives. Typical false positives result from crossing linear structures, calcified arteries and benign calcifications (Nishikawa, 2007). Qian et al. (2002) outline a method for grouping that analyses clusters based on the separation between calcifications rather than the cluster size, and found a reduction in the number of false positives compared to methods using cluster size alone. Global and local thresholding used in methods such as Nishikawa et al. (1995) may not handle image noise optimally. Instead, noise equalisation techniques can be used to improve the contrast of calcifications and improve detection rates (McLoughlin et al., 2004). The results of this method are shown in Figure 3.3.

Another method is multiscale processing, where wavelet transforms are used to filter the high-frequency microcalcifications from their surroundings. The wavelet is tuned to selectively highlight calcifications (Rizzi et al., 2012). Bazzani et al. (2001) combined a wavelet filtering approach with a statistical classifier to reduce the number of false positives. To make predictions, the classifier used the area, average pixel value, edge gradient, degree of linearity, and average local gradient of the calcifications. The differences in statistical properties of microcalcifications compared to the surrounding tissue are utilised by stochastic modeling methods (Jing et al., 2015). One of the earliest of such methods was proposed by Karssemeijer (1992), where a random field approach was used to compare neighbouring pixels and classify whether or not they formed a cluster. A more recent approach used spatial point process modeling to analyse the spatial distribution of clusters (Jing et al., 2010). Pixels are labelled to describe their properties, which are used to analyse calcifications according to various features. These include their amplitude, the average number of calcifications in an image, and the interactions of neighbouring calcifications within a cluster. These parameters inform a detection algorithm to locate clusters within the image.

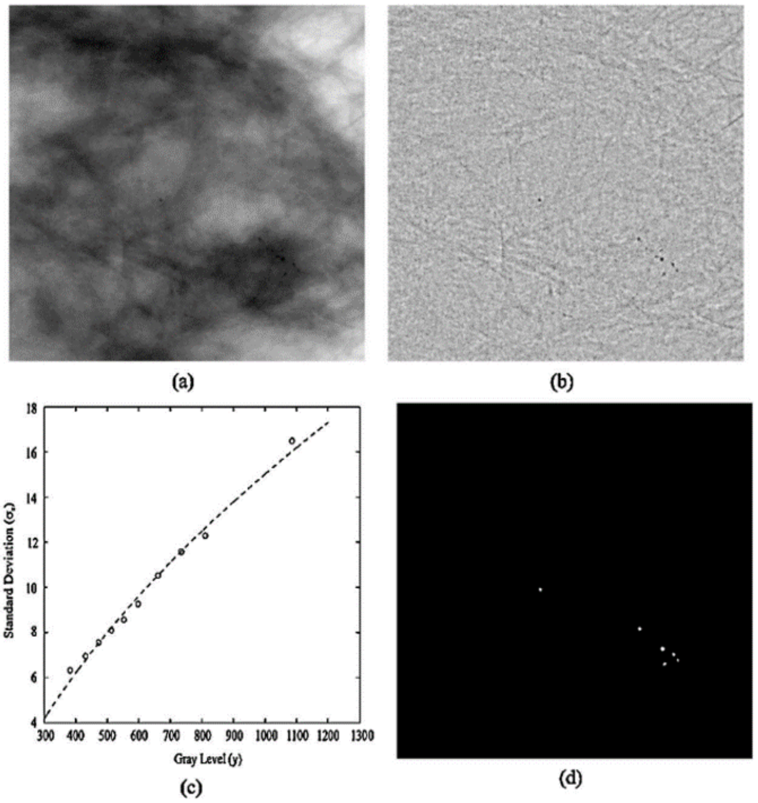


Figure 3.3: Stages of the algorithm developed by McLoughlin et al. (2004). (a) Patch of the input mammogram. (b) Microcalcifications highlighted as extremes of negative contrast at their location. (c) Standard deviation of contrast plotted against gray level across the mammogram, with a fit based on the square root of the noise. (d) CAD output showing detected microcalcification cluster after noise equalisation.

The majority of methods for the classification of microcalcification use machine learning. More recently, detection methods have focused on deep learning, which is discussed in Section 3.2.4. Machine learning methods use data from training images to form a model that typically makes a binary classification of whether or not a microcalcification is present at a given location. One of the most widely used machine learning techniques is a Support Vector Machine (SVM) (Yassin et al., 2018), where objects are represented as a point in N-dimensional space according to their features (where N is the number of features). Each point is separated into one of two categories by a hyperplane, and the SVM works to maximise the distance between points in each category. One study used a wavelet transform for feature extraction followed by an SVM classifier to achieve a sensitivity of 83.5% with 1.85 false prompts per image (Jian et al., 2012).

Another of the most popular machine learning methods is an artificial neural network (ANN) (Yassin et al., 2018). ANNs are composed of connected layers of nodes. The first layer receives the image data and processes the information according to a specific ‘transfer function’ and feeds an output a node in the next layer multiplied by a given weight, which in turn processes the data and passes it to nodes in the next layer, etc. The final layer outputs a classification for the given task. The weights are adjusted through training according to the classification error rate, in order to teach the network how to effectively classify the image data. ANNs were shown to be an effective method for classification of benign and malignant microcalcifications 25 years ago, achieving a better detection rate than radiologists (Jiang et al., 1996). Various network architectures have been proposed for classification, reduction in reader variability, and patient risk estimation (Ayer et al., 2013).

Commercial CAD systems use a combination of these methods and have a high sensitivity for locating microcalcifications. ImageChecker CAD 10.0 reports sensitivities of up to 99% for biopsy proven microcalcifications at a specificity of 29% – this is the standalone performance, not the performance of a reader using the software (Hologic Inc., 2014). Several studies have measured the standalone performance of commercial CAD systems that are routinely used in mammography clinics in the US, the results of which are given in Table 3.1. For the detection of microcalcification clusters, CAD sensitivity ranged between 83% and 100%, with a false prompt rate between 0.26 and 1.76 per case.

Table 3.1: Results from multiple studies that measured stand-alone CAD performance on three abnormality types. Two CAD systems routinely used in US clinics were measured: ImageChecker (Hologic, Inc.) and SecondLook (iCAD Inc.). Sens = sensitivity and FPs/case = false prompts per case. N/R = ‘not reported’ and ‘$-$’ indicates that it was not applicable in the study. The false prompt rate reported by some of the studies was underestimated since it was measured using only abnormal cases.

				Mass		**Microcalcification cluster**		**Architectural distortion**		**Total**	
Study	**CAD system**	**No. images**	**No. cancers**	**Sens (%)**	**FPs/ case**	**Sens (%)**	**FPs/ case**	**Sens (%)**	**FPs/ case**	**Sens (%)**	**FPs/ case**
Baker et al. (2003)[1]	ImageChecker	43	45	-	-	-	-	49	0.7	49	0.7
	Second Look			-	-	-	-	33	1.27	33	1.27
Malich et al. (2003)	Second Look	208	264	90	3.24	93	0.8	-	-	91.3	4.04
Yang et al. (2007)	ImageChecker	203	103	89	0.95	100	0.5	-	-	96.1	1.8
Kim et al. (2008)[2]	ImageChecker	93	83	91	0.81	100	0.32	-	-	95	1.14
				89	0.78	100	0.26	-	-	93	1.04
Sadaf et al. (2009)	Second Look	127	127	89	N/R	100	N/R	71	N/R	91	N/R
The et al. (2009)	Second Look	123	123	92	N/R	93	N/R	75	N/R	94	2.3
van den Biggelaar et al. (2009)	Second Look	1048	51	N/R	2.3	N/R	0.7	-	-	78	2.9
Bolivar et al. (2010)	Second Look	151	151	89	N/R	98	N/R	-	-	93	2.5
Scaranelo et al. (2012)	Second Look	122	124	-	-	96	1.76	-	-	96	1.76
Cole et al. (2012)[3]	Second Look	161	161	69	N/R	83	N/R	88	N/R	74	2.57
	ImageChecker			72	N/R	83	N/R	75	N/R	74	2.07
Murakami et al. (2013)	Second Look	152	152	98	1	100	0.8	75	N/R	91	1.8

[1] Baker et al. (2003) measured the performance of both CAD systems for the detection of architectural distortions.

[2] Kim et al. (2008) measured CAD performance for initial and follow-up examinations, giving two sets of performance measures.

[3] Cole et al. (2012) measured the performance of both CAD systems for the detection of all three abnormality types.

3.2.3 Mass and architectural distortion detection

A dense region with radiating spicules is indicative of a mass and radiating spicules with no central density are indicative of an architectural distortion, and these are the features CAD algorithms search for. Masses with an irregular shape, with ill-defined or spiculated margins are more likely to be malignant, and round or oval masses with well-defined margins are more likely to be benign (Bassett and Conner, 2003). Detecting masses and distortions is often more difficult than microcalcification clusters, due to the complex nature of normal features in mammograms, such as fibroglandular tissue and neurovascular structures. Image enhancement methods are often used to remove these structures from the mammogram, to improve the efficacy of mass detection. These processes must be careful not to impact on the spiculation of masses through smoothing, which would reduce the ability of algorithms to detect them, and so some methods may instead classify these regions rather than remove them (Karssemeijer, 2015).

In a mass detection scheme developed by Rojas-Domínguez and Nandi (2008), mammograms are initially enhanced to improve the contrast between structures and the background. The transform that is applied to each pixel is based on the statistics of the local neighbourhood of that pixel. Multiple binary image thresholds are applied to the image to segment image features. The characteristics of the features are used to rank and select potential masses. Another method that is used in both calcification and mass detection is wavelet analysis. A technique designed for mass detection in dense tissue was developed by Sakellaropoulos et al. (2006). Dense tissue is identified within the mammogram, and wavelet filtering is applied to those regions to extract features that could be used for mass classification. Image enhancement methods may produce a large number of false positives due to the amount of dense normal tissue in mammograms resulting in a large number of potential mass regions. False positives may be reduced by performing bilateral comparisons to measure the similarity of potential regions across breasts (Li et al., 2015).

Machine learning techniques are widely used for mass detection and classification, particularly SVMs and ANNs. The considerable variability in the appearance of masses poses a difficult problem for CAD systems, since it is difficult to define features that correspond to a wide range of lesion shapes, sizes and contrast. A method proposed by Campanini et al. (2004) omits the feature extraction step as a way to get around this problem. Instead, cropped image data that has been filtered with a wavelet transform to highlight structures in the image is classified using an SVM. A second SVM is used to reduce the number of false positives, and the overall sensitivity of the system was reported as 80% with 1.1 false positives per image. Other studies extract large numbers of features from images to better classify potential masses. Fauci et al. (2004) extract 12 features (based on the morphology, pattern and intensity) from ROIs and feed them into an ANN for classification with an area under the receiver operating characteristics curve (AUC) of 0.856. Varela et al. (2007) used 20 features based on morphology, texture, contour-related, gray level, and those from an image produced using an iris filter tuned for mass detection that transforms the image to highlight regions with a gradient towards a central point. The features were again used with an ANN, achieving a sensitivity of 88% with 1.02 false positives per image.

Many studies have directly compared the performance of machine learning techniques on the same dataset (Yassin et al., 2018). In a study by Lesniak et al. (2012), five machine learning classifiers (including an SVM and an ANN) were trained on a database of 2,516 mammograms and used for mass detection and reduction of false positives. The SVM was significantly better at both detection and false positive reduction compared with the other methods. García-Manso et al. (2013) compared the performance of an SVM and an ANN on the classification of masses in 2,620 cases, containing 2,324 benign and malignant masses. The performance of the classifiers was similar, with an AUC of 0.937 for the SVM and 0.925 for the ANN.

Methods may focus on the detection of spiculation (Karssemeijer, 2015), since this is a key feature of many malignant masses. Sampat et al. (2008) enhance spiculations in mammograms by applying spiculated lesion filters. These filters are designed as ‘matched filters’, based on the structures of spiculated masses. In addition to enhancing spiculations, a set of Gaussian filters are used to detect the central mass, as well as difference-of-Gaussian filters to suppress the normal linear structures. To obtain the overall algorithm output, the output from the normal structure filter is subtracted from the sum of the filter outputs for spicules and central masses. Instead of highlighting spiculated masses, Muralidhar et al. (2010) developed a method that traces the path of individual spicules using active contours that grow and deform along spicules, shown in Figure 3.4.

Taplin et al. (2006) demonstrated that the sensitivity of CAD was significantly worse for masses and architectural distortions compared to calcifications (67% versus 86%). As shown in Table 3.1, standalone CAD sensitivity for masses ranged between 69% and 98%, with a false prompt rate between 0.81 and 3.24 per case. This means segmentation methods are often not very specific, and normal features are considered as possible candidates for abnormalities.

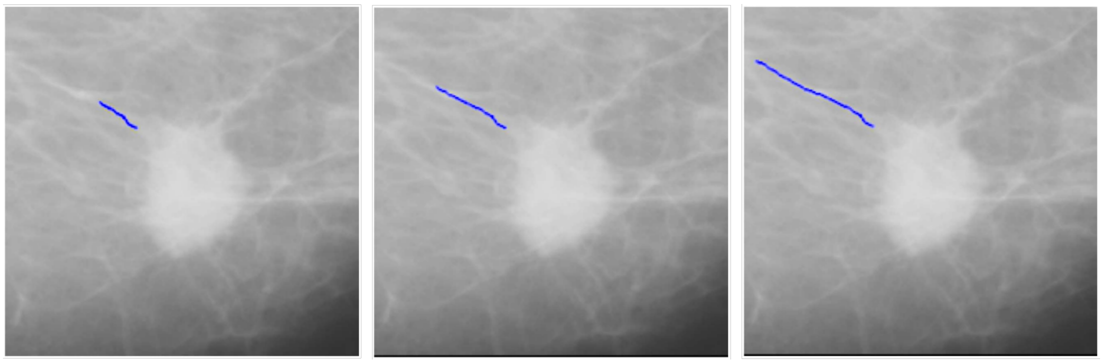


Figure 3.4: Visualisation of the active contour growth along a spicule over 3 iterations, adapted from Muralidhar et al. (2010). The process of growth and deformation continues until a stopping criteria is met (when the curvature exceeded 30°).

Since mass detection algorithms will often search for signs of spiculations, architectural distortions may also be detected as a result. However, without directly searching for distortions, this can often lead to low sensitivity. The standalone CAD performance for architectural distortions was measured in five studies, shown in Table 3.1. The sensitivity ranged between 33% and 88% with a false prompt rate between 0.7 and 1.27 per case, with the false prompt rate only reported by Baker et al. (2003). The relatively low performance of these systems means that it is desirable to incorporate methods specifically designed for the detection of architectural distortions.

Karssemeijer and te Brake (1996) describe a method for the detection of stellate objects in mammograms (lesions and architectural distortions). Initially, estimates of the orientation of line-based structures are made. These estimates are then used to identify radial patterns. This method achieved a sensitivity of 90% with 1 false positive per image. Breast tissue has been shown to approximate a fractal object, and Tourassi et al. (2006) use this for the detection of distortions. The fractal dimension (measure of pattern complexity) was estimated for both normal image patches and those containing a distortion, with a significantly higher fractal dimension estimated for distortions.

A similar method by Rangayyan et al. (2010) used fractal and texture analysis combined with various classifiers (including an ANN and an SVM) for the detection of distortions in prior mammograms of interval cancers. The best classifier achieved an 80% sensitivity with 7.6 false positives per image, an improvement over previous work using prior mammograms. Another study used three texture filters to enhance architectural distortions and suppress normal breast tissue structures (Yoshikawa et al., 2014). This produced a sensitivity of 82% with 1.06 false positives per image.

3.2.4 Recent advances

Recent studies have proposed promising CAD systems using deep neural networks (Kooi et al., 2017; Dhungel et al., 2017; Chougrad et al., 2018; Ragab et al., 2019; Rodríguez-Ruiz et al., 2019a). The major difference between conventional machine learning methods and deep learning is in feature selection and extraction. In deep learning, features are learned from the data itself, rather than using features defined by humans (LeCun et al., 2015). A comparison between the general architecture is shown in Figure 3.5. Consequently deep learning methods require large amounts of training data and this, combined with the increased number of network layers, requires high computational power (Ramadan, 2020). A review of studies that directly compare conventional machine learning and deep learning demonstrated that, in the majority of cases, deep learning outperforms conventional machine learning methods (Jiménez-Gaona et al., 2020).

In an effort to reduce the recall rate of women attending breast screening, the Digital Mammography DREAM Challenge (DM Challenge) was initiated, in which teams from across the world presented machine learning or deep learning methods for determining the cancer status of a participant given their mammograms (DM Challenge, 2017). The winners of the challenge achieved an AUC score of 0.8735 (Nikulin et al.,

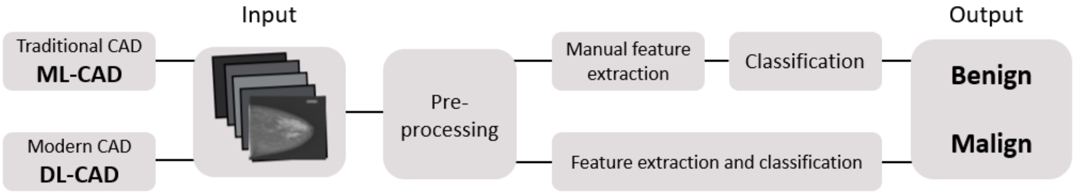


Figure 3.5: Comparison between the structure of traditional machine learning and deep learning methods, adapted from Jiménez-Gaona et al. (2020). ML-CAD requires features to be defined by a human, whereas deep learning CAD learns the features itself.

2017). Ribli et al. (2018) achieved second placed in the challenge with an AUC of 0.85 and reported a 90% sensitivity with 0.3 false positives per image on a separate dataset of 115 mammograms.

The high performance of deep learning CAD is promising and gives rise to other potential uses in breast screening that is not possible with traditional CAD. Deep learning CAD could be used in the same way as traditional CAD as a second reader, where the reader reviews the mammogram again with prompts after an initial unaided review. Furthermore, with a sensitivity and specificity comparable with expert readers, CAD could also be used as an independent second reader (Geras et al., 2019). Furthermore, CAD could be used as a pre-screener, selecting the cases that should be reviewed by an expert reader in an effort to reduce workload without a drop in detection rate (Raya-Povedano et al., 2021) – discussed further in Section 5.4.3. Pre-screening is possible by using case-based scores that indicate the probability of malignancy for that case, in addition to lesion-based malignancy probability scores, which is explored in Chapter 7.

A limitation of many studies measuring the performance of these systems is that they use image databases that are homogeneous, where the cases are all from a single manufacturer or do not contain a wide range of abnormalities, and so may not generalise well across clinics. However, Rodriguez-Ruiz et al. (2019b) measured the standalone performance of their deep learning CAD system with 2,652 cases from four different manufacturers and multiple reader assessments for each case. They demonstrated that their AI-CAD achieved an of AUC of 0.840, compared to an average AUC of 0.814 across 101 expert readers. However, this study used an enriched dataset, with a cancer prevalence of 24.6%. Therefore, there still needs to be large-scale clinical studies to compare performance with and without deep learning CAD, such as those conducted for conventional CAD (discussed in Section 3.5).

Deep learning CAD is a rapidly moving field across medical imaging and particularly breast imaging. It has been referred to as the new frontier (Gao et al., 2019) and CAD 2.0 (Kohli and Jha, 2018). However, the challenges that have faced traditional machine learning CAD remain when you are tasked with merging human and machine, and this will be discussed further in Section 3.6.

3.3 Analysis of performance

Before discussing how CAD is used and observer search performance is evaluated, it is important to define the metrics that are used for this analysis. The method used to analyse performance of CAD and readers is specific to the task and should be modified accordingly.

True or false positives or negatives can have different meanings depending on the task. In the context of mammography and CAD they are usually defined as by either case-level or region-level, defined in Table 3.2.

Table 3.2: Definition of performance metrics in mammography for both case-level and region-level.

	Image-level	**Region-level**
True Positive (TP)	Mammogram containing cancer correctly identified as a cancerous	Mark placed by a reader or CAD indicating a cancer on a cancerous region
False Positive (FP)	Normal mammogram incorrectly identified as containing cancer	Mark placed by a reader or CAD indicating a cancer on a non-cancerous region
True Negative (TN)	Normal mammogram correctly identified as normal	Absence of a reader or CAD mark on a non-cancerous region
False Negative (FN)	Mammogram containing cancer incorrectly identified as normal	Absence of a reader or CAD mark on a cancerous region

Using the definitions in Table 3.2, sensitivity is defined as:

$$sensitivity = \frac{TP}{TP+FN} \tag{3.1}$$

Furthermore, specificity is defined as:

$$specificity = \frac{TN}{TN+FP} \tag{3.2}$$

While it is useful to use sensitivity and specificity in observer studies, often the cancer detection rate (proportion of cancers in image set correctly identified) is used instead. This will usually be reported in addition to the recall rate of the women in the study (number of women recalled for an additional follow up scan) or the number of false positives responses per case.

Another important metric for observer performance analysis is the sensitivity index, or d'. It is defined as $d' = z(H) - z(F)$, where $z(H)$ and $z(F)$ are the z transforms of the hit rate H and false alarm rate F, respectively (Green and Swets, 1966). In addition to sensitivity and specificity, there is also the positive predictive value (PPV) and negative predictive value (NPV). The PPV gives the fraction of all positive cases that are true positives, and the NPV the fraction of negative cases that are true negatives. These are defined as PPV = TPs/(TPs+FPs) and NPV = TNs/(TNs + FNs). These provide a measure of performance that is dependent on the prevalence of true and false cases and give the likelihood of a positive or negative outcome.

One of the most common methods for evaluating observer performance is by plotting a receiver operating characteristic (ROC) curve (Figure 3.6) - a plot of the true positive rate (sensitivity) against false positive rate ($1-$specificity) (Hanley and McNeil, 1982). The area under the ROC curve (AUC) represents the probability that an abnormal image is distinguished from a normal image, and takes a value between 0 and 1. The AUC is described as the 'figure-of-merit' for ROC analysis, and is often used to compare performance between two conditions or imaging modalities. An AUC of 0.5 represents chance level, illustrated by the dashed line in Figure 3.6, and a value of 1 represents a perfect classifier of healthy and abnormal. It is not always appropriate to take the whole area under the ROC curve, for example when the low sensitivity regions are not of interest (such as for CAD), and instead the partial AUC (pAUC) can be used instead (Ma et al., 2013). The pAUC is the area under the curve above a chosen threshold, divided by the total area above that threshold.

The ROC does not take into account the location of abnormalities or cases with multiple abnormalities, limiting readers to a single rating per image. Therefore, it may be more appropriate to use the free-response ROC (FROC) method to evaluate performance. Here, readers are no longer limited to reporting a single lesion per case and provide a rating to each region they mark (e.g. from 1 to 5 or 1 to 100) denoting their confidence that the marked region is abnormal. TPs and FPs are defined according to lesion localisation – if a mark is placed within a predefined boundary from a lesion it is classified as a TP, otherwise it is a FP. The FROC curve is plotted as the TP fraction versus the number of FPs per image. The FROC curve does not have an associated figure-of-merit. Therefore, to obtain a meaningful equivalent of the AUC in ROC analysis, an alternative FROC (AFROC) curve can be plotted instead. The AFROC

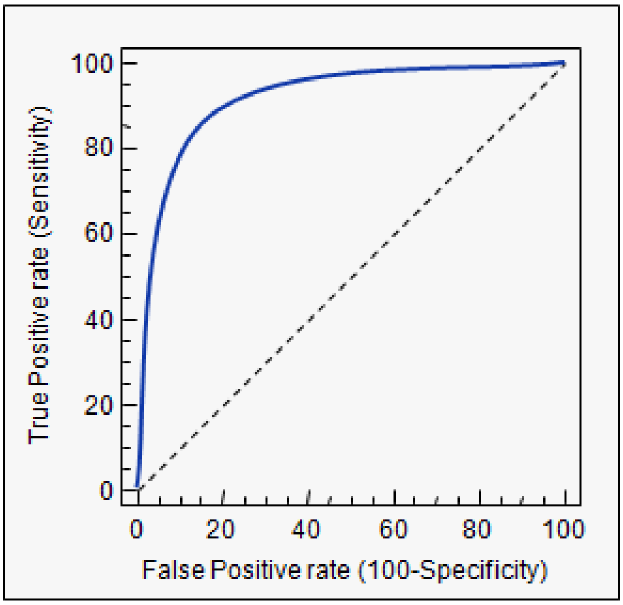


Figure 3.6: An example ROC curve. For an ideal observer, the curve would pass through the top left corner, which corresponds to 100% sensitivity and specificity. Image from MedCalc (2021).

curve is the number of marks indicating lesions divided by the total number of lesions versus the false positive fraction (FPF), and the area under this curve can be used as a figure-of-merit for FROC analysis.

However, in AFROC analysis, for cases with multiple responses, only the highest rated response is considered and the other false positive responses are ignored (Chakraborty and Berbaum, 2004). Furthermore, cases with multiple lesions have a stronger influence on the results than cases with a single lesion. To overcome this, jackknife free-response ROC (JAFROC) analysis was introduced (Chakraborty and Berbaum, 2004), which does not make the same assumptions. Software is available to run JAFROC analysis, available on Github (`https://github.com/dpc10ster/WindowsJafroc`). JAFROC ignores false positive responses on normal images, and alternative JAFROC models were introduced to take this into account (JAFROC1), which was shown to have the highest statistical power compared to other FROC methods (Chakraborty and Yoon, 2009).

Studies of the effectiveness of CAD can be categorised as longitudinal, where cancer detection rates are compared before and after the introduction of CAD into a screening unit, and cross-sectional, where reader performance is compared for reading without and then with the aid of CAD. Nishikawa and Pesce (2009) explain how these evaluation methods will yield different conclusions (see Figure 3.7): Each year, a radiologist is presented with 100 new cases of breast cancer, of which they detect 80 and miss 20, and the missed 20 are subsequently detected the following year (in addition to the 80 new cases that year). After 6 years, CAD is introduced and an additional 10 cancers of the new 100 cancers are detected (a total of 110 cancers for year 6). After that year, 100 cancers will be detected each year, 80 new cases unaided, 10 new from CAD and a further 10 missed from the previous year. A longitudinal study would conclude there is no difference in cancer detection rate from CAD being introduced (year 6 would be considered noise), whereas a cross-sectional study would report an increase of 11% in detection rate. The fact that CAD could help find extra cancers without increasing the detection rate (besides year 6) was due to the populations in two time periods being different, and so the reduction in cancer prevalence after year 6 is a direct result of introducing CAD. Therefore, cross-sectional studies are better suited to evaluating the efficacy of CAD systems.

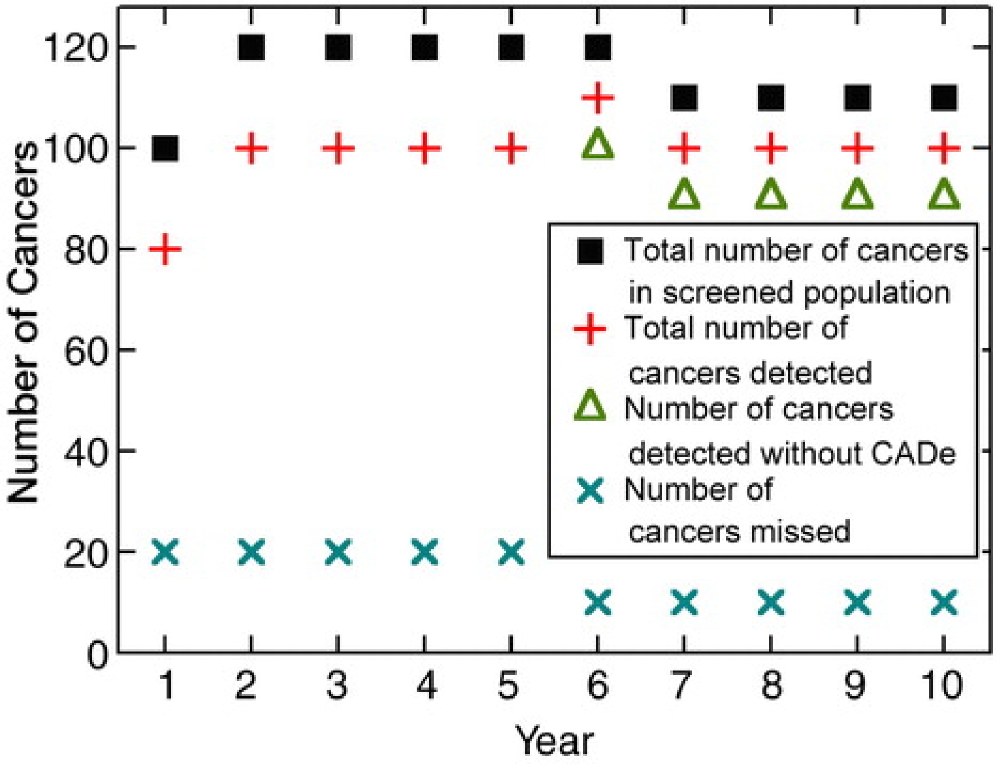


Figure 3.7: Number of cancers over time. CAD is introduced in year 6, and in longitudinal studies, the increase in cancers detected is dismissed as noise. Image taken from Nishikawa and Pesce (2009).

3.4 Modes of use and prompting techniques

The threshold that a CAD algorithm applies to determine which prompts to display on the image is called the *operating point*. The choice of operating point for an algorithm is a trade-off between sensitivity and specificity (Figure 3.8) – higher sensitivity comes at the cost of an increased number of false positives per image. CAD systems detecting both masses and microcalcifications use different algorithms and operating points for each abnormality type, which are either preset or chosen by the reader.

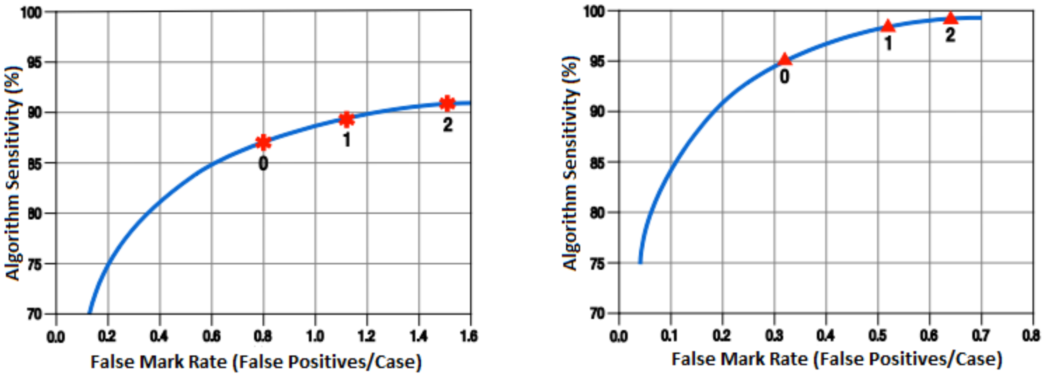


Figure 3.8: Algorithm sensitivity versus false mark rate for the ImageChecker 10.0 CAD system for masses (left) and microcalcifications (right). There are three choices of operating points for each abnormality: relatively low sensitivity and high specificity (0) and relatively high sensitivity and low specificity (2), and a midpoint between them (1). Image from (Hologic Inc., 2014).

The operating points determine the number of prompts that are displayed. CAD systems may also limit the maximum number of prompts that are displayed per image and per case. For example, with ImageChecker, the maximum number of prompts per image is 4 for microcalcification clusters, and 2 for masses and ‘malcs’ (mass and microcalcification cluster at the same location) (Hologic Inc., 2014). These are limited further when the images are processed as part of a case, at 8 per case for microcalcification clusters, and 4 for masses and malcs. When the number of prompts is limited, the prompts associated with lesser suspicious regions are removed.

The type of prompt used is an important factor for CAD systems. While some systems, such as ImageChecker (Hologic Inc., 2014), use solid prompts overlaying suspicious regions (Figure 3.2), other systems use prompts which outline the abnormalities (iCAD, 2016; ScreenPoint, 2021). This has the benefit of not obscuring the marked area. The prompt style can change the readers behaviour; for example, using ImageChecker’s EmphaSize feature which scales the prompt size by the probability of malignancy, larger prompts were shown to be significantly more likely to be recalled (Gilbert et al., 2008a). ScreenPoint has an interactive support function, where CAD prompts only appear when their location is queried by the reader.

The trust between a human and a computer aid will often determine how they use it (Parasuraman and Riley, 1997). Muir (1987) modeled the trust between humans and machine by extending the model of trust between humans, where humans were more likely to believe computer aids they consider reliable and less so those that let them down. Individuals may have an initial bias towards a computer aid leading to inappropriate levels of trust (Dzindolet et al., 2003). In this study, participants trust was improved when they were shown why the aid made mistakes. Therefore, understanding the trust between the reader and machine can be highly beneficial to the output of the system.

There are a number of ways that trust between the reader and the CAD system can be improved, such as providing confidence ratings on the prompts. This is something that is done with interactive CAD systems (such as ScreenPoint) and means that a reader is more likely to believe the output of a CAD system if the prompts they would have dismissed anyway have a low confidence rating (Jorritsma et al., 2015). Giger et al. (2002) proposed an 'intelligent workstation', where readers are provided with various visual prompts, such as likelihood of malignancy for a lesion and similar cases with known diagnoses. A clinical example of this is ImageChecker's PeerView mode, where the regions that are detected by CAD can be queried by the reader and the masses and individual calcifications are highlighted for further evaluation (Hologic Inc., 2014).

Similar to the prompt confidence ratings with interactive CAD, Cunningham et al. (2016) describe an 'analogue' CAD approach, where the probability that the prompt is marking a target is denoted purely by its colour (see Figure 3.9). They demonstrated that non-expert observers were more efficient at distinguishing between specific clusters of coloured dots using analogue CAD than traditional binary CAD that simply indicated whether a target was present or not. ScreenPoint (2021), for example, use a combination of prompt confidence scores and coloured prompts in their software, ranging from yellow (low confidence) to red (high confidence).

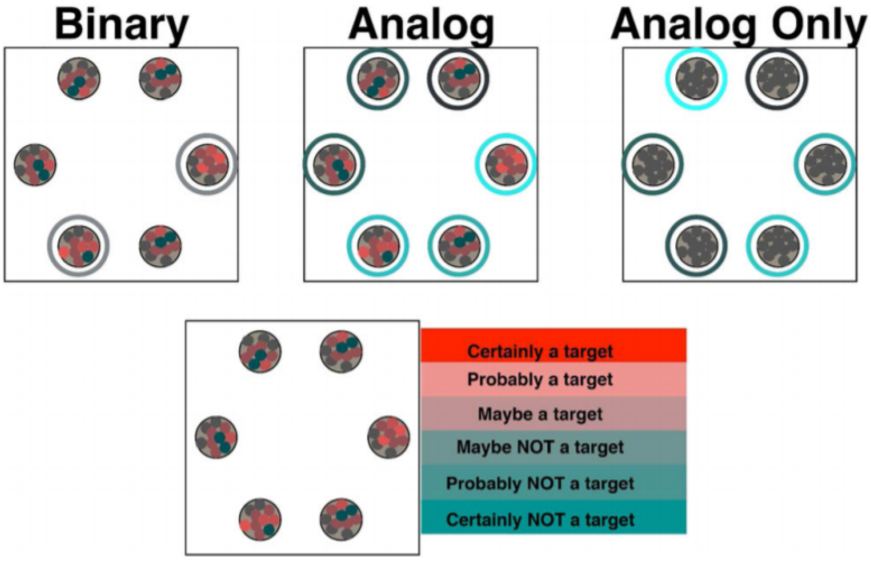
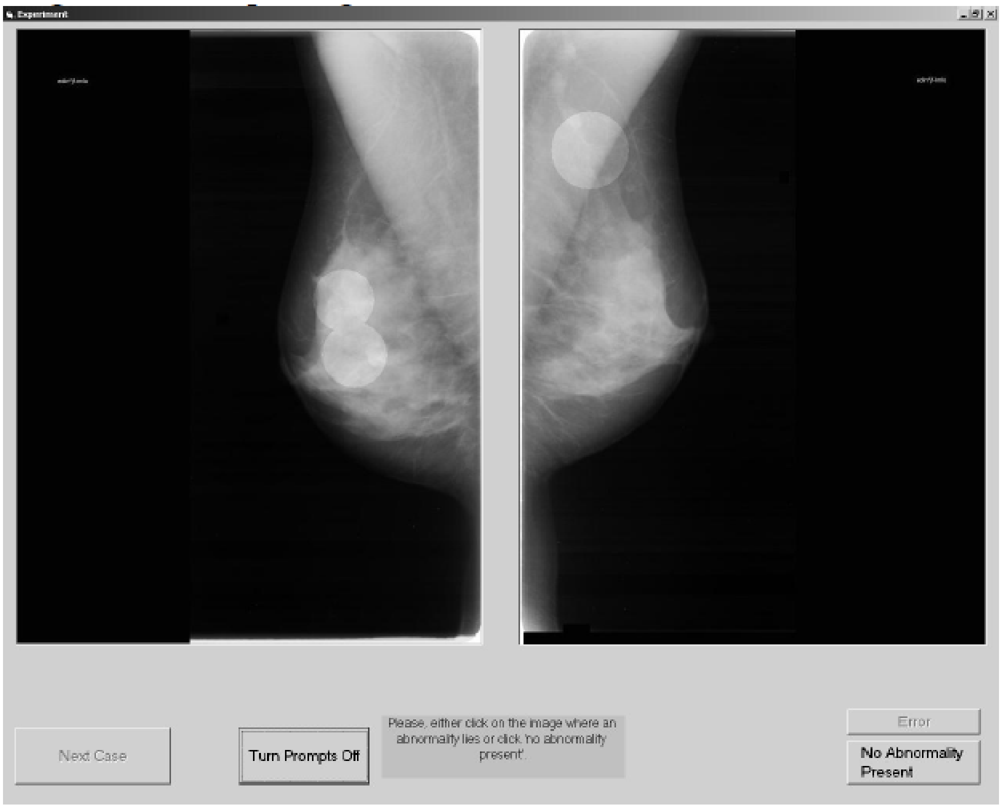


Figure 3.9: Image from Cunningham et al. (2016) showing analogue CAD prompt approach. The task involved participants selecting the item with the average colour most likely to be a target (red). Participants also selected their confidence in their decisions on the 6-point scale shown. Note the difference between binary and analogue CAD prompts, binary prompts are a single colour with a lower operating point so mark the two stimuli most likely to be a target. Whereas analogue CAD marks each stimuli with varying colour to denote likelihood of it being a target, ranging from grey (most likely a distractor) to turquoise (most likely a target).

Instead of markers, it has been proposed that highlighting may be more effective (Hatton et al., 2004; Kneusel and Mozer, 2017). Hatton et al. (2004) found that using CAD prompts which made subtle changes to the colour or brightness of target regions or the background (Figure 3.10) interfered less with the way that participants were looking at the images when compared to conventional prompts, which place markers directly over the region of interest. Therefore, subtle prompts more closely resembled their viewing behaviour under unaided conditions.



Please, either click on the image where an abnormality lies or click 'no abnormality present'.

Figure 3.10: Image from Hatton et al. (2004) showing one of their subtle prompt techniques – the suspicious regions have had their brightness increased relative to the background.

Kneusel and Mozer (2017) proposed a technique called 'soft-highlighting', which combines subtle and analogue prompts; targets are highlighted with a saliency based on the confidence level of the classifier. They found that soft-highlighting increased performance of naive observers searching for targets in satellite images, compared to 'hard-highlighting' where boxes were drawn around potential targets. An interesting method was proposed by Chen and Gale (2010*a*): it uses eye tracking to record search patterns of mammograms and combines this information with the CAD system to display more informative prompts. For example, different prompt styles can represent areas that have been fixated but not marked by CAD, areas that have been marked by CAD but not fixated, and areas that have been both marked by CAD and fixated and therefore more likely to be an abnormality (Chen and Gale, 2010*a*).

3.5 Usage and effectiveness

Only four of the 26 countries or regions in the ICSN discussed in Section 2.1 use CAD in their screening programmes: the US, Poland, Luxembourg, and the Navarra region of Spain (National Cancer Institute, 2012*b*). In Europe, double reading is standard practice, rather than single reading with the addition of CAD. Unlike the US, in Europe there is no reimbursement for using CAD, which could partly explain this difference. A major reason why single reading with CAD (or second reader CAD) has not replaced double reading is due to contradictory reports on the effects of doing so.

Skaane et al. (2007) reported that single reading with CAD achieved a higher cancer detection rate than independent double reading for both screen-film and digital mammography. However, they used a relatively small sample of 3,683 women with a total of 55 cancers and could not determine how CAD affected the recall rate. A larger study by Gilbert et al. (2008b) of 31,057 women with a total of 227 cancers found that single reading with CAD was equivalent to double reading in terms of sensitivity but resulted in a significant increase in recall rate. A meta-analysis reported that double reading increases cancer detection rate and recall rate, but double reading with arbitration (where readers confer on the cases or a third reader reviews them) increases detection rate while also decreasing recall rate (Taylor and Potts, 2008). Whereas, CAD did not have a significant impact on the detection rate and increased the number of recalls.

A review by Azavedo et al. (2012) concluded that there was insufficient evidence to suggest that single reading with CAD is equivalent to double reading. A review of five studies by Henriksen et al. (2018) showed that there was no significant difference in sensitivity for single reading with CAD compared to double reading, with only one study (Gilbert et al., 2008b) having a significant increase in recall rates for single reading with CAD relative to double reading. A deep learning CAD methodology developed by Google was able to achieve a non-inferior performance to double reading in terms of both sensitivity and specificity compared to 6 radiologists (McKinney et al., 2020). They simulated how this could be used clinically; replacing the second reader in instances where there is agreement between the first reader and CAD with the first reader decision taken as final, and the usual arbitration process where there is disagreement. This resulted in an equivalent performance but with an 87.98% reduction in workload for the second reader.

Several cross-sectional studies have evaluated the impact of CAD on detection and recall rates (Freer and Ulissey, 2001; Helvie et al., 2004; Birdwell et al., 2005; Dean and Ilvento, 2006; Ko et al., 2006; Morton et al., 2006; Georgian-Smith et al., 2007). The increase in detection rates due to CAD range from 0% (Georgian-Smith et al., 2007) to 19.5% (Freer and Ulissey, 2001) with a mean of 9.3%, and a mean increase in recall rate of 12.4%. The increase in sensitivity with CAD is promising but has not been convincingly replicated in clinical settings without an unacceptable increase in recalls.

The ability of readers to identify false CAD prompts will determine how prompts influence the recall rate. This ability, along with the reader's trust in the CAD system, will depend in part on the reproducibility of prompts. This describes how often a CAD system will prompt regions that are similar in appearance and structure. This has been measured directly by comparing the position of prompts across digital mammograms from the same women with a short interval between acquisitions. Kim et al. (2008) reported that true prompts had a significantly higher reproducibility than false prompts (85% versus 9%). Another study found that false prompts on normal digital mammograms had a reproducibility of 12% (Kim et al., 2009). These results are in agreement with previous work with screen-film mammograms that were repeatedly digitised and assessed by CAD, with a reproducibility of 13.8% for false prompts (Tiew et al., 2008). Changes in patient position between image acquisitions can result in changes in the distribution of breast tissue and therefore how the image is processed by CAD. The low reproducibility of false prompts means that readers will be less likely to recognise them as incorrect and thus more likely to mark them as a cancer.

Lehman et al. (2015) conducted a large-scale study investigating the effectiveness of single reading with CAD versus single reading alone. This study compared the performance of 271 radiologists who examined a total of 495,818 mammograms with CAD and 129,807 without. These mammograms were read between 2003 and 2009. No benefit with CAD was found: the sensitivity was 85.3% with and 87.3% without CAD, and specificity was 91.6% with and 91.4% without CAD. The impact of CAD on intrareader performance was also measured by comparing the 107 readers who read cases both with and without CAD. In this case, sensitivity significantly decreased with CAD from 86.9% to 83.3% (OR=0.53). Specificity was 90.7% with CAD and 89.6% without, but this was not significant (OR=1.02).

These results are counter-intuitive to the use of CAD as a second reader, where an improvement in both sensitivity and recall rate is expected. This is likely an outcome of CAD misuse in two possible forms (Nishikawa and Kyongtae, 2018). First, readers may be using CAD to reduce their recall rate by not recalling cancers that are not prompted by CAD, rather than using it to detect cancers they may have otherwise missed, which results in the lower sensitivity observed in Lehman et al. (2015). It may also be that readers use CAD as a concurrent (from image onset). This again can result in a reduction in sensitivity for unprompted cancers (Alberdi et al., 2004; Zheng et al., 2004).

Despite AUC values of the latest deep learning CAD algorithms continuing to increase, there are various human-CAD interaction issues that need to be addressed, which Nishikawa and Kyongtae (2018) argue cannot be resolved simply by improving the accuracy of the algorithm.

3.6 Human-CAD Interaction

The full potential of CAD will not be achieved if the reader-CAD interaction is not optimised. A key issue with CAD is the high number of false prompts, typically 0.5 per image (The et al., 2009; Cole et al., 2012), compared to an estimated 0.3% cancer prevalence in screening mammography (Evans et al., 2013). In current CAD systems high sensitivity comes at the cost of high false positives. Consequently, readers are more likely to dismiss prompts without properly reviewing them (Philpotts, 2009). This is known as under-trust, where readers are under-reliant on a useful aid (Parasuraman and Riley, 1997). For example, Nishikawa et al. (2012) showed that readers failed to act on 71% of correct CAD prompts marking cancers in mammograms. However, given that the system had a false-positive rate of 0.6 per image and the pressures on readers to maintain low recall rates, this result is not entirely surprising since deciding which prompts to act on is difficult (Philpotts, 2009).

An alternative to overcome the high false positive rate of traditional CAD is interactive CAD systems such as ScreenPoint (2021). Since on any given image the prompts are only seen by the reader if they query the locations where they are present, you can afford to operate CAD at an operating point with a higher sensitivity and a greater number of false positive prompts, as they are unlikely to be seen anyway. One study using interactive CAD reported to have a significant improvement in the partial AUC (pAUC) compared to unaided and conventional CAD (Hupse et al., 2013), and another showed a significant increase in AUC compared to unaided reading (Rodríguez-Ruiz et al., 2019a).

Less experienced readers tend to rely more on CAD (Nishikawa et al., 2012; Hupse et al., 2013). This is known as over-trust and is a form of automation misuse (Parasuraman and Riley, 1997). Over-trust in false prompts may also explain the increase in recall rates in many CAD studies. Another potential issue with over-trust is that the absence of a prompt could give the reader false reassurance that no cancer is present. In a study by Alberdi et al. (2004), the detection rate of cancers in mammograms decreased from 77% without CAD to 55% with CAD for incorrectly marked cancers (prompt in wrong location), and from 54% to 33% for unmarked cancers (no prompt present). In this study, overreliance on prompts may have caused a drop in sensitivity for the cases where prompts failed to mark a cancer either by being in the wrong location or were absent entirely.

Zheng et al. (2004) investigated how performance varied with different CAD operating points and found that CAD with 80% mass sensitivity and 0.5 false prompts per image improved performance, but 80% sensitivity and 1.2 false prompts per image significantly decreased performance. However, it is important to note that the readers in these studies were reading the cases with CAD initially, not as a second reader, which goes against recommended usage. In fact, when the CAD output was displayed after an initial interpretation, Zheng et al. (2004) reported that there was little difference in the overall performance of readers.

This reduction in sensitivity for targets that were failed to be marked by CAD was also observed for non-expert readers (Drew et al., 2012). Another study with non-expert participants reported that the presence of false positive prompts on images where targets were not marked by a CAD prompt led to a reduction in sensitivity (Ionescu et al., 2018). As with the studies with expert readers mentioned above, in the CAD condition of the studies by Drew et al. (2012) and Ionescu et al. (2018), prompts were displayed from image onset and not as a second reader. Therefore, this will have affected how participants search compared to how readers search typically using CAD in mammography as they do not get an initial unaided search without the presence of prompts.

3.7 Summary

CAD algorithms have seen a rapid increase in usage in the US over the past two decades after initial clinical studies boasted promising results, demonstrating improved cancer detection rates over single reading alone. This led to a rapid increase in the use of CAD systems across the US. Now, most mammography clinics in the US use CAD. However, this has not been adopted in Europe where double reading is standard practice, since there has not been sufficient evidence to suggest that single reading with CAD is equivalent in performance.

As CAD continued to grow in popularity across clinics in the US, studies began to highlight various issues. Studies were reporting increases in recall rates with little or no benefit in cancer detection rates, seemingly due to the high false prompt rate of the CAD systems. Readers were shown to dismiss correct CAD prompts in the majority of instances, and when prompts failed to mark a cancer, there was a significant tail-off in performance. There is a balance between reader-machine trust at play, which is a difficult relationship to maintain when the reader is unaware of why the CAD system has made certain decisions.

Therefore, providing readers with additional information can be beneficial to improve confidence in a CAD system. Interactive CAD systems withhold prompts until readers query their locations and may also provide lesion likelihood scores at the prompt locations. Indicating the probability of malignancy or providing similar cases with known outcomes are two examples of how readers can gain a better understanding of the mechanisms of the CAD they are using and in turn build better trust in the system.

Deep learning CAD methods are certainly going to be the future. However, they must be able to overcome the challenges of the interaction between reader and algorithm if they are to succeed where traditional CAD has fallen short. High accuracy with a reduction in the false positive prompt rate will be crucial, but it will not be enough by itself.

The next chapter will discuss in further detail the mechanisms behind the search task facing readers of mammograms. Chapter 5 will explore how eye tracking is used to study reader behaviour in mammography and with CAD.

Chapter 4

Visual Search

4.1 Introduction

Many tasks faced by readers of medical images require them to locate signs of abnormality. In these visual search tasks, the reader uses the ‘human search engine’ (Wolfe et al., 2015a). There are a number of different theories of human visual search, but before reviewing these it is first necessary to briefly discuss human vision and how the eye operates.

The main features of the eye are shown in Figure 4.1. Light incident on the eye is refracted by the cornea and passes through the pupil (the opening of the iris), and is focused by the lens onto the retina. The retina contains millions of light-sensitive receptor cells called cones and rods, approximately 6 million cones and 120 million rods. The distribution of rods and cones varies across the retina. At the fovea, there are no rods and a high concentration of cones, which are responsible for high spatial resolution, and therefore this region has the highest visual acuity (Atchison and Smith, 2000). Since only a small proportion of the overall visual field is focused onto the fovea, it must be moved around an image, such as a mammogram, to capture all of the fine details present. For expert readers viewing mammograms, Kundel and Nodine (2004) defined a useful field-of-view (equivalent to the FVF discussed previously) with a diameter of 5 degrees visual angle, centred around the gaze location.

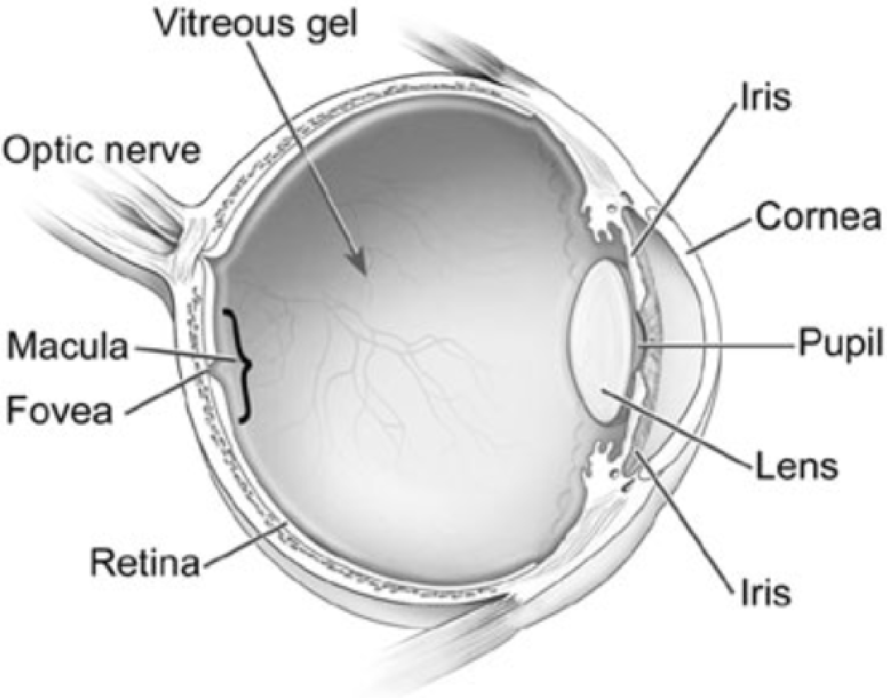


Figure 4.1: Diagram of the human eye, image from National Eye Institute (2019).

Fixed gaze on a single location for a period of time, usually defined as around 250ms (but can be shorter or longer than this), is known as a fixation. Between fixations, the eyes make rapid movements called saccades. The latency of a saccade, which is the time difference between the signal for a saccade to begin and the movement of the eyes, can be used to categorise them. Typical latency of saccades ranges between 150 and 500 ms (Klein and Ettinger, 2019). Saccades with shorter latencies (150 to 200 ms) are reactive saccades, where the eyes move in response to the appearance of a stimulus in the FOV. There are also express saccades with latencies around 80 to 120 ms

(Kingstone and Klein, 1993), which can be observed during tasks where participants must fixate a target away from an initial fixation point, if the fixation point is removed around 200 ms before the target appears. Those with longer latencies (>250 ms) are called voluntary saccades, which are purposeful movements to explore the visual environment. The amplitude of a saccade is often discussed in visual search literature, which is the difference between the start and end points of the saccade. The relationship between the amplitude and acceleration of saccades is described as the 'main sequence' (Gibaldi and Sabatini, 2020); for small saccades, velocity increases linearly with amplitude, whereas for large saccades the peak velocity is reached asymptotically.

Even when the eyes are fixating, they still make miniature movements of less than 1 degree visual angle in amplitude called microsaccades; the eyes are never truly stationary. Microsaccades are generally made involuntarily. In between microsaccades exist smooth eye movements around one tenth of the amplitude called drift. These are accompanied by another movement called tremor, a rapid movement that occurs approximately 100 times a second with amplitudes of around 0.001°. During a fixation, visual information is acquired, but at the same time, the next saccade is also being planned. This 'asynchronous parallel processing' improves the efficiency of search, since it means planning the next movement does not have to wait until all information has been processed (Hooge and Erkelens, 1996): as part of the planning process, it is believed that visual information at the location of the next fixation is processed before gaze has moved to that location (Godwin et al., 2021).

Fixations show where attention is focused and when visual information is being acquired, while saccades move this attention between locations. Measuring fixations and saccades made by observers while reading images gives a deeper understanding of how the images are being analysed, as it is possible to record search patterns as well as search performance.

4.2 Visual search theory

Visual search theories aim to explain the cognitive processes involved as we search environments for information. The leading theories are Feature Integration Theory (FIT), Guided Search (GS), and Attentional Engagement Theory (AET). These models are all ‘item-based’, where individual items are the central unit of search, but alternative ‘fixation-based’ models have also been proposed (Hulleman and Olivers, 2017).

Visual search is often studied with basic tasks where participants search for a target item amongst distractor items, with the total number of items present known as the ‘set size’ (Wolfe, 1998). Typically, a participant in these tasks responds to indicate whether they believe the target is present or absent in each trial. The time taken to respond is the reaction time (RT) and the percentage of correct decisions is the accuracy. Often the task will be analysed by plotting the RT versus the set size, called the ‘search slope’ (shown in Figure 4.2), which describes the change in RT as a function of set size. Search slopes are used as an indicator of search efficiency.

Theories of visual search may describe serial and parallel search, and feature and conjunction search. Serial search is where items are attended to one after another and parallel search involves the processing of multiple items at once (Palmer, 1995). In feature search the target is distinguished from the distractors by a single feature (colour, orientation, size, and motion). For example, search for a red circle among distractor green circles. Search where the target shares one or more features with the distractors, is known as conjunction search. An example of conjunction search would be searching for a red circle among red squares and green circles; the target cannot be defined by the colour or shape alone but rather a conjunction of both (red and circular).

Figure 4.2 shows the search slopes for various tasks. Feature search results in a flat search slope (efficient), since the target is easily distinguished from distractors, regardless of the number of distractors. Conjunction search results in steeper search slopes (quite efficient), where the common features between targets and distractors mean that search takes longer with more items present. Inefficient search slopes are produced with ‘configuration search’ tasks, where the targets and distractors are composed of different configurations of vertical and horizontal lines (e.g. digital 2 among digital 5s or T among Ls). Very inefficient search slopes, where the search display is processed in excess of 30 ms per item, may result from search involving the conjunction of two orientations. For example, search for a vertical rectangle containing a tilted bar among distractors of vertical rectangles with horizontal bars and horizontal rectangles with tilted bars (Bilsky and Wolfe, 1995).

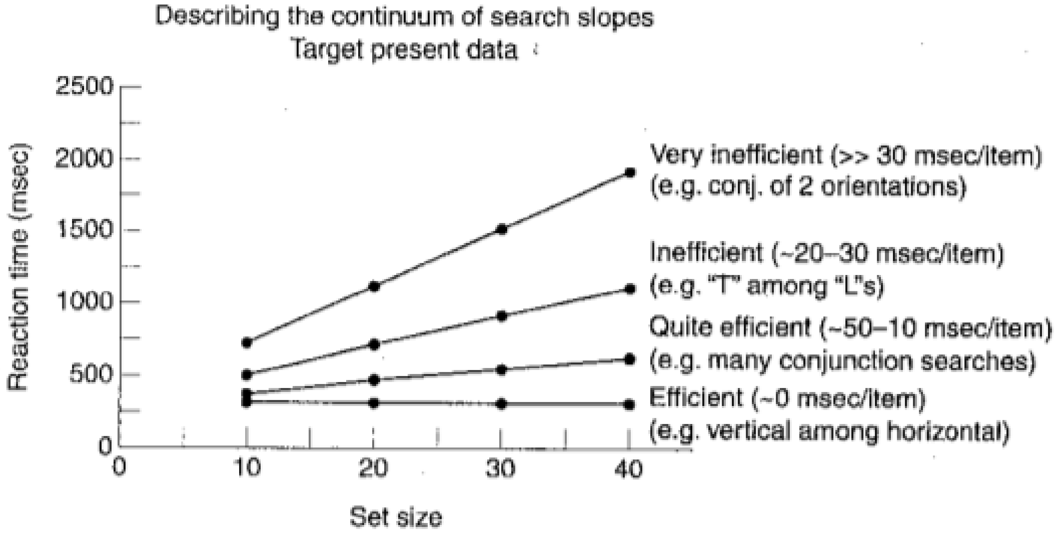


Figure 4.2: Search slopes of various efficiencies, from Wolfe (1998). A conjunction search involves detecting a target amongst distractors that share one or more feature with the target.

Search slopes are plotted for target present and target absent trials separately. Error rates are generally higher for target present trials than target absent. Rather than speed-accuracy trade-off, this reflects the fact that participants will end their search with a target-absent response when they did not find the target. But this is an incorrect response only for target present trials. Reaction times are slower for target absent trials. In principle, an observer must attend to all items before a target absent response can be given. For target present trials though, on average, only around half of the items need to be attended before a target present response (Wolfe, 1998).

The 'quitting threshold' determines when observers' terminate their search. When the target has been detected, this is an obvious point to stop searching. However, it is more complicated when the target has not been found and will depend on a variety of factors, such as the memory of previously visited items and the preceding trials. Observers do not have a perfect memory of the items that have already been visited and items are often re-examined, even with a set size as small as six (Young and Hulleman, 2013).

Feature Integration Theory, proposed by Treisman and Gelade (1980), explains the mechanisms of feature search and conjunction search. In FIT, features such as colour and orientation, are preattentively registered in separate 'feature maps'. Feature search is where targets differing from distractors by a single feature can be found by inspecting the relevant feature map. For targets that consist of a conjunction of features, attention must be applied serially so that feature maps can be bound together using a master map with the locations of the items. Feature searches are more efficient than conjunction searches since they are independent of the number of items in the display and the feature map can be monitored for activity signifying the presence of a target. Whereas in conjunction searches, inspecting the feature map does not help since, to decide whether an item is a target, attention must be applied serially to bind features of an item.

Wolfe et al. (1989) proposed Guided Search as an alternative to FIT. GS does not make the same distinction between parallel and serial search, and instead combines signals from feature maps into a single activation map. Attention is guided first towards the location with the highest activation, and then the next highest, and so on until the target is found or a target absent decision is made. There have been multiple updates to GS, and the latest version is GS 6.0 (Wolfe, 2021a). GS 2.0 assumes that search has perfect memory and that the search slope represents the time it takes to process an item (Wolfe, 1994). The assumption of perfect memory was shown to be false (Horowitz and Wolfe, 1998) and was removed in GS 4.0 (Wolfe, 2007), which now assumed that the search slope represents the selection rate of items. GS 4.0 also introduced the 'car wash model' for the processing of items: selected items enter processing one at a time, but multiple items are processed in parallel (since the time taken to select an item is less than the time taken to process it). In GS 5.0, Wolfe et al. (2015b) discussed that observers were capable of multi-tasking in their search for multiple targets, where they would keep track of target locations and mark a target at the same time as searching for further targets.

The most recent update, GS 6.0 (Wolfe, 2021a), describes how items are stored in a map that is constantly updated and a new item is selected every 50ms, with search being guided by items nearer to the point of fixation. Unlike previous GS models, 6.0 accounts for eye movements with the inclusion of a functional visual field (FVF) that plays a pivotal role in fixation-based search models such as Hulleman and Olivers (2017), discussed later in this section. Another key addition to GS 6.0 is the various forms of guidance, which is discussed below.

An overview of the GS 6.0 model is shown in Figure 4.3. With reference to Figure 4.3: (1) The visual system receives the image from the retina and (2) all information is processed through the non-selective pathway and information about the scene is extracted. (3) For the identification of individual objects, information is passed through the selective bottleneck. (4) and (5) The selective bottleneck is fed via five types of guidance. Bottom-up guidance describes search driven by image features such as salience, even without a target in mind readers may be drawn to particular regions (guided by the stimuli). With top-down guidance, attention is directed by the reader themself (determined by the properties of the target specific to the search task). Value guidance describes how features that are more valuable (e.g. participant is rewarded for finding these) will attract more visual attention. What was previously searched will influence current visual attention (history guidance). Finally, with scene guidance, attention is directed by the background as opposed to the stimuli. The various types of guidance form an attentional priority map that directs search through the selective bottleneck, resulting in a generally efficient search process. (6) The object that is currently being attended to is held in the Visual Working Memory (VWM), which also holds the top-down guiding template.

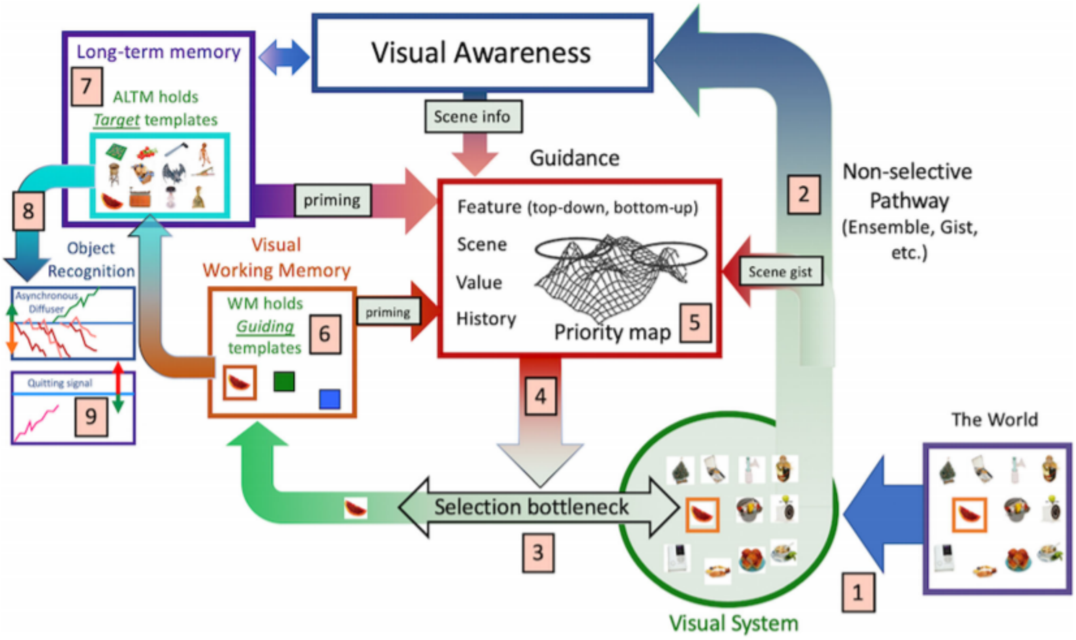


Figure 4.3: Schematic representation of Guided Search 6.0, from Wolfe (2021a).

(7) The target template is held in long-term memory and objects in the VWM are compared to identify targets and discard distractors. (8) Search items are passed into a parallel processing stage around every 50ms, where multiple items are processed at once (over several hundred milliseconds) to either recognise it as a target or dismiss it – referred to as the 'asynchronous diffuser'. (9) A quitting threshold determines when the search is terminated, with the threshold lowering after true-negative responses since the scene is now easier to search with a distractor dismissed. On the other hand, the threshold is raised when a target is missed.

Attentional Engagement Theory is another alternative theory to FIT, proposed by Duncan and Humphreys (1989, 1992). FIT asserts that attention must be applied serially to bind features together that are present at a given location during a conjunctive search. This is not the case with AET, there is no serial stage of search and there is no qualitative difference between the processes behind parallel and serial search. In AET, stimuli compete to enter the VWM, weighted by their target similarity. Search efficiency decreases when the similarity between targets and distractors increases, and when dissimilarity of distractors increases. Müller et al. (1994) present a model called parallel SEarch via Recursive Rejection (SERR), a computational implementation of the fundamental ideas of AET. Templates units are used to group targets and distractors by their similarity. During search, templates will 'fire' – if a distractor template fires then the locations of distractors that fit this template are inhibited, while a target template firing will lead to a target present response. Distractors are recursively rejected until there are no items left or the target is detected. GS 2.0 (Wolfe, 1994) included how search efficiency varies with target-distractor similarities in reaction to AET (Chan and Hayward, 2013).

These item-based models make two implicit assumptions: shallow search slopes (reaction time versus set size) are the most informative about the cognitive processes of visual search, and eye movements are not an important factor in the search process. It is argued by Hulleman and Olivers (2017) that these assumptions have prevented all aspects of search being explained by current models, whereas a fixation-based model is not restricted by these assumptions. Hulleman and Olivers (2017) presented a simulation of a fixation-based model, with four fixed parameters and one free, based on the results of previous experiments. The main assumption in this model is that there is a functional visual field (FVF), defined as the area of the visual field around a fixation from which a target can be expected to be detected. The processing may not necessarily be complete before attention is diverted elsewhere.

The size of the FVF depends on the task difficulty, as shown in Figure 4.4. It is larger for easy tasks and smaller for hard tasks. Items within the FVF are processed in parallel and fixations are set to a constant duration (250ms). Since memory of previously visited locations in visual search is limited, avoidance of re-fixation on previously visited locations was limited. Finally, once 85% of items had been visited, the search was terminated. When compared to experimental data for three tasks of varying difficulty, the model was shown to correspond well for both reaction times and number of fixations as functions of the display size. The variability in reaction time and number of fixations as a function of display size was also well matched by the simulations.

Young and Hulleman (2013) and Hulleman and Olivers (2017) provided experimental evidence for the need for a fixation-based theory by investigating how different difficulty search tasks affected observer performance. The results of Hulleman et al. (2019) form another piece of evidence that there is a clear difference between medium difficulty search and hard search. In this paper, the targets and distractors were formed

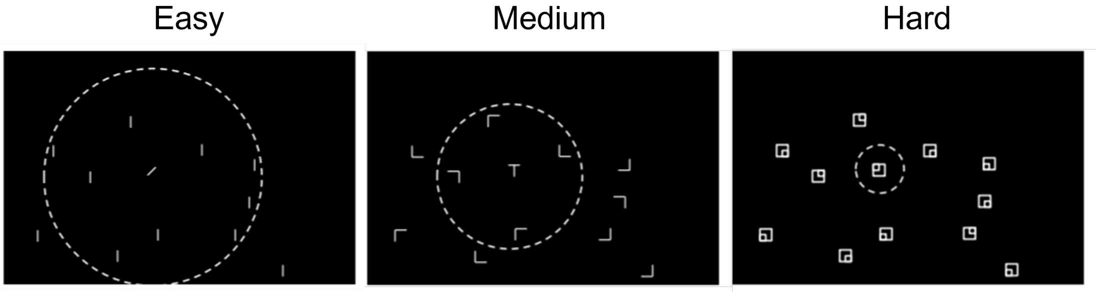


Figure 4.4: Estimated size of the FVF (dotted circle) for tasks of different difficulties. The easy task (left) is search for a diagonal amongst vertical lines, the medium task (middle) is search for a T amongst Ls, and the difficult task (right) is search for square with a small square in the top left corner amongst squares with a small square in a different corner. Image adapted from Hulleman and Olivers (2017).

from sets of lines. The tasks were completed with half of the search items created from the same line set as the target (50% eligibility) and with all items from the same set (100% eligibility). The difference in eligibility was achieved by rotating distractors relative to targets. This design was used since item-based models predict that performance will improve for 50% eligibility. When half of the items are rotated, FIT predicts that the orientation map would suppress distractors in the master map resulting in a smaller set size. In GS, the bottom-up and top-down guidance will be dictated by orientation, boosting the activation of non-rotated items. Finally, in AET, a group of distractors would be coded by their orientation and would no longer compete to enter the VWM since they do not match the target template, and therefore the target can be detected more easily.

Hulleman et al. (2019) demonstrated that for a medium difficulty search task, search performance was reduced for 50% eligibility compared to 100% eligibility. However, in difficult search tasks, search performance was improved in the 50% condition. This suggests that for difficult search tasks, item-based models are compatible with this result, and suggest that single items are processed per fixation. For medium difficulty tasks, the size of the FVF is increased, and multiple items are processed per fixation – in contrast with item-based models where similar results would be obtained for tasks of different difficulties. This work demonstrates the importance of a search model to account for more than just the target template, since the surrounding items also have an impact on detectability, as demonstrated with the medium difficulty search task. These results do not directly follow from a fixation-based theory, but unlike an item-based theory they can be understood within a fixation-based theory.

In GS 6.0, while Wolfe (2021a) disagrees with the fundamental idea of a fixation-based search model, the importance of the FVF is acknowledged. The FVF loads multiple items into the asynchronous diffuser (stage 8 in Figure 4.3), but the main search process is governed by individual item processing. However, the definition of the FVF differs between GS 6.0 and that described by Hulleman and Olivers (2017). In GS 6.0, attention is restricted to the area around the fixation and cannot roam the entire search display, and does not provide an explanation for why the FVF reduces in size when the search difficulty increases (Hulleman et al., 2019).

These cognitive models are typically based on laboratory studies using search tasks such as finding Ts amongst Ls, rather than complex medical images, since they are concerned with the fundamental mechanisms of visual search. The search of images used in visual search literature can differ greatly from medical images. Visual search studies often use a single target and generally do not involve a diagnosis stage. Yet, there are visual search studies that have modelled the conditions in complex search scenarios such as airport baggage screening and radiology. In particular, multiple target search has been investigated and the impact of low target prevalence was highlighted.

Hybrid search (Wolfe, 2012) and mixed hybrid search (Wolfe et al., 2017a) describe the search of multiple targets, where targets are either specific items or categories of items (discussed further in Section 4.3). In radiology, readers often search through multiple image views for various types of abnormalities. This sort of complex search has been modelled through hybrid foraging search (Wolfe et al., 2016). Here, there are multiple targets to search for and the number of each of these targets is not known. In a study in which observers completed a hybrid foraging task, readers searched the image in ‘patches’ (an example of one of these patches is shown in Figure 4.5), moving between each patch to review items and verify if they were one of the desired target types via a memory search (Wolfe et al., 2016). In doing so, around 25 to 33% of targets remained undetected when moving to the next patch. The readers also had a tendency to select a single target type and go on ‘runs’ of detecting those targets, instead of the other types that were also available. These single-type runs were found to be more efficient in terms of reaction times rather than searching for multiple types at once, since readers memories were reconfigured in such a way that the classification of that target type would be faster.

It has also been investigated whether target prevalence has an impact in hybrid search (Wolfe et al., 2017b). The effect of target prevalence was initially reported in visual search literature by Wolfe et al. (2005). In this study, observers searched for semi-transparent objects in noisy backgrounds at a target prevalence of 1%, 10%, and 50%. It was found that the miss error at 1% prevalence was around double the error at 10% prevalence and four times higher than at 50% prevalence. Measuring the effect of prevalence is key for radiology, where the rate of targets is often well below 1%. In a hybrid foraging study, observers searched for four target types that had varying prevalence rates of between 7% and 53% (Wolfe et al., 2017b). Targets that were more common were detected at a significantly higher rate than less common targets. Therefore, prevalence has a clear impact for both single and multiple target search.

These hybrid foraging tasks are more similar to real world search than those typically used in basic visual attention research. The fact that there are parallels between

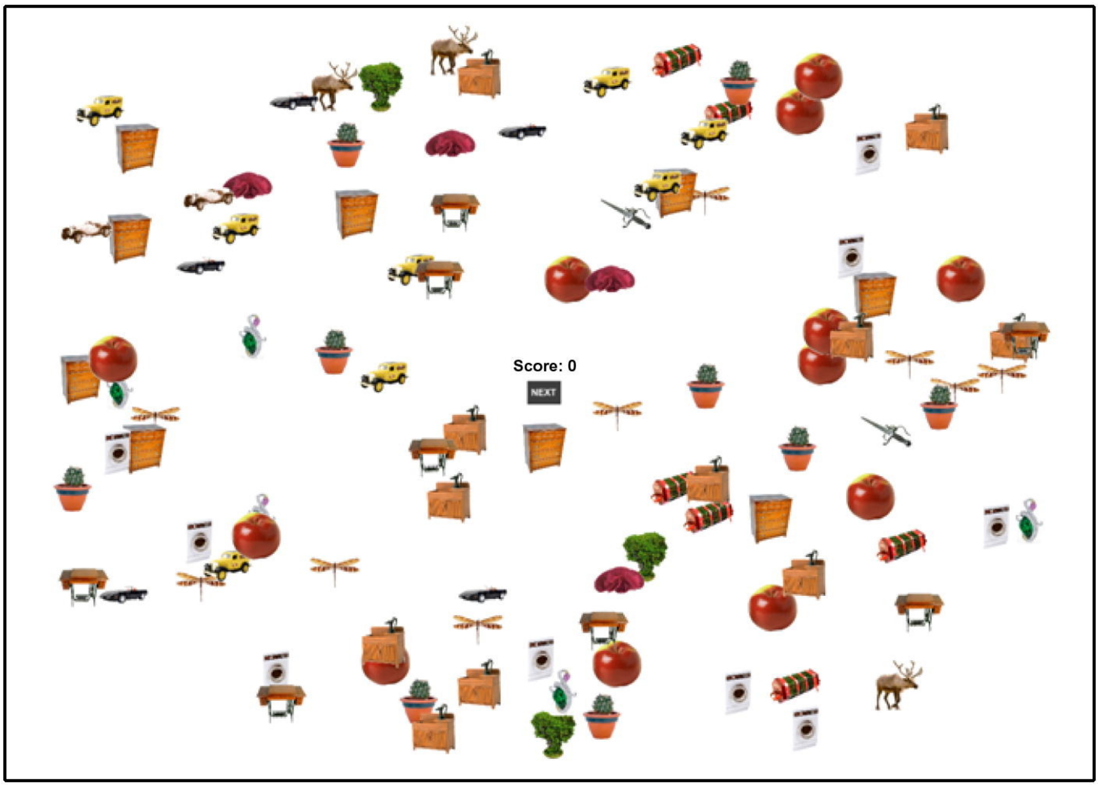


Figure 4.5: An example of a patch used in the hybrid foraging task in Wolfe et al. (2016). The items would move around the display. Observers were required to hold 8, 16, 32, or 64 possible targets in memory. The set size ranged between 60 and 105 across patches, with a target prevalence of 20–30%. There were multiple of each target type present across patches.

hybrid foraging and basic search suggests that the fundamental processes described in visual search theories are likely to be applicable to real world search. By better replicating real world search such as medical image interpretation, a deeper understanding of the processes can be gained, and importantly methods to reduce errors and improve performance can be investigated.

Before hybrid (and hybrid foraging) search entered the visual search literature, different approaches were developed to describe search in medical settings by monitoring the search behaviour of medical experts. Outside of the basic visual attention research, Nodine and Kundel (1987) outlined radiologists' visual search of medical images with five stages. This initial work was based on the search of chest x-rays for lung nodules. A later study evaluated this approach for mammography and altered the order of search stages (Kundel and Nodine, 2004). The first stage is *global impression*: the image is analysed globally, lasting a few hundred milliseconds. This initial global impression has been shown to be enough for readers to determine whether a mammogram contains a cancer or not, with an above chance level of accuracy (AUC = 0.75), when shown the image for just 500ms (Evans et al., 2016). The next stage is *foveal verification*: areas of the image that attracted attention in the global analysis are inspected (foveal vision) and located, which takes a few seconds. Then *discovery search*: given a specific task, the observer begins a cognitively guided search, examining areas with a high probability of finding the search target of that task. This is followed by *reflective search*: the whole image is re-examined to confirm initial suspicions and to aid in decision making on any abnormalities. Finally, *post-search recall*: the observer reports on findings.

4.3 Factors affecting visual search

A variety of different factors can be detrimental to visual search, some of which were discussed in Section 2.4 in the context of mammographic interpretation. These factors can lead to a range of reader errors. The most common of which are perceptual errors, where retrospectively apparent findings are simply missed, accounting for an estimated 60% to 80% of total errors (Bruno, 2017). As previously stated, perceptual errors can be classified as search, recognition or decision errors, depending on how long (if at all) targets were fixated. Examples of search and decision errors are shown in Figure 4.6. The cause of these errors may be that targets are subtle or masked by normal tissue, or due to an ineffective visual search (Krupinski, 2010). Kim and Mansfield (2014) defined 12 types of diagnostic errors in radiology, and in their study found that underreading, where otherwise obvious findings are missed, accounted for 42% of errors.

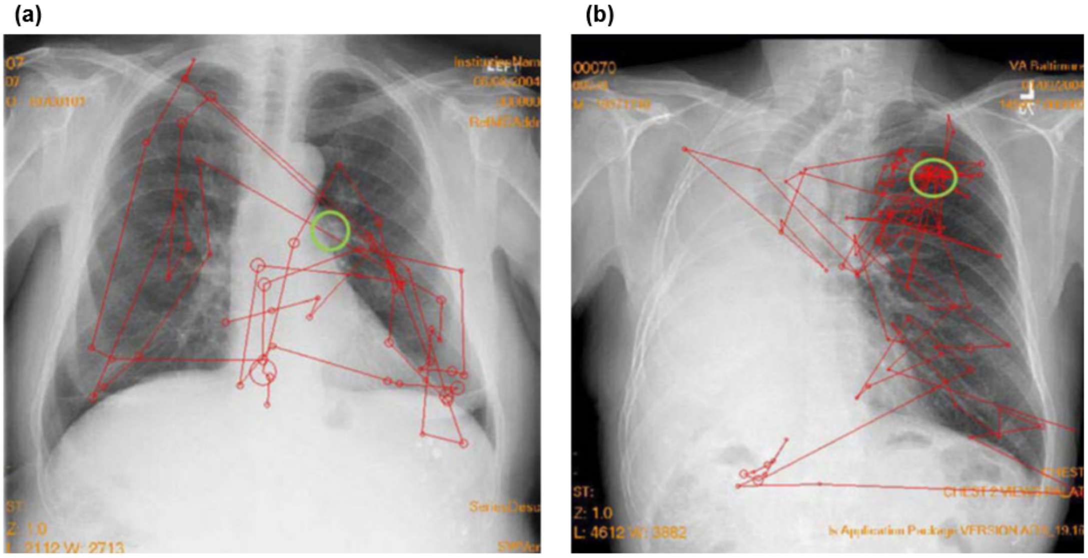


Figure 4.6: Example of perceptual errors in radiology, from Krupinski (2010). The red circles are fixations and lines between them are saccades, while the green circle marks the position of the tumour. (a) Search error: the reader does not fixate the tumour and does not report it. (b) Decision error: the reader fixates the tumour multiple times but does not report it.

The prevalence effect is a well known effect in visual search literature (Wolfe et al., 2005; Horowitz, 2017), where search performance decreases as targets become rarer. Evans et al. (2013) compared the performance of expert readers at different prevalence conditions (see Figure 4.7). A total of 100 cases were inserted into the normal workflow of 14 readers in a hospital clinic over a 9-month period resulting in a prevalence of around 1%. Six of those readers later interpreted the same 100 cases at 50% prevalence in a laboratory setting. There was a significant increase of 18% in the number of false negatives when moving from the high to low prevalence condition. Therefore, low prevalence is likely to account for a portion of the missed cancers in screening mammography.

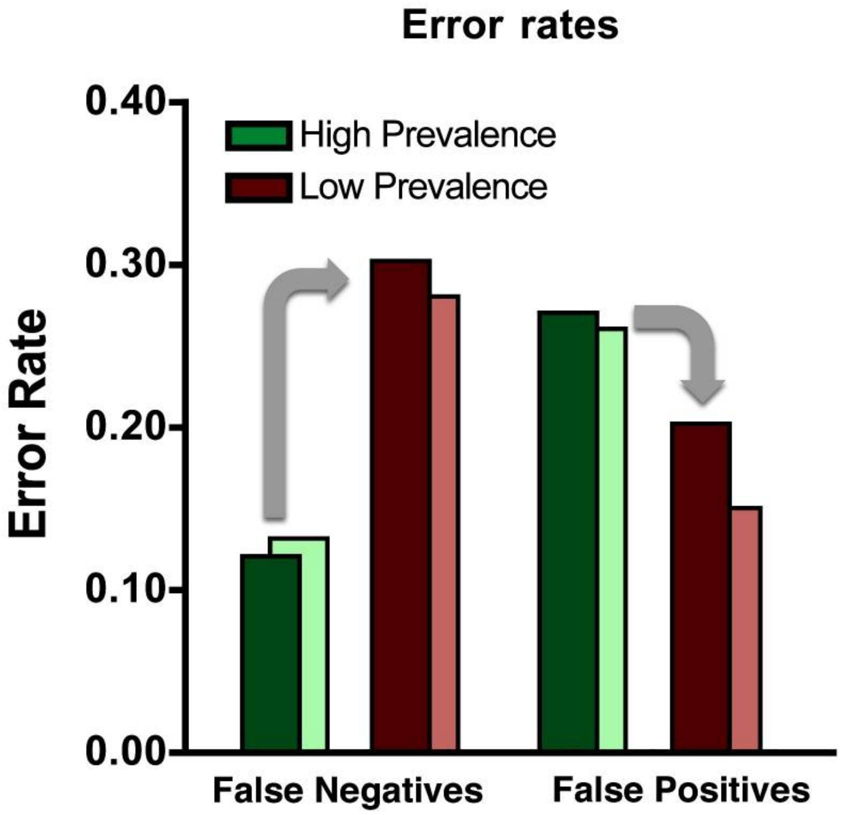


Figure 4.7: Results from Evans et al. (2013) demonstrating the low prevalence effect. Green bars represent the 50% prevalence condition and red bars the 1% condition. The different shades of each colour are a result of filtering out readers to account for case familiarity, since some cases were used in both the high and low prevalence sets. The dark bars represent the data for all readers (N=14) and the light bars are the average rates for the 6 readers who completed both conditions (light red) or when cases used in both prevalence conditions were removed (light green).

The effect of prevalence on the ability of 14 expert readers to detect abnormalities in chest x-rays was investigated by Gur et al. (2003). A total of 1,632 cases were used in this study. Five prevalence levels were used, ranging from 2% to 28% across three main abnormality types. Readers were required to give a rating between 1 and 100 to indicate their confidence that an abnormality was present in an image. In contrast to the results found in mammography, they found that there was no effect of prevalence on the AUC values of readers or case review time. In a follow-up study, the data was reanalysed to examine the effect of prevalence on reader confidence ratings (Gur et al., 2007). They observed a significant trend in the confidence rating as a function of prevalence, with confidence ratings increasing as the prevalence decreased. Since the study was completed in a laboratory setting, it is unclear on whether they would generalise to a clinical setting.

Instead of changing the true target prevalence, Reed et al. (2014) examined how changing the readers' expectation of prevalence can affect their behaviour when reviewing chest x-rays. Readers were split into three groups, with each group reading a set of 30 cases (containing 15 abnormalities) twice over two sessions, separated by a few days. Each group was told a different prevalence at the beginning of each session: 9, 22, or not told, in one session versus 15 (true prevalence) in the other. Across all groups, higher prevalence expectation on normal images led to a significant increase in duration of image scrutiny and number of fixations. For normal images, the confidence ratings of one group (9 versus 15) was significantly increased at higher prevalence expectation. This is the opposite trend to that observed by Gur et al. (2007), however, it compared the expectation of prevalence as opposed to changing the true prevalence. Therefore, the expectation of prevalence caused readers to inspect normal images more thoroughly and increased their confidence that the image was normal, despite the fact actual prevalence remained unchanged. However, despite higher confidence and longer interpretation times at higher prevalence expectation, no significant differences in sensitivity, specificity, false positives, or false negatives were observed. In mammography, prevalence is much lower than the 9% used for the low prevalence condition in this study and therefore there may have been an impact on performance too had a lower target rate been used. As demonstrated by Wolfe et al. (2005), the performance is reduced drastically at 1% prevalence, with errors almost doubling compared to 10% prevalence (error rate of 30% versus 16%).

Studies have also investigated how prevalence can interact with CAD prompts. Russell and Kunar (2012) conducted a series of experiments examining the viewing behaviour of non-expert participants searching for Ts amongst Ls in a high (50%) an low (2%) prevalence condition. The first experiment was conducted without CAD and reported a low prevalence effect where miss error was higher at low prevalence. In another experiment with 9 participants aided by CAD prompts, there were 240 trials in the high prevalence condition and 5,000 trials in the low prevalence condition. Miss errors were higher in the low prevalence condition, for all target present images (prompt marking target, prompt marking non-target region, and no prompt present), with the fewest errors made when prompts marked the target. Errors were also reduced when allowing participants to self-correct their responses, i.e., go back and change their response if they realise they made a mistake after they had moved onto the next image. Prompts reduced the low prevalence effect, with participants more willing to mark targets that were prompted compared to the no prompt condition, however it was not completely eliminated.

A later study by the same group with a similar methodology investigated the low prevalence effect in the search of mammograms with CAD, using non-expert readers trained to read mammograms (Kunar et al., 2017). In one experiment, a single mass target was used, and in another a range of masses were used. Both studies had 24 participants, with 80 mammograms in the high (50%) prevalence condition and 1,000 in the low (2%) prevalence condition. With a single target to search for, the low prevalence effect was observed, consistent with Russell and Kunar (2012). However, the low prevalence effect was not observed when searching for a range of masses. This was due to a higher than anticipated error rate in the high prevalence condition, rather than a reduction in miss errors in the low prevalence condition. The CAD prompts caused participants to miss a higher number of masses when the prompts were on non-target regions, and believe that masses outside of prompts were more likely to be benign. These results suggest that participants had become overreliant on CAD. There was a higher miss error and false alarm rate when a range of targets was used, which was hypothesised to have occurred due to a weakening of target representations in the Visual Working Memory (discussed below). This may have in turn caused participants to rely even more on the prompts.

Visual search tasks that involve more than one target require multiple target representations to be held in the Visual Working Memory (Oberauer, 2002). VWM describes the capacity of visual information to be stored at any one time and accessed for search tasks. These target representations must be held in focus while searching, with the limit on the number of targets that one can retain within their VWM was argued to range from just one (Oberauer, 2002) to four (Cowan, 2001). However, Wolfe (2012) demonstrated that while search efficiency decreased for the number of targets memorised, participants were able to search for as many as 100 targets with a miss error of 2% and 19% for set sizes of 1 and 16, respectively, suggesting it is possible to retain many target representations at once. This may not be the case for complex targets (as is the case in mammography), where there is a trade off between the complexity of the targets and the number of items that can be stored (Alvarez and Cavanagh, 2004;

Eng et al., 2005). Studies have demonstrated that visual search tasks involving two targets incur a dual-target cost, resulting in a less accurate search when participants were tasked with finding two targets compared with just one (Menneer et al., 2007; Mestry et al., 2017). This reduction in accuracy was argued to be due to a weakening of the target representations in the VWM for the non-preferred target (Mestry et al., 2017).

When observers are engaged in a specific task, they may miss important (sometimes obvious) but unexpected information – a phenomenon known as 'inattentional blindness'. In one study with readers searching for lung nodules in chest CTs, 83% of observers failed to see an image of a gorilla inserted into the CT (Drew et al., 2013), shown in Figure 4.8a. However, the gorilla had a lower brightness than the nodules and was only present in 5 slices out of the total 239, but does demonstrate that even expert observers are not exempt from this limitation of human visual search. Expert medical readers may argue that a small gorilla is not something that would be found in a chest CT and therefore means it would not be regularly detected. However, it has also been found that readers tasked with detecting lung cancer missed signs of breast cancer and enlarged lymph nodes, at a rate of 66% and 30%, respectively (Williams et al., 2020), shown in Figure 4.8b. When those same readers were instead directed to search for a range of abnormalities, these miss rates were reduced to 3% and 10%, respectively. Although, this instruction is not particularly relevant in mammographic screening where the task is already pre-defined to search for a range of abnormalities.

An analogue of this problem was presented by Wolfe et al. (2017a), described as 'mixed hybrid search'. Hybrid search involves searching displays for multiple targets, and mixed hybrid search involves searching for both specific targets (e.g. this hat or this dog) and categorical targets (e.g. clothes or animals). These categorical targets are analogous of the unexpected (incidental) findings in medical imaging. In experiments with non-expert readers, the miss error for categorical targets was higher than for specific targets; when both target types could appear on any trial, the miss rate was 3.6% for specific targets and 36.6% for categorical.

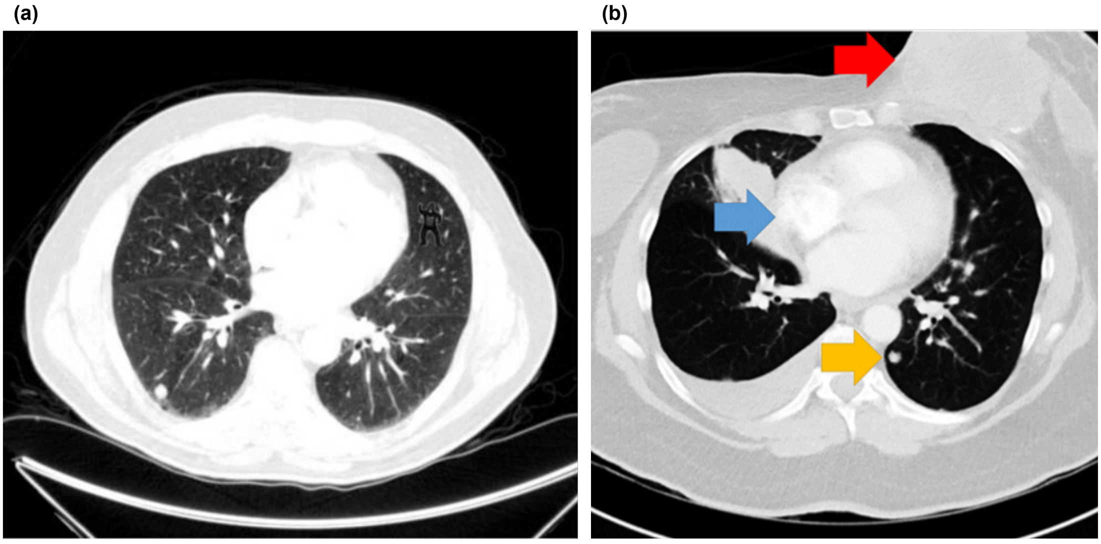


Figure 4.8: Images used to test inattentional blindness of expert readers. (a) chest CT containing a gorilla from Drew et al. (2013) and (b) chest CT containing the target lung nodule (yellow arrow) and unexpected abnormalities of a breast mass (red arrow) and enlarged lymph nodes (blue arrow) from Williams et al. (2020).

Nartker et al. (2020) investigated methods to reduce the miss error rate of categorical targets in mixed hybrid search. There were three methods used: non-search trials containing outlined categorical targets were inserted between trials to remind the observers of their presence; asking the observers to respond separately regarding the presence of specific and categorical targets; and finally requiring participants to record their responses on a six-response checklist to indicate the presence or absence of each target (3 specific and 3 categorical targets). Only the last method led to a reduction in the error rate for categorical targets (18% for last method versus >30% for first two methods). The use of a checklist resulted in a significant increase in the average trial time compared with the other methods, which would be problematic in a clinical environment. Furthermore, there are many possible incidental findings in imaging modalities such as lung CT, thus requiring readers to mark a checklist for the presence/absence of all possible findings is not feasible.

In mammographic screening and other imaging modalities, the whole image must be searched adequately to find all signs of cancer. However, there is also pressure on readers to get through a large number of cases, and so the reader must decide when to terminate the search and make a decision on the case. If a reader has already detected a sign of cancer, they will be less likely to detect an additional sign (if present) - an effect known as 'satisfaction of search' (Berbaum et al., 1990). Krupinski (2010) estimated that between one fifth and one third of errors can be attributed to satisfaction of search. Although a more recent study, with 20 readers interpreting 64 chest CTs with and without the addition of simulated nodules, found no evidence of this effect (Berbaum et al., 2015). While accuracy was unchanged, readers reported fewer abnormalities in the condition with simulated nodules, as measured by a significant reduction in the centre of false positive range of ROC operating points.

It has been suggested that satisfaction of search may be due to a 'suppression of recognition' (Kundel and Nodine, 2010). That is to say, when a reader fixates on an abnormality that is dissimilar to an abnormality that has already been detected, the reader is less likely to perceive it as abnormal. Preliminary analysis of this concept by Mello-Thoms et al. (2014) found evidence to support it; previously made decisions biased readers decisions on abnormalities they fixated but did not report, suggesting they were too dissimilar in appearance to the reported abnormality to also be accepted as abnormal. However, this study did not follow Berbaum's protocol, in which satisfaction of search cannot be credited as the reason for any abnormality being missed if it cannot also be demonstrated that it would not have been missed without a distractor present (Berbaum et al., 1990). Therefore, further research is needed to confirm the idea of recognition suppression.

In addition to satisfaction of search, where search decisions are influenced by previous fixations and decisions within the same case, reader behaviour can also be biased by previous cases or information provided before they review a case. Prior knowledge of the locations of potential nodules in chest radiographs (reported previously by at least one other reader) disrupted the usual visual search of radiologists and led to a reduction in accuracy compared to an independent review of the images (Swensson and Theodore, 1990). When readers of chest scans were told that the set of images originated from patients who developed a lung tumour 6 months after the image was taken and provided the location of that tumour (for normal images, fictitious locations were used), there was a significant improvement in nodule detection compared to an initial read without this information (Littlefair et al., 2016). However, this information also led readers to over-read the cases and specificity decreased on the potential tumour locations. The increase in false positives is likely due to the increased expectation to find an abnormality at those locations, even when there was not a nodule to be found.

The effect of prior knowledge has also been explored in mammography. A study by Elmore (1997) analysed the differences in performance when readers were provided with clinical history for each case versus when no information was provided. The overall diagnostic accuracy was 67% without patient history and 72% with, but this difference was not significant. This may have been due to low statistical power with 10 readers reviewing just 50 cases. However, clinical history did influence the recommendations made by readers. For example, providing details of the patient's family history of breast cancer caused more readers to recommend a follow-up. Reviewing cases with clinical history available is standard practice in the majority of radiology, including mammography.

A review article of sixteen studies of various imaging modalities, only one of which used mammograms (Elmore, 1997), concluded that clinical history should be provided (Loy and Irwig, 2004). In the majority of these studies, reading with clinical information leads to a significant improvement in accuracy and in no instances was there a significant reduction in accuracy. They also suggest that there should be efforts made to improve how this information is provided to readers (Loy and Irwig, 2004), such as presenting the information to readers after an initial review (Littlefair et al., 2016) – in a similar way as second reader CAD. Not only does the information provided directly (or made available) to readers affect behaviour, but it has also been demonstrated that judgements on the current case are influenced by the judgements and features of previous cases in the sequence of cases examined (Alamudun et al., 2018). By monitoring eye movements during review of mammograms, the visual search behaviour and decisions of readers on previous cases were shown to be significant predictors of behaviour on the current case being reviewed.

4.4 Summary

Visual search of an image or environment comprises of a series of fixations, where the eyes are relatively still and visual information is gathered, and rapid movements called saccades that move fixation to a new location. The underlying processes are complex and are aimed to be described by visual search models.

Models are either item-based and fixation-based, depending on the central unit of search. The most established item-based model is Guided Search. The general principle of this model is that the visual system takes in the image, the whole image is processed to extract scene information which contributes to an attentional priority map that directs search for individual items. Target templates held in long-term memory are compared with the item that is currently being fixated, and it is either accepted or discarded.

Fixation-based models challenge the idea that search is dictated by the processing of individual items, and is instead governed by fixations. The key idea is that there is a functional visual field around fixations, in which parallel processing of items within that region occurs. The size of the FVF varies according to the difficulty of the task, decreasing in size as task difficulty increases. The processing that occurs within the FVF describes the statistical probability that a target signal is present, and attention may be diverted before this process is complete.

The search of mammograms has also been described. Search is thought to begin with a global impression, where the whole image is rapidly analysed and suspicious areas of the image are flagged up, which are then attended to individually over a longer period. The image will then be searched in a process guided by the task, attending to areas of highest suspicion. Finally, the whole image is examined again to confirm any initial suspicions before decisions are reported.

The search of medical images can be affected by the conditions of the task or the information available, and may increase the number of perceptual errors made. The low prevalence effect has been shown to impact expert readers interpreting mammograms, where miss errors increase when the prevalence of cancer decreases. Previous work has demonstrated that CAD prompts that fail to mark targets increase the miss error compared to unaided viewing and this effect was shown to worsen at low prevalence. Readers engaged in a specific task are also more likely to miss incidental findings, and when several of the same targets are present in an image, they are less likely to find further targets once the first has been detected. In mammography, the availability of clinical history for a given case was shown to not impact overall accuracy but did increase recall rate where history suggested an increased cancer risk.

In order to study visual search in detail, eye tracking is often used. Chapter 5 discusses the methodology behind eye tracking techniques and various studies using it in mammography, with CAD, and elsewhere.

Chapter 5

Eye Tracking

5.1 Introduction

Eye tracking aims to calculate the location of a participant's gaze, either by measuring eye movements relative to the head, or by measuring the orientation of the eye in space. Eye tracking has a wide range of research and commercial applications, including cognitive psychology, marketing, defense, medicine, sport, and computer vision. There has been a wide variety of eye tracking studies in radiology (van der Gijp et al., 2016) and in breast imaging specifically (Gandomkar and Mello-Thoms, 2019).

The first measurements of eye movements were obtained in 1879 by observing a participant with the naked eye as they read text, where the experimenter(s) counted the number of saccades per line and listened to the sound of the contracting muscles in the eye using a rubber tube placed on the eyelid (Wade, 2010). The first measurements with experimental apparatus were highly intrusive, involving a cup pressed onto the eye, with the lids held open, and the head heavily constricted (Huey, 1898; Delabarre, 1898).

A technique that was developed in the 1930s, and is still in use, is electro-oculography (EOG), whereby electrodes are affixed around the eyes to measure the voltage caused by rotations of the eye (Mowrer et al., 1935). The negatively charged retina and the positively charged cornea form a dipole, and when the direction of gaze is changed, the direction of the dipole also changes, which can be measured by the electrodes around the eye. Typically, the spatial resolution of this technique is >1 degree visual angle. Another, more intrusive method, where thin copper wires wrapped into a coil and embedded in a silicone annulus (called scleral search coils) are applied to the eye, with a local anaesthetic applied to the eye (Robinson, 1963). When the participant is sat within large coils generating a magnetic field, eye movements will induce a voltage on the coil which can be measured. With two magnetic field coils, horizontal and vertical eye movements can be measured. This technique has a high spatial resolution (<0.5 degrees visual angle) but is not as common now due to its invasive nature.

Video based eye trackers are advantageous in that they do not require the application of anything to the participants face or eyes, although some may still require them to be restricted in some capacity or actually wear the system if it is head-mounted (e.g. glasses). They work on the principle of using reflections of infrared (IR) light from the surface of the cornea, the relative position to the pupil is recorded by a camera, and this is used to determine gaze position. Spatial resolutions are typically <1 degrees visual angle. This will be discussed in further detail in the following section.

5.2 Eye tracking methodology

The most common method for eye tracking is corneal reflection (CR) with video based systems. An IR light source is directed towards the eye, which produces a reflection on the cornea (see Figure 5.1). A camera records the eye and measures the relative position of the reflection and the centre of the pupil. Image processing algorithms are used to find the locations of the reflection and pupil centre.

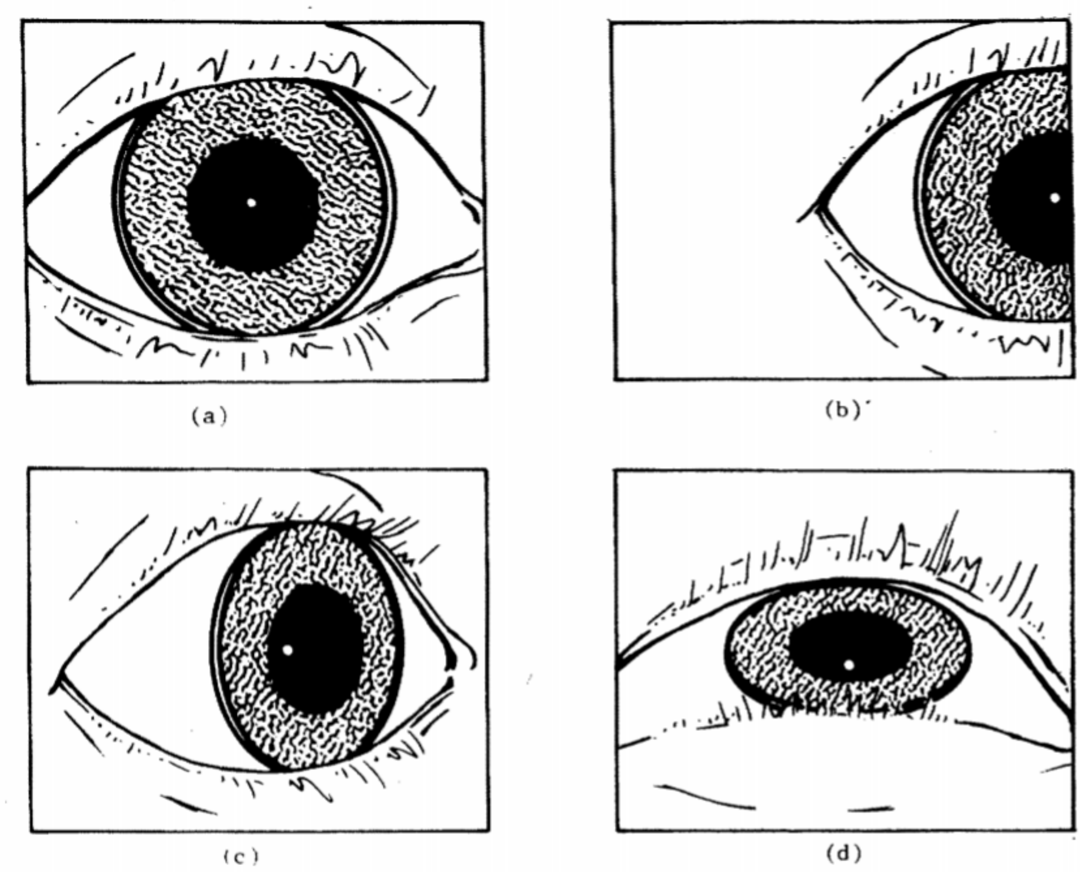


Figure 5.1: Corneal reflections (white dot on the pupil) for various eye positions. Reflections are shown for (a) eye looking straight ahead, (b) eye looking straight ahead but translated horizontally, (c) eye looking to the side, and (d) eye looking upwards. As shown for each of the four diagrams, depending on the rotation or translation of the eye, the corneal reflection will appear in a different position relative to the pupil. Image from Merchant et al. (1974).

Two techniques for CR exist: dark pupil and bright pupil, which depend on the position of the camera relative to the IR source (see Figure 5.2). When the IR source and the camera are coaxial, the light reflects off the retina and creates a bright pupil (similar to red eye) with a high contrast between the pupil and the iris. When the IR source is offset from the optical path then the pupil appears dark. The relative position of the CR and pupil centre can be used to determine the gaze direction, since these change with pure eye rotation, but are relatively independent of head movements (Duchowski, 2017).

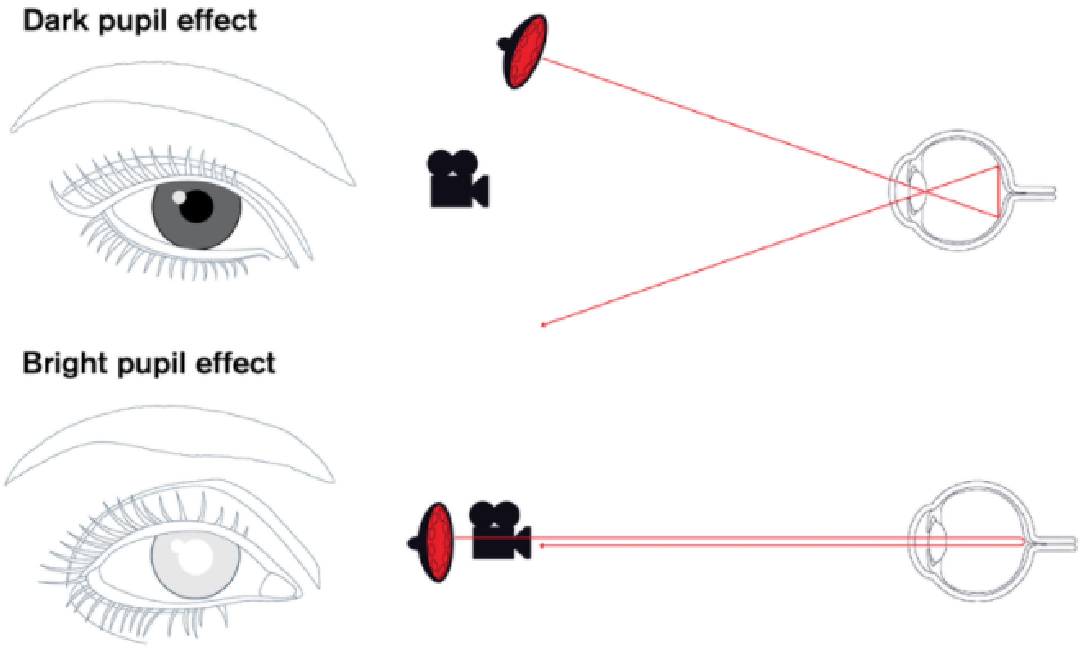


Figure 5.2: Schematic of eye tracking setup for bright and dark pupil corneal reflection. Image from Tobii Pro (2017).

With reference to Figure 5.3, the separation between the pupil centre and CR, d, is given by $d = K\sin\theta$, where θ is the angle between the optic axis (line of gaze) and the light source, and K is the distance between the centre of corneal curvature and the pupil centre. To map the gaze direction onto a screen, the observer screen distance must also be known. Furthermore, the tracker requires a number of ground truth measures of the line of gaze, acquired through a calibration procedure. Using these reference points, the images of the eye are then linked by an algorithm to locations on the screen.

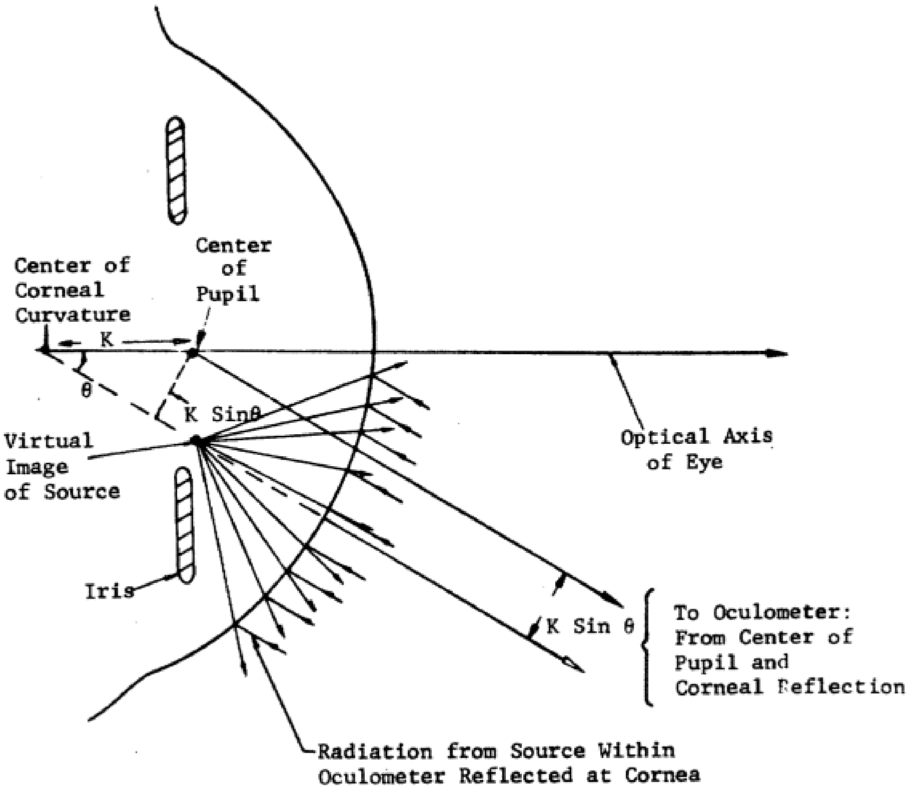


Figure 5.3: Diagram for the calculation of gaze direction. Image from Merchant et al. (1974).

There are two types of eye trackers: remote and head-mounted. Remote trackers are desk-mounted and record the eye from a distance, and head movements are mostly restricted. Head-mounted systems use cameras and illuminators close to the eye to track gaze direction, and co-register this data to video from a forward facing scene camera to show where the reader is looking. Although, with EyeLink II, the scene camera is optional and can connect directly to the computer for on-screen tracking (SR Research, 2021). Remote trackers are typically used for screen-based experiments and tend to have a higher spatial resolution compared to head-mounted trackers, whereas head-mounted trackers are commonly used for studies in natural environments. However, because remote systems often restrict head movements and have a narrower field-of-view than head-mounted trackers, head-mounted systems are often used instead for large displays like those used in radiology.

While head-mounted systems use IR illuminators for pupil tracking, not all require the use of CRs for gaze estimation. Since the head is fixed relative to the eye camera and IR source, the CR is not necessary (Klein and Ettinger, 2019). The general approach for gaze estimation without the use of CRs is to build a 3D model of the eye and to estimate the position of the eyeball centre using the pupil centre. From there, the optical axis can be estimated and gaze estimation can be made using further measurements of the pupil from the eye camera(s).

There are general considerations for every eye tracker regarding the setup and use to maintain high spatial accuracy during recording. Arguably, the most important of which is the calibration, since it underpins the model for gaze mapping. This usually involves the subject looking at a series of points, or in some cases a single point. For remote systems, the calibration procedure provides reference points to map raw eye tracking data to screen coordinates, and for head-mounted systems to points in the environment video. Calibration may be followed by a validation sequence to verify the accuracy of the calibration, and if necessary, the calibration is repeated.

Godwin et al. (2021) provides an in-depth review of important methodological considerations and common mistakes to avoid in eye tracking experiments. These range from the setup and calibration to the data analysis, to advise on how to most accurately track participants and appropriately extract information from the data.

Various factors can cause issues with the calibration procedure. Eye make-up such as mascara can interfere with the pupil detection, since it makes eyelashes thicker and darker, causing problems if they overlap the pupil edge. It is often advised that participants remove any eye make-up prior to a recording session. However, mascara is only an issue when the edge of the pupil is close to the eyelid; when there is a reasonable amount of white (sclera) visible around the pupil, mascara does not interfere. There may also be a problem when the top or bottom part of the pupil is covered by the eyelid.

Glasses and contact lenses commonly cause issues with many video-based eye trackers by interfering with the reflections from the IR source or creating additional reflections to the corneal reflection that will defeat the tracking algorithm. Soft contact lenses are generally not an issue compared to hard lenses. Solutions for obtaining an accurate calibration with glasses and contact lenses include changing the distance of the eye camera(s) from the eye and tilting the angle relative to the eye, or (for monocular tracking) changing the eye that is being tracked (Klein and Ettinger, 2019). Glasses usually work fine with the EyeLink1000 desktop eye tracker, as does the Pupil Core eye tracking headset when using the provided extensions to the eye-camera arms to increase eye-camera distance.

If calibration accuracy starts to drift during the experiment, some trackers allow for drift corrections to be performed during the experiment. This usually consists of the participant fixating a single point and the experimenter pressing a button once they have done so. If the fixation point is beyond a given tolerance from the drift correction point, the calibration procedure is repeated. The eye video is also shown during the recording session for most video trackers so researchers can manually monitor the participant to check that their eye(s) are still being tracked correctly. For example, the pupil and corneal reflection are still in focus, or that the eye has remained within the ROI window (particularly important for head-mounted trackers where the headset may be knocked during use).

Fixation durations can vary greatly, both between and within participants. Since there is no definition of a fixation in terms of duration, algorithms detect saccades and everything that is not registered as a saccade is thus a fixation. The fixation location is defined as the average X-Y location of all samples that are between consecutive saccades. The minimum and maximum threshold for the duration of fixations must be set by the algorithm. A tracker may split a single fixation into two separate fixations, resulting in two short fixations at a single location. Alternatively, it could group two fixations into a single long fixation at the same location. Of course, how fixations are registered will also depend on the tracker's definition of a saccade. Godwin et al. (2021) argue that a minimum duration of 60ms captures all short fixations, and fixations with a duration below this value are not long enough for any meaningful information to have been acquired for them to be included in the analysis. An upper limit of 1200ms was suggested, although this is less clear on how it should be decided since it is somewhat arbitrary and dependent upon the experimental design. Another suggested method was to remove fixations with durations greater than, for example, 2.5 standard deviations from the mean.

In addition to the duration of fixations, it is important to only include fixations that are within the area of the display that are of importance to the analysis, and experimenters should remove fixations according to their location (Godwin et al., 2021). For experiments on displays in radiology, the most appropriate way to do this would be to remove fixations falling outside of the images. Of course, fixations falling outside of the image may be due to measurement errors, so experimenters should take this into consideration when excluding fixations from analysis. Finally, it is often of interest to measure whether a region or object has been fixated. Since fixations are normally distributed around the centre of objects (Godwin et al., 2021), and there is some error in the accuracy of the tracker, the analysis must provide some tolerance to account for this. In radiology, and in visual search literature in mammography, a 2.5° visual angle radius is associated with the useful field-of-view (Kundel and Nodine, 2004). Therefore, this value can be used as a tolerance around objects for accepting or rejecting whether they were fixated.

5.3 Eye tracking devices

The experiments of this thesis have used two eye trackers, one remote and one head-mounted. The details of each of these will be explored in more depth in this section.

5.3.1 EyeLink1000

The EyeLink1000 is a remote desk mounted eye tracker, shown in Figure 5.4. As shown, there are infrared illuminators that produce a CR, which is then recorded by a high-speed camera. The position of the tracker and camera is adjustable with a number of knobs and screws, making it possible to change the angle of the illuminators and camera. The EyeLink setup screen is shown in Figure 5.5. From here, there are a number of important settings that can be altered, including: the thresholds for the pupil and CR that can be set automatically (and manually adjusted), the sample rate, and the power of the illuminator. The eye camera should be adjusted so that the eye is positioned centrally in the top camera image in Figure 5.5. The lower camera image in Figure 5.5 shows the dark pupil and the CR; the camera lens should be adjusted so that both of these are in focus, such that the pupil appears as a full circle and without shadows around other parts of the eye (particularly around the eyelashes).

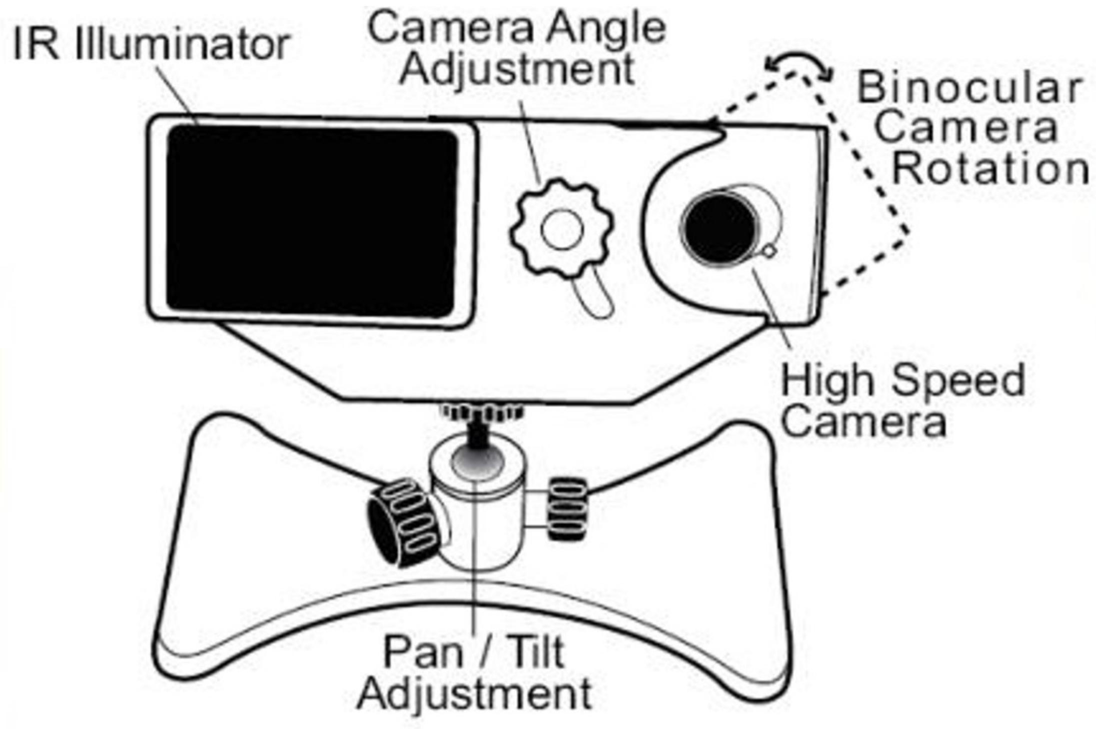


Figure 5.4: Diagram of EyeLink1000, taken from SR Research Ltd (2009).

The setup screen (Figure 5.5) is also where the calibration is initiated from. This is a 9-point calibration procedure where the position of each point is randomised. The process is automated, with the position of each point changing once the participant has fixated it. Once the calibration is completed, a validation procedure can be performed, which follows the same process and checks the accuracy of the calibration in calculating gaze position. If any fixation deviates by greater than 1° visual angle then the calibration should be repeated. Once an accurate calibration has been accepted, the gaze positions can be calculated using the pupil and CR centres (as described in Section 5.2).

The technical specifications for the EyeLink1000 are provided in Table 5.1. The accuracy is typically high (<0.5°) and a sampling rate of 1000 Hz for gaze location was used in our experiment. Participants are supported by a chin and forehead rest to

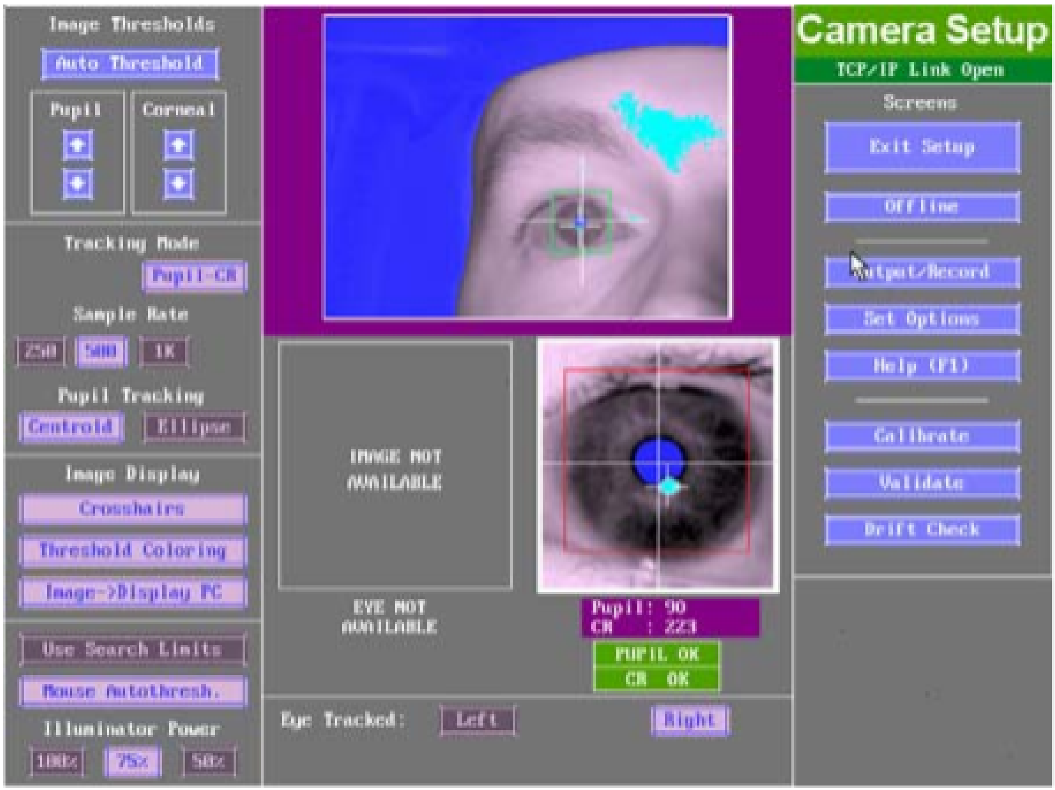
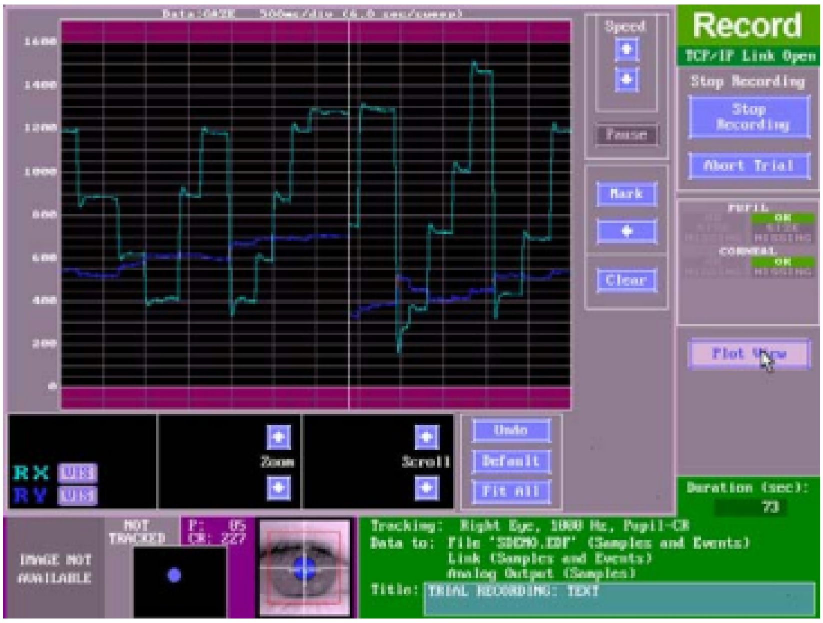


Figure 5.5: Screenshot of the EyeLink setup screen. The lower camera image shows the pupil centre crosshairs and CR centre crosshairs. Image taken from SR Research Ltd (2009).

restrict head movements. It is important to keep the camera within the optimal camera-eye distance (40−70cm) for data quality purposes. The data output is in the form of a single EyeLink data file (EDF). Raw gaze position and pupil size are generated automatically, with button and message events generated by the experiment programme and sent to the EyeLink system to be recorded. The real-time operator feedback is what can be monitored during the experiment, shown in Figure 5.6, where the eye camera image is shown along with the tracking status (‘OK’ or ‘MISSING’). The stimuli that the participant is viewing can also be displayed on this screen with the gaze cursor overlayed. Alternatively, a trace of their gaze coordinates as a function of time can be viewed instead.

Table 5.1: Specifications for the EyeLink1000 tracker (SR Research Ltd, 2009).

Property	
Average accuracy	0.25°–0.5° typical
Sampling rate	250, 500, or 1000 Hz
Spatial resolution	<0.01° RMS @ 1000 Hz
Allowed head movements	±25 mm horizontal or vertical, ±10 mm depth
Eye tracking principle	Dark pupil corneal reflection
Gaze tracking range	32° horizontally, 25° vertically
Optimal camera-eye distance	40–70 cm
Data output	raw eye position gaze position pupil size button events message events
Real-time operator feedback	Gaze position cursor or position traces Camera images and tracking status



Tracking: Right Eye, 1000 Hz, Pupil-CR
Data to: File 'SDEMO.EDF' (Samples and Events)
Link (Samples and Events)
Analog Output (Samples)
Title: TRIAL RECORDING: TEXT

Figure 5.6: Screenshot of the EyeLink record screen. The trace of the participant's gaze coordinates is displayed in real-time. The lower camera image shows the pupil centre crosshairs and CR centre crosshairs. The box on the right signals the tracking status. Image taken from SR Research Ltd (2009).

5.3.2 Pupil Labs

Pupil Labs - Pupil Core Headset (referred to as Pupil Core from here onwards) is a wearable eye tracking headset, shown in Figure 5.7. The headset consists of two 200 Hz eye cameras and a forward facing 120 Hz world camera mounted on a lightweight frame (total weight 22.75g). Each of the eye cameras can be adjusted along an arm to extend them further away from the eye or bring them closer, or they can be moved around a ball joint to change the angle relative to the eye. The world camera can be tilted up or down to change the field-of-view (FOV). Pupil Core allows for free movement of the head, and can be connected to either a laptop or mobile device via a USB-C connection for data collection. All of the Pupil Labs software is open-source, making it possible to adapt the software to specific needs and even add custom features. There are many third-party features available (`https://github.com/pupil-labs/pupil-community`) that can be easily added to the software provided by Pupil Labs.

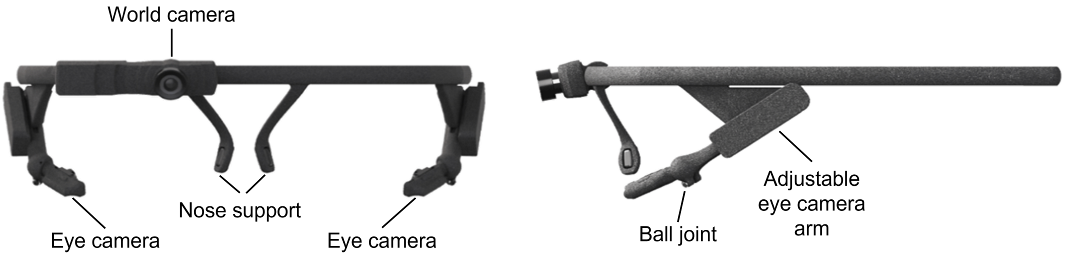
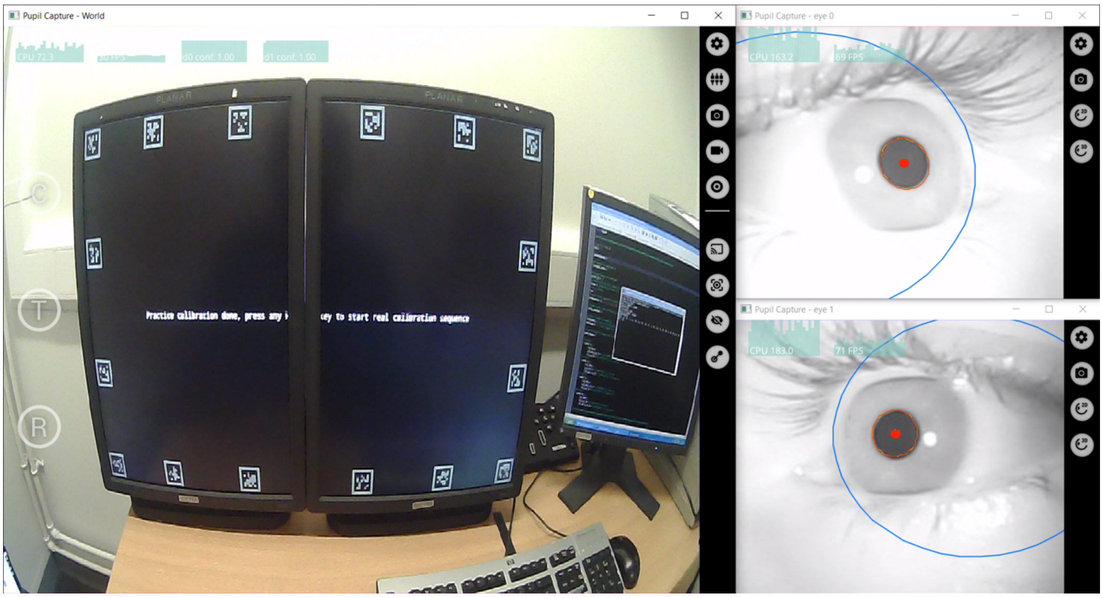


Figure 5.7: Photograph of Pupil Core. The world camera can be titled vertically about the frame to align with the desired FOV. The two eye cameras can be moved around a ball joint and along the camera arm either closer or further from the eye. Image provided by Pupil Labs with permission to reproduce, with added annotations.

Figure 5.8 shows a screenshot of the Pupil Capture software used for the setup of Pupil Core. The eye camera images can be viewed to ensure that the cameras are properly positioned. Pupil detection is an automated process, discussed in further detail below, but may require fine tuning. This can be done by altering the ROI around the eye, changing the minimum intensity value of a pixel to be considered as part of the pupil, or setting the minimum and maximum pupil sizes. Once the participant is positioned in the experiment environment, the world camera should be adjusted so that the experiment screen or task is in view in Pupil Capture, and the camera lens can be focused if necessary.



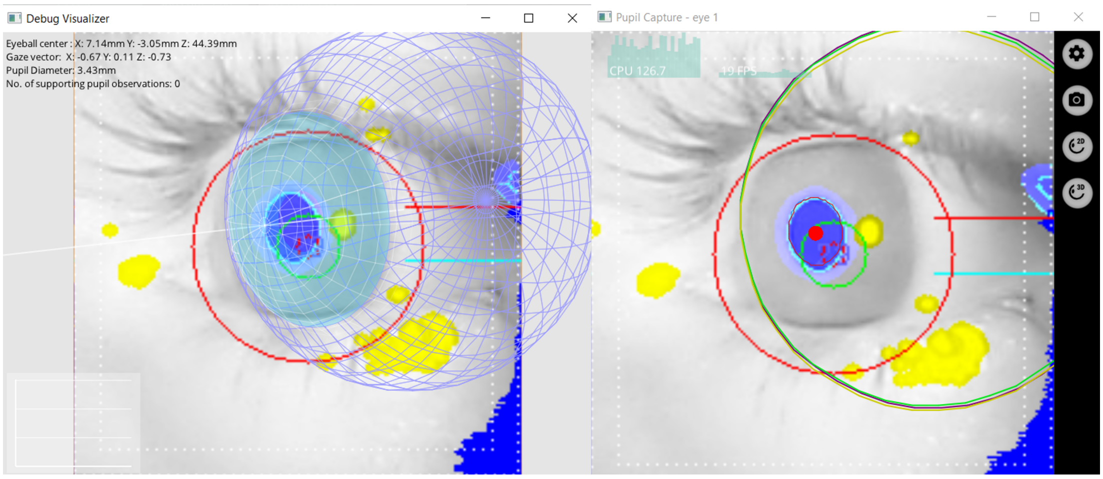
Practice calibration done, press any key to start real calibration sequence

Figure 5.8: Screenshot of Pupil Capture software used for data collection with Pupil Core. The eye camera images are shown on the right: red circles around the pupil indicate that they have been properly detected.

There are three different calibration types to choose from with the Pupil Core. First, which is the standard in visual search tasks, is the screen marker calibration, where a series of calibration targets are displayed on a screen to be fixated. For search tasks that are not screen-based, calibration can be performed either using a single target (printed or otherwise displayed), or by fixating a series of natural features in the environment. The accuracy of the calibration can be estimated in Pupil Capture with an accuracy test, repeating the calibration procedure and displaying the errors between estimated gaze and the reference targets.

Surrounding each eye camera are IR LEDs that illuminate the eye (dark pupil method), and from these camera images both the pupil boundary and centre are extracted. This process uses an algorithm involving a series of image processing transformations (Kassner et al., 2014). The eye algorithm view in Pupil Capture is shown in Figure 5.9. First, the image is converted to greyscale and an initial rectangular estimate of the pupil region is made. Next, an edge detection process is followed to find contours, which are then filtered based on the pixel intensities. The dark pupil region is identified by thresholds set by the user (discussed previously), detecting the lowest pixel intensities in the histogram of the image. Other reflections in the image are filtered out that do not fall inside the dark pupil region. The remaining edges around the dark pupil region are connected to form sub-contours, and subsequently used to create various candidate pupil ellipses via ellipse-fitting. Finally, the properties of the ellipses are assessed and if they meet a confidence threshold, the best of those that are above the threshold are used to define the pupil contour and centre.

In addition to the 2D pupil detection, Pupil Capture also models the eye in 3D using the camera images of the eye. The initial steps are based on a method developed by Swirski and Dodgson (2013): the position of the eyeball is estimated without the use of corneal reflections, and instead uses multiple measures of the pupil in a process of ‘least-squares intersection of lines’. This method is outlined in Figure 5.10, where the eye is modeled as a sphere of radius R. A series of images taken over a period of time from the eye camera provides a set of 2D pupil edges (using the process described previously), and the lines normal to the pupil centre are parallel to the gaze direction. This set of lines, normal to the gaze direction and through the centre of the eyeball, provides an estimate for the position of the eyeball.



```
Eyeball center : X: 7.14mm Y: -3.05mm Z: 44.39mm
Gaze vector:  X: -0.67 Y: 0.11 Z: -0.73
Pupil Diameter: 3.43mm
No. of supporting pupil observations: 0
```

Figure 5.9: Screenshot of eye camera image showing algorithm view in Pupil Capture. The two red circles indicate the pupil min and max sizes set by the user, and the small green circle is the current pupil size. The debug visualiser (left) also shows the eyeball model.

However, additional steps are added to the estimation of eyeball position, expanding on the model by Swirski and Dodgson (2013). Pupil use a two-sphere model, accounting for the curvature of the corneal surface, detailed in Dierkes et al. (2019). The model expands upon that shown in Figure 5.10, with sets of 3D pupil contours defining sets of intersecting lines through the eyeball centre to estimate its position. Once the eyeball position has been estimated, the gaze estimation can take place. This is done by estimating the centre point of the pupil in 3D, and therefore the line through the pupil centre and eyeball centre gives the optical axis. Finally, a correction is applied to account for the effects of corneal refraction, which can incur systematic errors into the calculation of eyeball and gaze position (Dierkes et al., 2019).

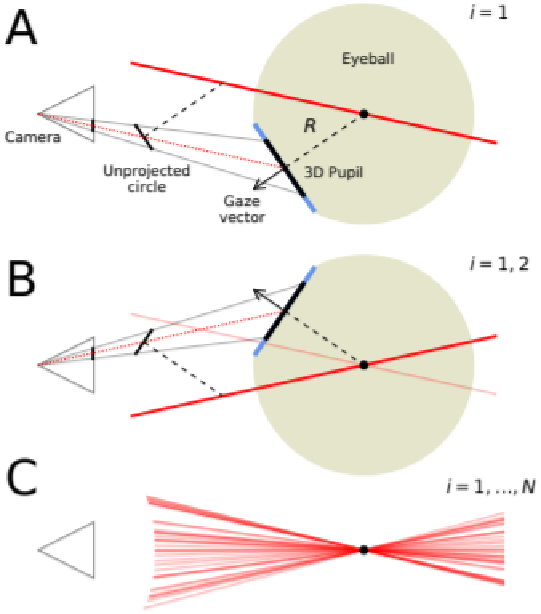


Figure 5.10: Diagram of method of estimating eyeball position using the intersection of lines based on pupil measurements proposed by Swirski and Dodgson (2013). (A) Initially, the 3D pupil surface is mapped to a 2D ellipse (black circle shown in side view). The red dashed line is formed through the centres of the circular unprojections of the 2D pupil ellipse from the eye to the camera, passing through the pupil centre. The eyeball to pupil centre distance, R, is fixed, and therefore the eyeball centre is constrained to lie on the red solid line, parallel to the red dashed line. (B) A change in gaze direction gives rise to an independent constraint for eyeball position. (C) Eyeball position is estimated by a least-squares intersection of constraining lines derived from a set of pupil contours from N gaze direction measures. Diagram taken from Dierkes et al. (2019).

Calibration and gaze mapping can be performed within Pupil Capture with the gaze position overlayed on the world video during recording, but calibration can also be done post-hoc once the recording has finished. This is done in Pupil Player, the analysis software for Pupil Core, shown in Figure 5.11. An offline calibration can be performed if, within the recording calibration, targets have been displayed and a correct procedure has been followed by the participant (i.e., reference locations known to have been fixated by the participant). Calibration targets are automatically detected by the software and can be manually removed or added if required. A range of the targets can be chosen to be used as a validation set. Once the offline calibration has been completed, gaze mapping and fixation detection can be (re)calculated using this. Another feature in Pupil Player is the option to add annotations, which allows the user to add a custom label to a given timestamp to flag the occurrence of an event in the recording.

Pupil Capture and Player also offer surface tracking, where surfaces can be defined by using QR-code like markers within the experiment environment. An example of this is shown in Figure 5.11, where surface markers are displayed around the edge of each monitor. Using the markers, four surfaces were defined by selecting appropriately positioned markers to define them (minimum of 2 markers required for surface definition) and then creating the desired shape within the environment. The surface will then remain fixed within the video according to the chosen parameters and visible while the markers are within view. Once the surfaces have been defined, gaze and fixations can be mapped onto the surfaces and heatmaps of the recording are produced.

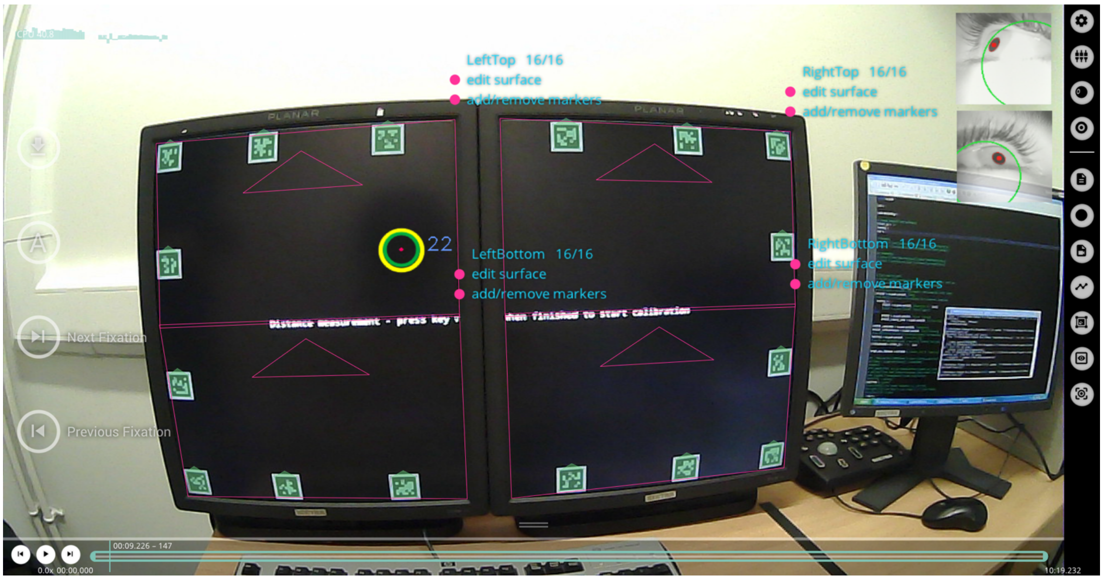


Figure 5.11: Screenshot of Pupil Player software used for data analysis with Pupil Core. On the left-hand side are the hotkeys to export the data, add an annotation, and skip through fixations. On the right-hand side is the menu to change the settings, manage plugins, and initiate the offline calibration and gaze mappings. The QR-like markers around the edge of the screen are highlighted in green when properly detected. They are used to define surfaces, which in this case are the four pink rectangles over the two displays. The gaze position is displayed on the left screen (green circle with red dot at centre) and a yellow circle indicates a fixation at this point. The eye camera overlays are displayed in the top right of the screen, showing the detected pupil features.

Through the use of surface markers, Pupil Player is able to build a 3D model of the environment to support head pose tracking (an example is shown in Figure 5.12). This provides the head pose of the world camera within a 3D coordinate system with one of the markers set as the origin. The data output of the head pose tracking provides the rotation and translation of the headset in the 3D coordinate system relative to the origin marker.

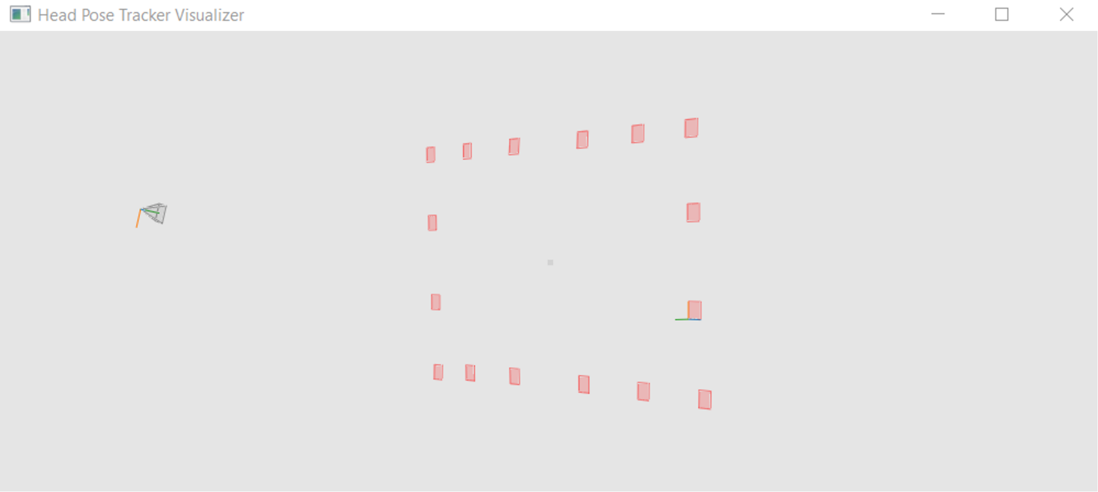


Figure 5.12: Screenshot of head pose tracking visualiser in Pupil Player software. The markers around the edge of the screens are shown in red in this view, with the Pupil Core headset on the left. The marker third down on the right is set as the origin for the coordinate system.

The technical specifications of the Pupil Core headset are given in Table 5.2. The typical accuracy is stated as 0.60°, but will vary depending on the experimental setup and has been shown to drift over the course of the experiment (Ehinger et al., 2019). Head movements are not restricted with Pupil Core, however rapid movements may cause slippage of the headset and will lead to the software briefly losing track of any surface markers in the environment. In addition to the gaze position and fixations (both in the world video and mapped onto surfaces), the pupil size and position data are also exported from Pupil Player. As shown in Figure 5.8, while the recording session is taking place, the researcher can monitor the gaze position on the world video (provided a calibration has already occurred) and the eye camera images, which display the pupil ellipses and centre, and the respective algorithm confidence threshold values.

Table 5.2: Specifications for Pupil Core (Pupil Labs, 2021*a*).

Property	
Average accuracy	0.60° typical
Sampling rate	200 Hz eye camera 120 Hz world camera
Spatial resolution	<0.02° RMS
Allowed head movements	Unrestricted
Eye tracking principle	Dark pupil with 3D model
World camera FOV	100° horizontally, 74° vertically
Data output	gaze position pupil size and position fixations annotations surface tracking data head pose tracking data
Real-time operator feedback	Gaze position cursor Camera images and pupil contours Pupil tracking confidence

5.4 Visual search studies

5.4.1 Studies in mammography

Eye tracking has been used in studies in mammography to gain a better understanding of how readers search mammograms, how different workstation setups affect performance, and ultimately to find methods to improve diagnostic performance. Kundel et al. (2007) recorded the eye positions of trained readers of mammograms with varying levels of experience whilst they interpreted 40 mammograms, half containing a subtle cancer (see Figure 5.13). Half of the cancers were fixated in around 1 second, with a median time to fixate a cancer of 1.13 seconds. In another similar study, the majority of cancers (approximately 57%) were fixated within 1 second (Kundel et al., 2008). The fact that abnormalities were fixated so quickly indicates that there was some initial rapid interpretation of the image and supports their model (Kundel and

Nodine, 2004) of a global analysis of the entire image followed by a closer examination of regions identified as suspicious.

This global approach was shown to be more proficient than an exclusive use of the ‘search-to-find’ technique where a focal search is carried out by scanning the image to recognise and evaluate abnormalities (Kundel et al., 2007). Furthermore, it demonstrated that the process of a global impression leading to a rapid review of the suspicious region was occurring before the scanning phase of the model (as opposed to an earlier model where scanning was believed to occur prior to foveal verification (Nodine and Kundel, 1987)).

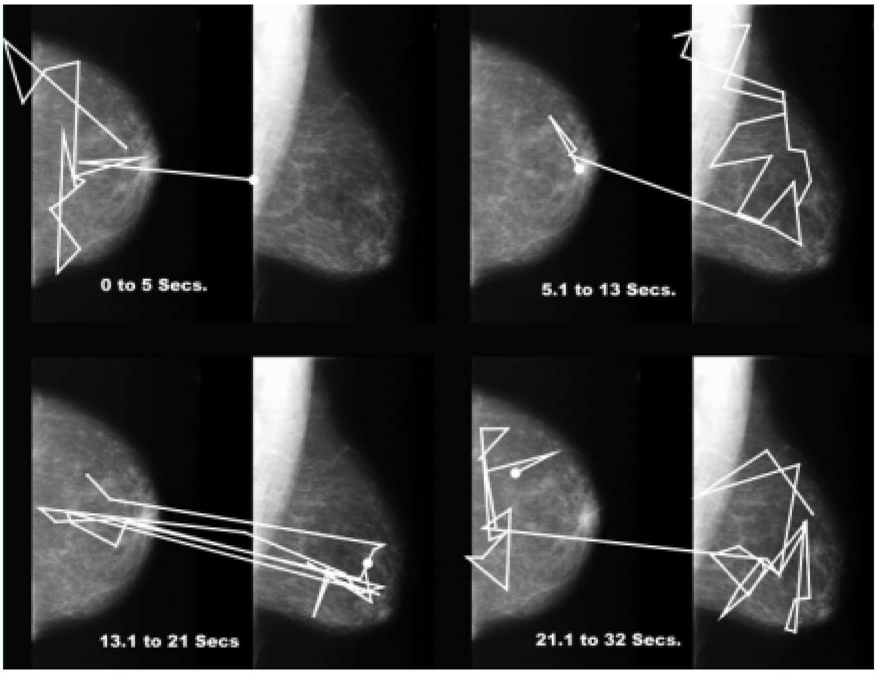


Figure 5.13: Image from the study by Kundel et al. (2007) showing the visual search path of a mammographer viewing an abnormal mammogram. The search starts with an initial saccade to the mass in the CC view, followed by a long saccade to the mass in the MLO view. Comparisons between the views are made and finally both views are searched completely.

The search strategy of the reader also depends on their level of expertise (Kundel et al., 2007). Less expert readers were not able to employ the initial global impression of the mammograms that would rapidly guide them to potential abnormalities. Instead, they adopted the slower search-to-find method, which also resulted in more errors (Kundel et al., 2007). Once it was established that experienced readers follow this model of search, Mello-Thoms (2009) investigated the impact of the initial global perception of the image being incorrect. When readers were initially drawn to a false positive area, they were significantly more likely to make false negative errors (fixate abnormalities without reporting); effectively they became blinded to the characteristics of true positives in those mammograms due to an initial erroneous view.

The commonalities between visual search strategies of expert readers of mammograms was investigated by Mello-Thoms (2008), and while the agreement between pairs of readers varied, there tended to be a similarity between search patterns once readers had fixated the location of response. Studies of how the most experienced readers search images could be particularly useful in the training of new readers. Although it is important to note that there is not going to be a search approach suitable for all images, rather, useful techniques such as how comparisons between views could be formulated (Mello-Thoms, 2008).

A later study by Gandomkar et al. (2018) investigated the use of eye tracking parameters to pair radiologists with the aim of optimising overall performance. They extracted 14 metrics from the eye tracking data and grouped readers by experience before pairing, with an example of how three of these metrics are obtained demonstrated in Figure 5.14. The eye tracking metrics differed greatly between experience groups compared to within groups. In >50% of all possible pairings, the maximum benefit was not achieved, suggesting that double reading may not always provide improved outcomes and depends on the combination of readers. In terms of maximising performance, pairing readers using cognitive measures, rather than their AUC values, was more robust to small sample sizes (Gandomkar et al., 2018).

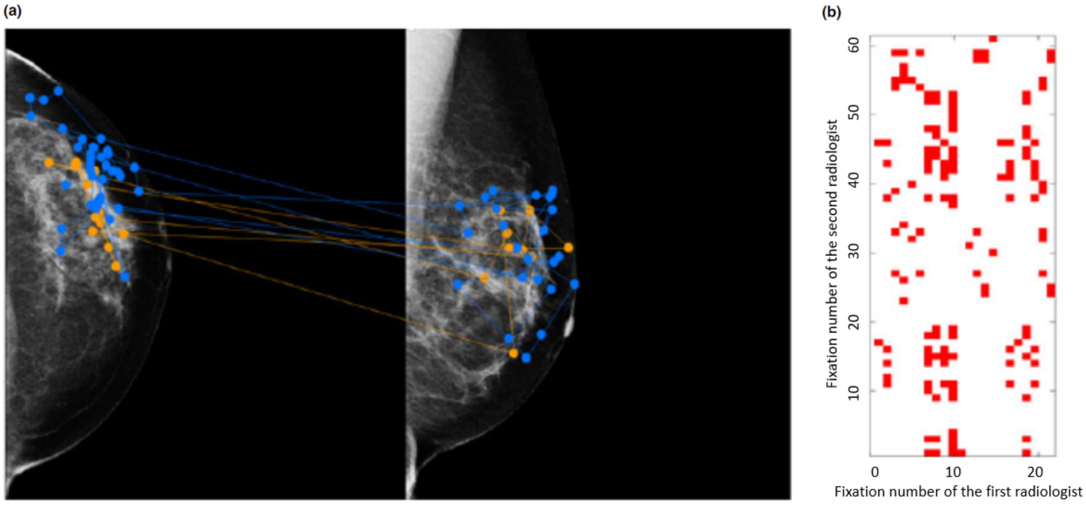

(a)
(b)
Fixation number of the second radiologist
10
20
30
40
50
60
0
10
20
Fixation number of the first radiologist

Figure 5.14: Cross-recurrence plot for two radiologists, adapted from Gandomkar et al. (2018). (a) Scanpaths from two readers (one in blue and one in orange), where the circles are fixations and lines are saccades between fixations. (b) Cross-recurrence plot corresponding to the scanpaths. This plot indicates the order that fixations occurred for each reader. A point in the *i*th column and *j*th row indicates that these fixations were within 2.5°. From these plots, three metrics were extracted: recurrence (regions fixated by both readers), determinism (similarity in their sampling strategies), and laminarity (for regions covered by both readers, the proportion of consecutive fixations).

Further studies have explored the differences in visual search behaviour of readers varying in experience. Less experienced readers were shown to fixate for shorter periods than those with more experience, and also looked in different regions of the image (Lévêque et al., 2019). Less experienced readers were also shown to have a more dispersed search (fixations had a greater spread across the image) than experienced readers who tended to focus for longer on single regions (Leveque et al., 2021).

In this study, all cases were lesion-free and so from these results it is not possible to know whether more experienced readers would spend more time fixating abnormal regions. Similarly, in an earlier study (Krupinski, 1996) experienced readers detected lesions earlier and spent less overall time in the image, covering less of it than more novice readers. These results align the search model discussed by Kundel et al. (2007), with less experienced readers opting for the slower search-to-find technique, and may also indicate that experienced readers have a better understanding of where they should look in an image.

There has been a number of studies that have tracked readers' gaze to gain a deeper understanding of the errors made in mammography. Krupinski (1996) reported that experienced readers made more decision than recognition errors, whereas this pattern was reversed for less experienced readers. This suggested that the latter were not fixating for long enough on potential abnormalities to register them as such. Mello-Thoms (2003) observed similar results for less experienced readers, however reported that experienced readers made equal numbers of decision and recognition errors. This contradiction is likely due to the difference in case sets between studies: single view versus two-view (Mello-Thoms, 2003). The time spent interpreting the image can also impact the error rate, with longer interpretation times (Saunders and Samei, 2006) and total dwell time (Voisin et al., 2013b) both being correlated with an increased chance of making a diagnostic error. Carney et al. (2012) reported that for each additional minute of viewing a cancer case, the chance of a true positive increased by 12%. However, for normal cases, longer interpretation times were associated with an increased risk of a false positive response, which was significant across all confidence levels. Each additional minute of viewing increased risk of a false positive by 20% for readers who were not at all confident in their decisions and by 42% for readers who were very confident.

A detailed study into the errors made in mammography and visual search of mammograms in general was conducted by Wolfe et al. (2021b). Seventeen readers of varying experience reviewed 80 (60 target present) single-view mammograms, tasked with detecting the presence of a single mass. No differences in performance or eye tracking parameters were found between readers of different experience levels. The proportion of errors made were in line with Krupinski (1996), with 25% search, 25% recognition, and 50% decision for trials where targets were missed without a false positive response. To investigate the nature of these errors in more detail, the proportion of saccades that went to the target region immediately following a fixation at various distances from the target was measured, shown in Figure 5.15a. As shown, the proportion of saccades that went to the target was $\approx$33% when the fixation was 2° from the target. This rises a small amount to around 50% when looking at the proportion of the next three saccades from a fixation 2° from the target (Figure 5.15b). Therefore, it suggests that even when attention is focused inside the UFOV in mammography of 2.5° radius estimated by Kundel and Nodine (2004), it does not necessarily mean that a target will be detected. This result may also go some way to explain inattentional blindness, discussed in Section 4.3, since readers may fixate near a target without processing it as such.

A reasonable assumption with search errors, where the target is not fixated and subsequently unreported, may be that the image was not searched as completely as in true negative responses or compared with recognition and decision errors. However, in terms of image coverage, this was not found to be the case (Wolfe et al., 2021b). Coverage did not vary between cases where a true negative response was given and cases where the target was missed. On cases where a target was missed and without a false positive response, the image coverage did not vary between error type. Search errors, it seems, did not result as a consequence of readers quitting search too early (Wolfe et al., 2021b). The high target prevalence of 75% and the experimental setup may have impacted the way in which observers searched the mammograms, since the images were single-view with no access to other views for a single case or indeed prior mammograms as would be available in the clinic. However, the interesting insight into how targets are missed is unlikely to change under more realistic circumstances, but this should be verified.

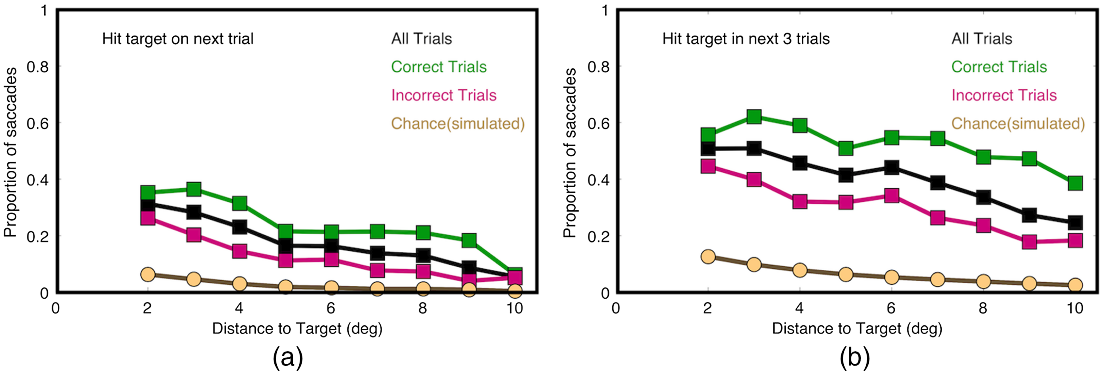


Figure 5.15: Image from the study by Wolfe et al. (2021b) showing (a) proportion of next saccades that hit the target and (b) proportion of the next 3 saccades that hit the target, versus distance of current fixation from the target. Distributions are shown for all trials, and split further by correct and incorrect trials. A simulated chance distribution is also shown, which estimates how an observer would perform if attention was moved around the image to check locations where a mass may be expected, rather than moving to a location that has already raised suspicion.

The type of image and the way they are presented to readers is important to search behaviour and performance. Radiologists were found to have different visual search behaviour, in terms of time to hit first lesion and median dwelltimes on cancers, for digital mammograms compared to digitised screen-film mammograms (Mello-Thoms, 2010). Digital mammograms reduced the number of cancers that did not attract any visual attention and increased the number of cancers that did attract visual attention for over 1 second. It is conventional for mammography workstations to use two clinical monitors for displaying the mammograms, however using a larger and higher resolution single monitor may be more beneficial for visual search (Krupinski, 2016). Using a single monitor significantly reduced search time without a reduction in diagnostic accuracy.

Drew et al. (2015) conducted a study with 23 expert readers comparing the conventional method of viewing past and present mammograms side by side on a 21 inch clinical monitor, with a method where images could be 'toggled' between one another. There was an average 6 second reduction in the time taken to reach a decision with the toggle mode, and a small but not significant improvement in diagnostic accuracy of 5%. Eye tracking was not used in this study, but it is likely the improvements were partly due to fewer long saccades and subsequent fixations between side by side past and present images to compare regions of interest, relative to toggled images where gaze is fixed at a single location. The features of the image, in addition to how they are displayed, will also impact the visual search of readers. Mousa et al. (2014) investigated how breast density changed the visual search patterns of seven radiologists. Lesions that were overlying fibroglandular dense tissue in both low and high density images were fixated faster and for longer than lesions located outside dense regions. These are the regions where lesions may typically be hidden by dense tissue and so the gaze of the reader is likely to be attracted first.

An alternate modality that requires the reader to 'toggle' through a stack of images is DBT, discussed briefly in Section 2.3. Since DBT images are 3D, visual search behaviour must adapt. Aizenman et al. (2017) tracked eye movements of radiologists reading single-view DBT and digital mammography cases on a 21 inch clinical monitor, finding significantly improved sensitivity and specificity for DBT. Readers made longer fixations in DBT, but similar amplitudes of saccades. A smaller area of the images was covered for DBT than digital mammography, but the 3D nature of DBT images complicates this calculation somewhat. For the DBT images, Aizenman et al. (2017) marked regions defined by a useful field-of-view (UFOV), for slices visited during each fixation and also the slices directly above and below. Coverage could then be calculated by dividing the marked regions by the total image volume.

Dong et al. (2018) studied eye movements of radiologists experienced in reading DBT images, with the aim of gaining a deeper understanding of their visual search strategies that could potentially be used as a training tool for future trainees starting to use DBT. Readers were presented with the synthetic 2D-DBT image first, followed by the 3D-DBT image. Gaze analysis was performed to track overall visual coverage and on AOIs, and scrolling behaviour was recorded throughout. Readers used the initial 2D view to perform a global-focal scan, identifying possible abnormalities. They then scrolled through the 3D view to inspect those regions in greater detail, spending more time fixating them than in the 2D view. Another study using eye movement data in DBT is Jiang et al. (2017), which recorded naive reader (physicists) gaze while reading DBT images to extend a visual search model observer for mass detection.

There is a need to improve the perceptual ability of less experienced readers through training from experienced readers. Nodine et al. (1999) studied performance of various levels of expertise groups reading mammograms. Less experienced readers struggled with perceptual recognition and decision making (identifying differences between normal, benign and malignant image regions) due to lack of 'perceptual-learning experience' during training. Methods have been proposed to use gaze information to improve diagnostic accuracy, with the idea that false negatives tend to have significantly longer dwell times than true negatives (Krupinski, 2010), and so feeding back areas that attracted attention but were not reported on could lead to a reduction in missed cancers. These are known as computer assisted perception (CAP) tools.

Nodine et al. (2001) implemented a method called ‘computer-assisted visual search’ (CAVS), where regions that were fixated for 1 second or longer were used as feedback after initial unaided viewing. CAVS increased the detection and localisation of lesions by 12%, but this was not significant (this may have been a power issue with only 6 readers and 40 cases). Similarly, Tourassi et al. (2010) describe an interactive information-theoretic CADe (IT-CADe) system for mammography which recorded regions that were fixated for longer than 1 second but not reported by the reader. These regions of interest are compared with a database of mass and normal templates to determine the probability of that region being malignant. This system demonstrated potential to reduce false negatives, detecting 4/6 decision errors and 5/8 search errors.

An interesting idea presented by Gandomkar et al. (2017) combined the gaze data of radiologists with extracted image features, of both the mass (if present) and the surrounding background, to classify locations marked by radiologists as TPs or FPs. This method improved performance of radiologists in the study; lesion localisation accuracy increased by an average of 12%, along with an average of 44.5% decrease in the number of FPs per image from 0.46 to 0.21. Chen and Gale (2010*b*) proposed using the gaze data of an experienced reader of mammograms as a training tool for inexperienced readers. However, Chen and Gale (2010*b*) state that this alone is unlikely to improve performance significantly, but it may be useful to combine CAD output with regions that attracted expert readers’ gaze.

Recent work has investigated using gaze recordings of expert readers to train deep learning algorithms. Mall et al. (2018) used the eye tracking data of eight radiologists to train a deep learning model to predict reader behaviour. The areas of the image that were fixated directly, indirectly (peripherally), or not at all, were combined with the behavioural data (decisions and their confidence in that decision). These combinations of fixated regions and decisions were used to train a model to predict the readers decisions and the areas of mammograms that are likely to either attract the attention of readers or to be ignored. These were able to be accurately predicted by the model. Since for a given mammogram, the visual search pattern of a single reader is known to be variable in repeat sessions, Mall et al. (2019a) modelled attention level rather than, for example, scanpath. These attention levels were: 'foveal areas' where 3 sequential fixations occur within 2.5° of each other, 'peripheral areas' where <3 sequential fixations occur within 2.5° of each other, and 'never fixated areas' where none of the 8 readers fixated. In Figure 5.16, two of the attention levels (foveal and peripheral areas) used to train the deep learning network are visualised. It was demonstrated that the model was able to accurately predict the radiologists' attention levels and decisions on mammograms.

Finally, gaze data was used to explore the underlying features of regions for different error types: search, recognition and decision errors (Mall et al., 2019b). The features for each error type were used to train machine learning methods to predict the type of missed cancer and compared that to a CNN. The machine learning methods were better suited to determining the type of missed cancer (false negative) compared to the CNN, which learns the features by itself. By identifying the type of error that is associated with a false negative region, it provides a better understanding of why the cancer was missed and therefore training can be customised accordingly to improve performance for those regions (Mall et al., 2019a).

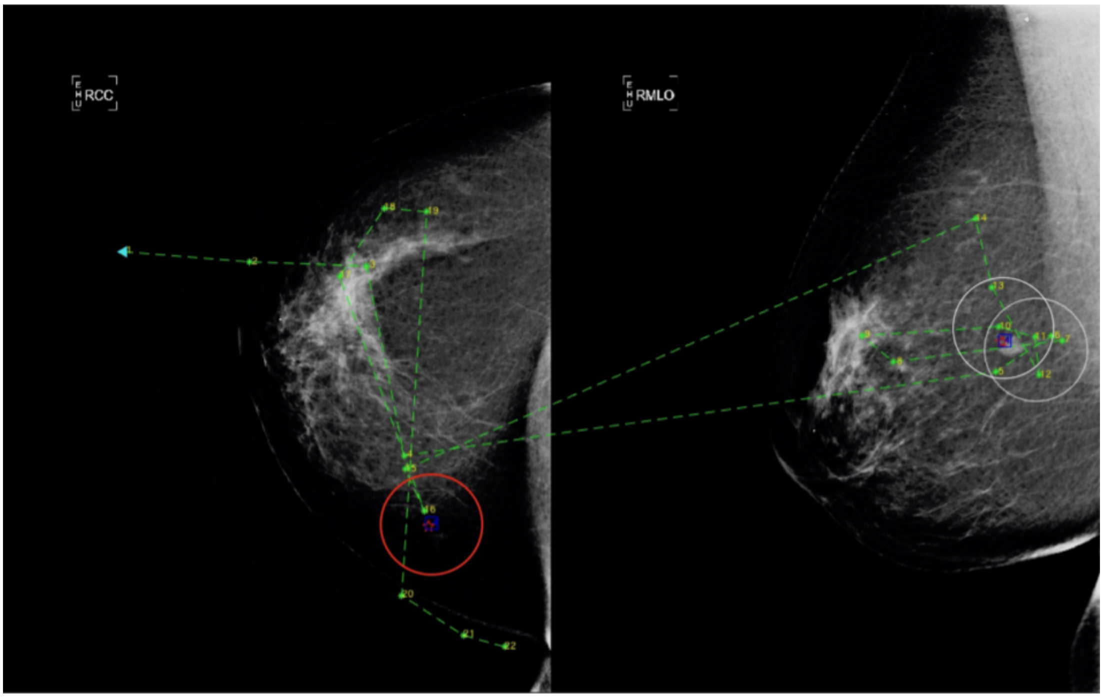


Figure 5.16: Images demonstrating how different attention levels were defined (Mall et al., 2019a). Scanpath of a reader for a two-view mammogram. The red star indicates an abnormality and a blue square where a reader has marked it as such. The green points are fixations and dashed lines saccades between them. The white circles are foveal areas and the red circle is a peripheral area. For this case there were areas that were never fixated.

5.4.2 Studies with CAD

Prompts have been shown to affect the visual search patterns of naive observers (Hatton et al., 2004; Drew et al., 2012). Drew et al. (2012) performed an eye tracking study investigating the effect of CAD on visual search of novice observers searching for target Ts amongst distractor Ls in a 1/f$^{2.4}$ noise distribution. CAD improved sensitivity by 7% and decreased specificity by 3%. However, when CAD missed a target, sensitivity fell to 56% (significantly lower than the no-CAD group) compared to 97% when CAD marked a target. This could be explained with the eye tracking data. Where targets were not marked by CAD, observers spent less time fixating it compared to the no-CAD group, or never even fixated it. Observers in the CAD group also had a much lower total image coverage (see Figure 5.17), mainly fixating around the prompts. This was also seen in the study by Hatton et al. (2004) for naive observers trained to interpret mammograms; more of the image was searched when there were no prompts present. A similar result was found for CT colonography, where observers gaze was attracted by CAD marks but areas not marked by CAD were viewed less than observers not using CAD (Helbren et al., 2015). This effect was stronger for readers with less experience.

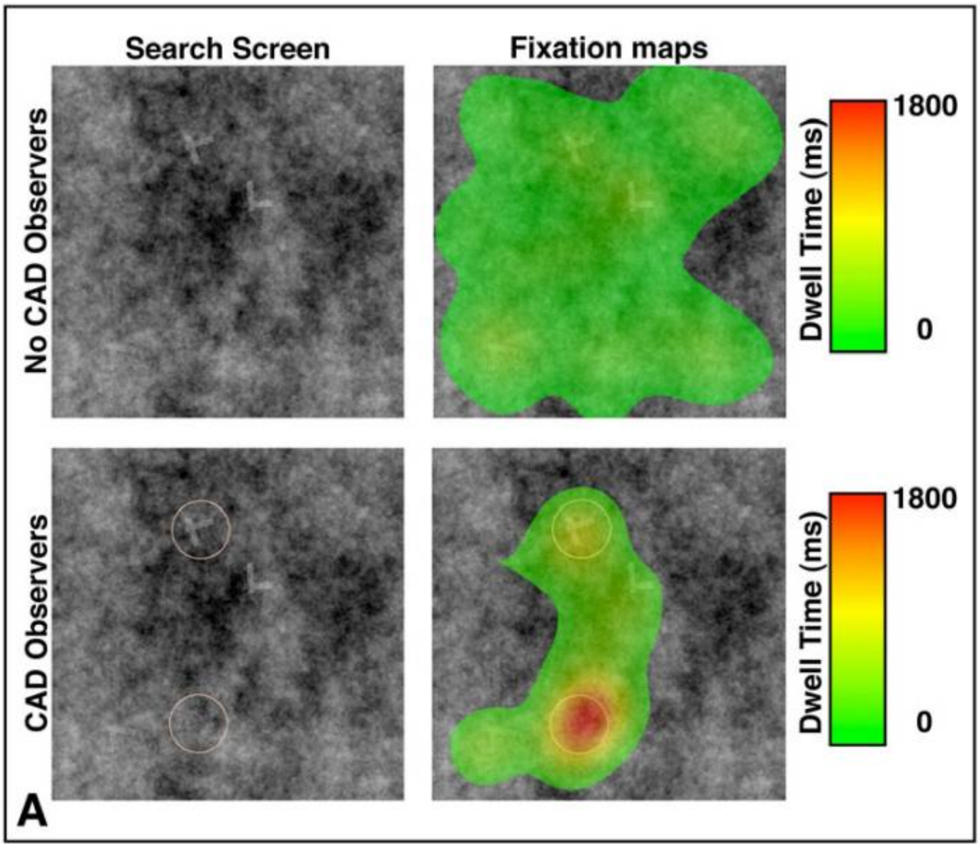


Figure 5.17: Image from Drew et al. (2012) showing the visual search patterns for a target absent trial, for a search of a target Ts amongst distractor Ls. Two CAD prompts are present in the CAD condition (bottom row), focusing the participants visual attention around them.

The main aim of Hatton et al. (2004) was to compare subtle prompts to traditional CAD prompts (shown in Figure 5.18). They found that with subtle prompts the number of bilateral comparisons were most similar to the unaided condition. Furthermore, participants turned the subtle prompts on and off a fewer number of times, since they were not overlaying important features of the mammogram. As mentioned, participants focused their visual attention primarily around the prompts (see Figure 5.18), as quantified by the percentage of overall time viewing prompts and percentage of total fixations on prompts. Both of these metrics revealed significant differences in search behaviour between prompts and no prompts, but no difference in the type of prompt used. Subtle prompts provide an alternative to typical CAD prompts and may be less disruptive to visual search in terms of switching prompts on and off and reducing viewing time compared to traditional prompts, while still attracting attention to areas readers might have missed in unaided viewing. However, since subtle prompts result in significant reductions in image coverage, CAD should still preferably be used as a second reader.

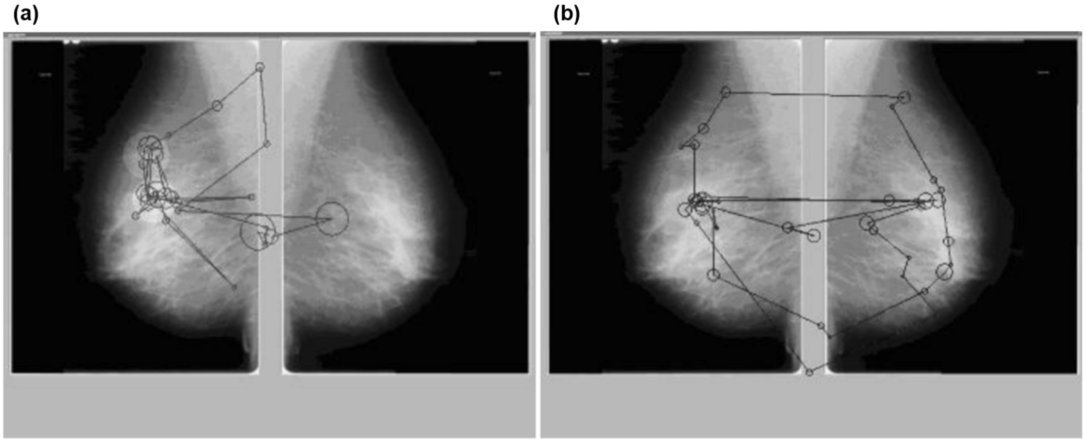


Figure 5.18: Scanpath of a non-expert participant reading a mammogram with (a) subtle prompts versus (b) no prompts, from Hatton et al. (2004). The black circles are fixations, with the longer the fixation duration the larger the circle. The lines joining fixations are saccades.

The effect of target prevalence on the visual search behaviour and error rates of non-expert participants completing a task (searching for target Ts among distractor Ls) with and without CAD was investigated by Drew et al. (2020). Two separate experiments were conducted, one with traditional CAD prompts (referred to as binary CAD) and the other with an interactive CAD approach, in which prompts only appeared when participants clicked on the regions where they were present. In each experiment there was a low prevalence (10%) and a high prevalence (50%) condition, both completed with and without CAD. Eye tracking revealed that binary CAD caused participants to search less thoroughly, with a significantly lower cumulative dwelltime on distractors compared to the no-CAD condition, and a significantly lower fraction of distractors fixated in target absent images. For interactive CAD, the opposite effect was observed for distractor dwelltime, with participants fixating for longer in the CAD condition than the no-CAD. The overall image coverage is visualised in Figure 5.19 for each condition, demonstrating similar results to Drew et al. (2012) for binary CAD.

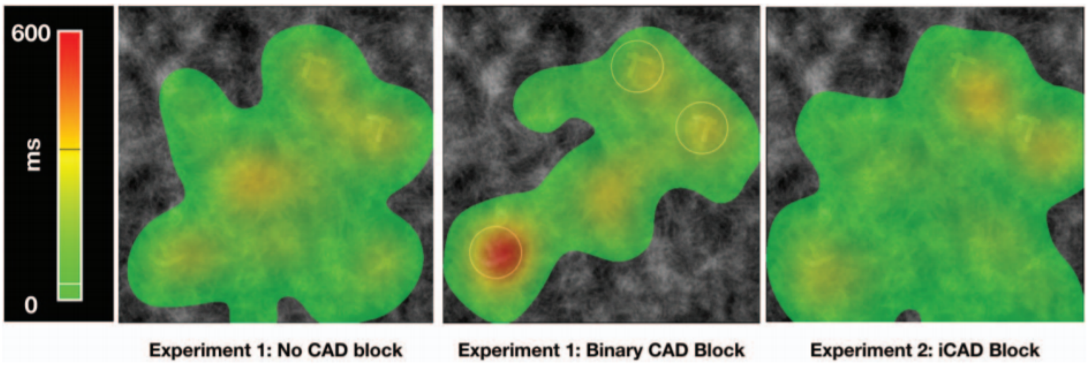


Figure 5.19: Heatmaps of visual search of participants in no-CAD (left), binary CAD (middle), and interactive CAD (right) conditions, from Drew et al. (2020). Each heatmap shows the mean fixation dwelltime for the participants that completed that condition.

Drew et al. (2020) reported that binary CAD where the target was prompted led to a significant improvement in the detection rate, with no effect of target prevalence. Interactive CAD did not improve detection rate at high prevalence, but did lead to a small benefit at the low prevalence condition. For targets that were not marked by CAD, for binary CAD there was a significant increase in miss error for those targets compared to the no CAD condition, which was significantly greater at low prevalence. This result is in agreement to that reported by Kunar et al. (2017), who hypothesised that participants became overreliant on CAD prompts leading to an increase in miss errors when CAD failed to mark the target. In the study by Drew et al. (2020), the miss rate for participants using interactive CAD did not increase when CAD failed to mark targets, even at low prevalence. Therefore, although interactive CAD only led to a modest improvement in detection rate at low prevalence, it eliminated one of the main disadvantages of CAD.

The nature of computer feedback has been investigated in a number of studies, in mammography (Giger et al., 2002; Hatton et al., 2004; Gilbert et al., 2008a; Hupse et al., 2013) and in the visual search literature (Cunningham et al., 2016; Kneusel and Mozer, 2017; Drew et al., 2020). A study by Drew and Williams (2017) investigated how providing simple feedback based on the eye tracking data of non-expert participants affected their performance and behaviour while searching natural scenes for simple targets (ovals and rectangles). Various types of feedback were examined over a number of experiments. Firstly, participants could press a button to reveal unfixated regions and visited regions, displayed by splitting the image into a 6 by 4 grid and highlighting the corresponding rectangles. Another feedback type was displayed automatically while the participants were viewing the images, initially overlaying the image with opaque grey rectangles, which became more transparent as they were fixated. The final feedback type involved displaying the 10 regions which they had visited least, after they had submitted a decision on the image.

With most feedback types, no reliable benefit in performance was found. An improvement in accuracy was observed for the feedback where the 10 regions least fixated were displayed post trial. This improvement was observed at 25% prevalence but not at 50%. This improvement at 25% was concluded to most likely be a false positive, since the feedback only highlighted 8% of the target locations, and therefore was unlikely to have been the reason that participants were able to improve their performance. The lack of any apparent benefit may be due to the nature of the images or the fact that the feedback was not linked with any information regarding target location. For example, systems in mammography that could combine the gaze data of expert readers with CAD output (abnormality location probabilities) and feed that back to the reader, such as those described by Chen and Gale (2010*a*), will likely have better success. Furthermore, the methods described in Section 5.4.1 may make better use of combining eye tracking data by combining it with AI algorithms.

5.4.3 Studies without eye tracking

Other methods of improving diagnostic performance and efficiency which do not involve eye tracking have been proposed. For example, to reduce the workload of radiologists, it has been investigated whether CAD could be used as a pre-screener (Astley et al., 2003). Here, a CAD algorithm sorted cases into either ‘possibly abnormal’ and ‘almost certainly normal’. Radiologists would then only review the abnormal cases and a small subset of normal cases. They found the CAD system only missed 3 of approximately 90 cancers, but over 70% of cases had CAD prompts. Since then, it has been argued that CAD should never be used as a pre-screener, because the standalone sensitivity of CAD would result in an unacceptable number of missed cancers (Philpotts, 2009). However, the study by Astley et al. (2003) is over 18 years old, and so pre-screening may be more feasible with present or future CAD systems.

In fact, a recent study by Raya-Povedano et al. (2021) explored triaging cases with AI to reduce the workload in both mammography and DBT using retrospective analysis of a previous study with 16,067 participants interpreted by experts. The impact of triaging cases was simulated using the AI risk score calculated for each case; low-score cases (<8/10) were not read, the rest were double read, and any cases not originally recalled by readers but in the top 2% of the most suspicious according to the AI were automatically recalled. Compared to double reading of DBT cases, this strategy resulted in a non-inferior sensitivity (increase of 3.2% for DBT and 2.6% for mammography), a reduction in the recall rate (16.7% for DBT and 16.9% for mammography), and a reduction in workload (72.5% for DBT and 71.5% for mammography). Earlier work in this group demonstrated that this method was capable of reducing workload without a change in reader AUC (Rodriguez-Ruiz et al., 2019c). These results are similar to another simulation study which used a deep learning method to triage cases, where readers would only review cases not deemed to be cancer free (Yala et al., 2019). Again, a reduction in workload was accompanied by an improvement in specificity without a reduction in sensitivity.

One method in which all mammograms would still be reviewed by at least one reader is where mammograms are sorted automatically into those which should be single-read and double-read (Balta et al., 2020). An AI score between 1 and 10 denoting the likelihood that a case contained a cancer was used to triage which cases would be single-read and which would be double-read. Different thresholds were evaluated to determine the optimal score to use for triaging, and it was found that single-reading for cases with scores between 1 and 7 produced the best results. Using this threshold, there was no change in cancer detection rate, a 0.56 percentage point decrease in recall rate, and a 32.6% decrease in workload. A similar methodology was used by Lång et al. (2021) to improve the detection rate of interval cancers in a screening cohort. By recalling all women with an AI score of >9 (approximately 10% of cohort), interval cancers were found to reduce by 19.3%.

Another potential improvement could be to follow a similar approach to airport baggage screeners, where fictional weapons are projected into luggage. This is used as a method of quality assurance for screeners to maintain and improve detection rate of real weapons. This could be applied to mammographic screening by inserting fictional patient data with known diagnoses into the normal work-flow of readers. However, only 31% members of the Association of University Radiologists in the US agreed that this was a good idea (Phelps et al., 2017), due to concerns with the increase in workload. Since a considerable number of fictional cases would need to be inserted to see a significant improvement, concerns over increased workload would be valid.

5.5 Summary

Eye tracking has long been used to study visual search behaviour, with the earliest devices developed in the late 19th century. Most modern devices are much less intrusive than their predecessors, making use of video-based technology to record the eye and experiment environment. Two main types of eye trackers exist: remote desktop trackers, which are typically used for screen-based experiments, and head-mounted trackers that have a variety of uses, from tracking gaze across large displays to experiments with participants moving in natural settings.

The two eye trackers that have been used in the experiments in later chapters are the desktop EyeLink1000 and the head-mounted glasses Pupil Core. To estimate gaze position, the EyeLink1000 uses dark pupil tracking with corneal reflections, and is able to achieve typical accuracies of $<0.5^\circ$ at 1000 Hz. The Pupil Core creates a 3D model of the eye, recording the eyes at 200 Hz and the experiment environment at 120 Hz, with typical accuracies of around 0.6°. Both trackers provide the gaze coordinates and fixations in their relative coordinate systems, and Pupil Core offers surface tracking to map that data to user-defined surfaces within the world video.

Important considerations should be taken into account when setting up an eye tracking device. Remote trackers should be positioned at the recommended distance from the participant, and the eye camera of any tracker should be positioned so the eye is centrally in view and focused. The calibration procedure is vital to achieving high accuracy and is often followed by a validation procedure to check it was followed correctly. Data cleansing for analysis will also improve the quality of data. This may involve setting appropriate fixation duration limits and removing those which fall outside of these thresholds or removing all fixations that are outside of the AOI.

Many studies have used eye tracking to gain insight into how readers search mammograms, and how their search is affected when the setup or nature of the images is altered. Eye tracking has been used to model the visual search of mammograms, highlighting that readers begin search with a global analysis before fixating regions that were identified as potentially abnormal or where the prior probability of a lesion is highest. Less experienced readers were shown to follow the slower search-to-find strategy, resulting in a higher number of errors. Due to the differences in search behaviour between readers of varying experience, it may be possible to train those of least experience by learning from the search of the most experienced readers. Gaze patterns have also been combined with CAD outputs to make feedback more informative for readers. Machine learning and deep learning methods have used eye tracking data to train models to predict image regions that will attract visual attention and the error types of missed cancers.

CAD prompts have been shown to affect the performance and visual search of non-expert participants reviewing images. A well-reported result in CAD literature is that the absence of a prompt on a target leads to a significant reduction in sensitivity for those targets, compared to unaided reading. Eye tracking demonstrated that displaying CAD prompts from image onset led to a significant reduction in image coverage, with participants focusing their attention on the prompts. Alternative prompt types and techniques such as subtle or interactive prompts have been demonstrated to have less of an impact on visual search, and lead to fixation patterns that more closely resemble unaided viewing. In non-experts, the cost of CAD missing targets increases at lower prevalence using traditional prompts, but is not seen for interactive CAD, with a small improvement in overall accuracy.

CAD algorithms may also be used earlier on in the screening process, triaging cases before they are read by radiologists. Early studies suggesting this idea were met with reservations on the basis that CAD systems were not accurate enough for this to be viable, but recent improvements report promising results. It may be possible for AI to decide which cases should be single read or not read at all (those with low levels of suspicion) and those which should be double read (those with high levels of suspicion). More unique concepts have also been explored, such as inserting fictional patients into the usual workflow of the radiologists to increase cancer prevalence, in a similar way as airport baggage screeners having fictional weapons projected into the luggage they are checking. These ideas, however, must take into account the increase in workload of an already over-worked profession if they were to be welcomed.

The next chapter explores the effect of using CAD as a second reader on visual search and behaviour. In the studies that have been discussed combining eye tracking and CAD, the prompts have been displayed from image onset, which is not the intended use of CAD for mammography. It is investigated whether the initial unaided viewing is affected by a secondary viewing with CAD – the so-called ‘safety-net’ effect.

Chapter 6

Safety-net effect with CAD

6.1 Introduction

In Section 3.3, we discussed how studies investigating the efficacy of CAD can be categorised as either longitudinal or cross-sectional. Cross-sectional studies are where readers sequentially review cases without and then with CAD, and longitudinal studies are where the cancer detection rate is measured before and after CAD is introduced into a clinic. As discussed in Section 3.3, cross-sectional studies are better suited for evaluating CAD systems (Nishikawa and Pesce, 2009). This is because longitudinal studies may not detect a change in cancer detection rate due to the prevalence of cancers changing after the introduction of CAD. Whereas, this is not an issue with cross-sectional studies since the same cases are read without and then with CAD.

The results of several cross-sectional studies are given in Table 6.1, where single reading with CAD had a higher sensitivity than single reading alone on average by 9.3%, and achieving a similar sensitivity compared to double reading. However, the recall rate is where there is a significant cost, with a 12.4% and 14.7% increase compared to single and double reading, respectively.

Most CAD systems are designed to be used as a second reader, where the reader first searches the mammogram unaided, and then reviews the image using the CAD output. In the majority of cross-sectional CAD efficacy studies, CAD is indeed operated sequentially this way: the reader searches a case initially unaided and gives their verdict (which is taken as the unprompted no-CAD decision) and then again with the aid of prompts (taken as the CAD decision), with the assumption that these two conditions are independent from each other.

Table 6.1: Comparison of single reading with CAD to single reading alone and double reading, in terms of sensitivity and recall rate. These are the percentage changes between no-CAD and CAD conditions, not percentage point differences. For the single reading versus single reading with CAD, it is the average and range of several studies.

Method	Single Reading with CAD	
	Sensitivity	Recall rate
Single reading[a]	+9.3% (0.0−19.5%)	+12.4% (6.3−25.8%)
Double reading[b]	–0.6%	+14.7%

[a] Average percentage increase in sensitivity and recall rate going from single reading to single reading with CAD (Freer and Ulissey, 2001; Helvie et al., 2004; Birdwell et al., 2005; Dean and Ilvento, 2006; Ko et al., 2006; Morton et al., 2006; Georgian-Smith et al., 2007).
[b] Change in sensitivity and recall rate going from double reading to single reading with CAD (Gilbert et al., 2008b).

Operating CAD in such a way may lead to a hypothesised safety-net effect (Astley and Gilbert, 2004), where the initial unaided search is adversely affected by the fact it is preliminary to a further search with the aid of CAD, and may therefore be less thorough than if CAD was not available. The aim of this study is to evaluate whether there is evidence for the existence of such an effect. If so, it may be that clinical studies that implement CAD as a second reader to evaluate its efficacy may overestimate its benefit, since the performance in no-CAD conditions will be underestimated. The results discussed in this chapter have been published in Du-Crow et al. (2019).

6.2 Methods

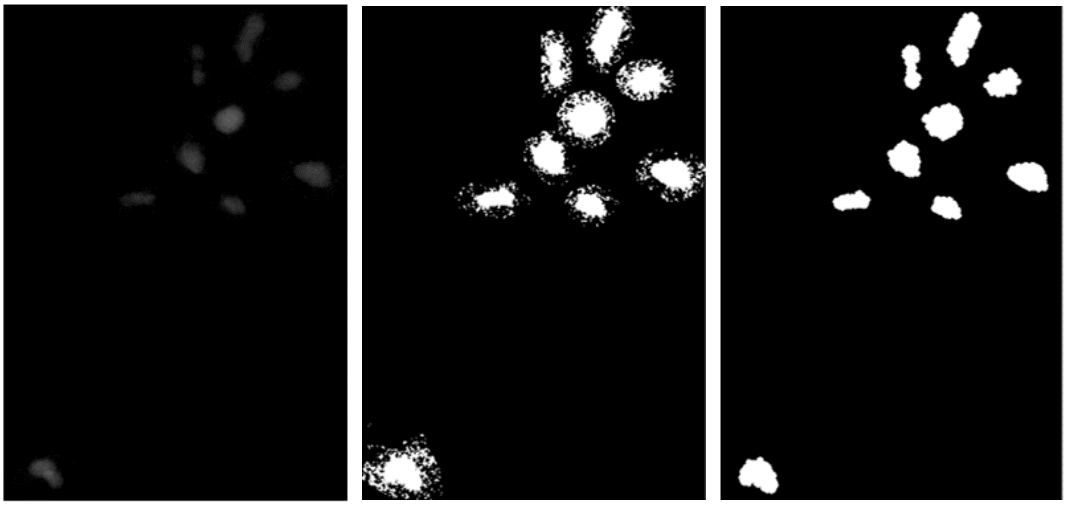
This experiment was a visual search task with non-expert participants searching for microcalcification clusters in 1/f noise distributions. Each participant searched 100 images in a no-CAD condition and 100 in a CAD condition, with the order varied amongst participants. Eye-movements were tracked throughout the experiment. The prompts used in this study were manually overlaid on the images and were not generated by a CAD algorithm. The operating point was chosen to replicate CAD in mammography, discussed further in Section 6.2.3.

Fifty-two participants (median age 21, age range 18-59, 37 female) were recruited for the study and informed consent was obtained. Four participants' eye-movements could not be accurately calibrated so only behavioural data was collected. Sixteen participants were undergraduate psychology students and received course credit for taking part, while the rest (students, university staff, and members of the public) received £10 in exchange for their time. The study was approved by the University of Manchester Research Ethics Committee (2018-4586-6410).

6.2.1 Stimuli

The images used in this study comprised of synthetic mammogram-like backgrounds combined with malignant microcalcification clusters used as targets. Clusters were extracted from magnified images of slices of mastectomies (Warren et al., 2012). We were provided with a total of 117 of these clusters, one example is shown in Figure 6.1a. Figure 6.1b shows a binary view of the cluster, to give a clearer picture of the shape of the individual calcifications. The cluster ROIs, defined as the bounding box around the edge of the calcifications, ranged from 163×163 pixels to 873×903 pixels.

Two sets of clusters were formed by matching the cluster properties between the sets. The following properties were used: mean cluster pixel intensity, number of calcifications in cluster, total area of three largest calcifications, mean calcification area in cluster, and maximum calcification area. However, before these could be calculated, we first dilated and eroded the clusters in Matlab using the *imopen* function with a disk-shaped structuring element with a radius of 3 pixels. The result of this process is shown in Figure 6.1c.



(a) Example cluster (b) Binary cluster (c) Smoothed cluster

Figure 6.1: Example of a microcalcification cluster. Cluster in its (a) original form, (b) as a binary image to highlight the full details of the calcifications, and (c) smoothed using Matlab's *imopen* function for cluster property analysis.

Once clusters were prepared for analysis, 80 were randomly chosen from the 117 and split into two groups of 40 (set A and B), and the cluster property distributions were measured for each set. This process was repeated until their properties were matched, as determined by a Kolmogorov-Smirnov test. The results of this are given in the first five rows of Table 6.2. Once the two sets of clusters were established, they could be inserted into backgrounds. The backgrounds were created using open-source code in Matlab, described as a fractal surface generator (Methven and Qi, 2018). They were $1/f^{1.5}$ noise distributions, which resemble the glandular component of mammograms but lack linear structures. One hundred backgrounds were generated initially (set A) and then rotated 180 degrees to form a second set of 100 (set B). Forty images were randomly chosen from image set A to be used for cluster insertion, along with the corresponding rotated images from image set B, giving a target prevalence of 40%. Participants viewed both sets of images, one with CAD and one without.

Table 6.2: Results from Kolmogorov–Smirnov test for splitting calcification clusters into two sets. H is the test decision for the null hypothesis that the distributions of properties in the two cluster sets are from same continuous distribution, where H=1 rejects the null hypothesis at the 5% significance level. The condition for the samples to be used was $p > 0.99$ for the first 5 properties, measured for clusters not yet inserted into backgrounds. The final two properties were measured once the clusters has been inserted into backgrounds to ensure they were similar.

Cluster property (not inserted)	**H**	**p**
Mean cluster pixel intensity	0	0.9998
Number of calcifications in cluster	0	0.9998
Total area of 3 largest calcifications	0	0.9998
Mean calcification area in cluster	0	0.9998
Maximum calcification area	0	0.9998
Cluster property (inserted)	**H**	**p**
Cluster region image entropy	0	0.9239
Cluster-background contrast	0	0.8403

For displaying the images on the experiment computer, they were displayed at 800×800 pixels. However, initially, backgrounds were created as 7500×7500 pixels in size. The clusters were inserted by multiplying the background image ROI with the cluster ROI, and by not initially resizing the clusters, this process resulted in clusters that were blended better with the background. Each cluster was placed at a random point within a 5×5 grid on the background. Given that there were large variations in the sizes of the clusters, the grid was programmed in such a way that the entirety of the cluster would remain inside the image if it was placed at the edge of the grid.

Once the position was chosen, the cluster pixels were multiplied by the background region pixels to insert it into the images of set A or B. The image was then resized to 800×800 pixels. Finally, the cluster region image entropy and cluster-background mean pixel contrast within the cluster region were compared to confirm the similarity between the target sets (see the bottom two rows of Table 6.2). To avoid the predictability of target position within the sets of clusters, we used each of the possible 25 grid positions once in each set and 15 were used twice. An example of one of the images with a cluster inserted, highlighted in yellow, is shown in Figure 6.2.

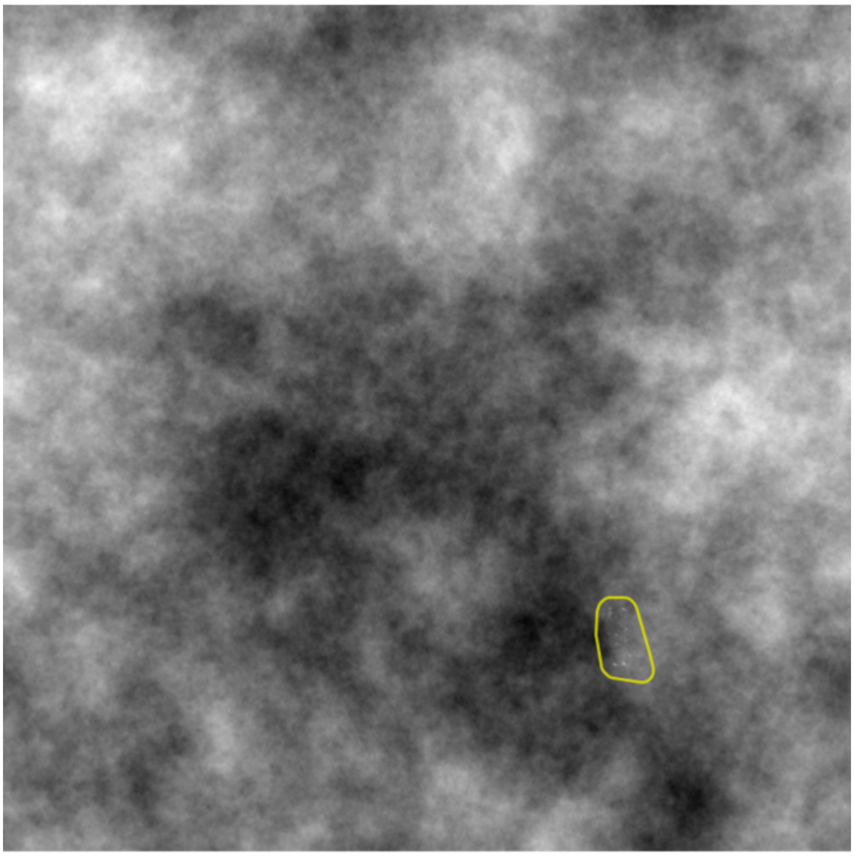


Figure 6.2: Example of experiment image with microcalcification cluster inserted (highlighted in yellow).

6.2.2 Experimental setup

The experimental setup is shown in Figure 6.3. Participants' head movements were restrained with a chin rest, which was adjustable in height, at a viewing distance of 73cm. The images were displayed on a ViewSonic VX2268WM LCD monitor with a resolution of 1680×1050, visual angle 36° × 23° at a viewing distance of 73cm. Eye movements were tracked with an EyeLink1000 desktop eye tracker, described in Section 5.3.1. The experiment code was written using PyGaze (v0.6.0) (Dalmaijer et al., 2014), with a PyGame back-end (v1.9.2).

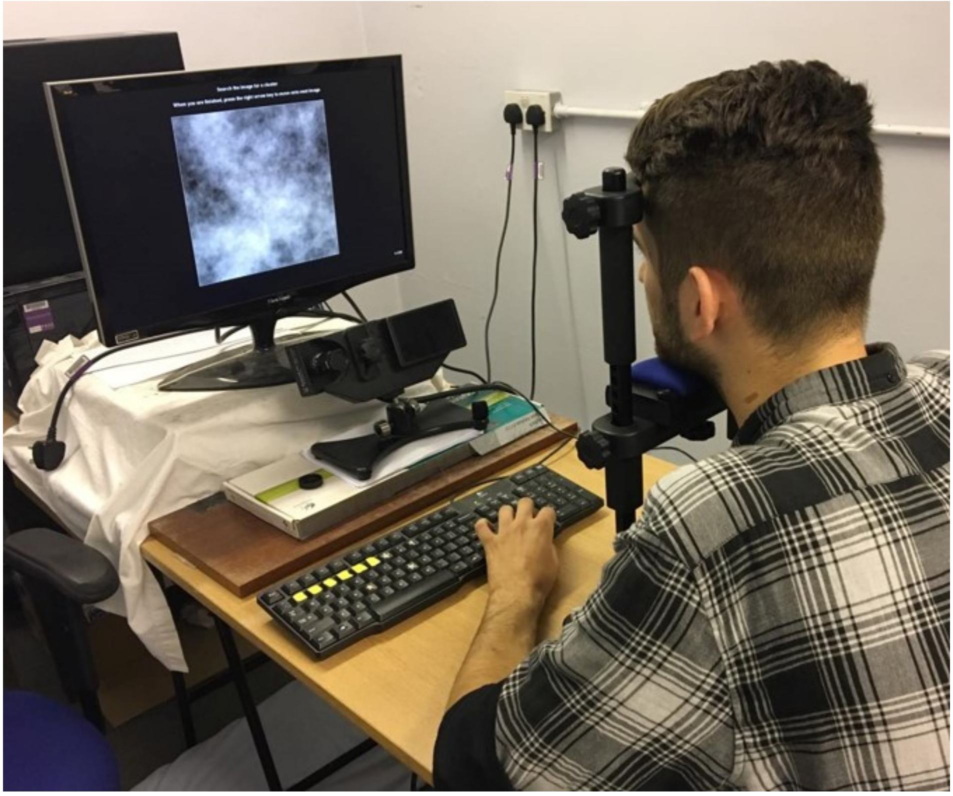
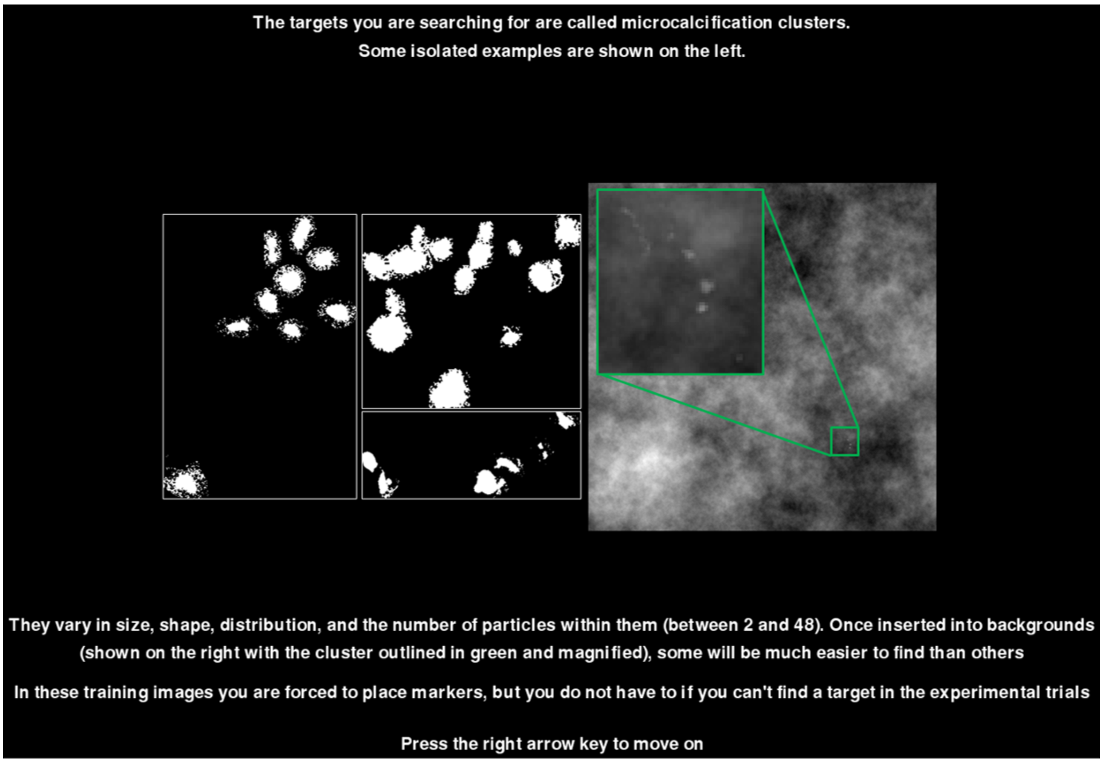


Figure 6.3: Photo of experimental setup. Participants' head movements are restrained with a chin rest. When the experiment was taking place, the lights were turned off.

6.2.3 Experimental procedure

An inclusion criterion of the experiment was participants must have normal or corrected to normal vision (20/20 or higher). We tested this using the Freiburg Vision Test (FrACT) (Bach, 1996). This was done with a visual acuity test and a contrast sensitivity test. The eye tracker was able to accommodate participants wearing glasses and contact lenses.

A training set of images was used to familiarise the participants with the experimental procedure, the appearance of the images and targets, and how to operate CAD. The initial training screen, Figure 6.4, was used to introduce the participants to micro-calcification clusters. It was noted on this screen that in the training, they were forced to click somewhere on the image as a way of making them engage with the training set. However, it was made clear that in the experiment set, if they did not find a target, they did not have to place a marker on the image.



The targets you are searching for are called microcalcification clusters.
Some isolated examples are shown on the left.

They vary in size, shape, distribution, and the number of particles within them (between 2 and 48). Once inserted into backgrounds (shown on the right with the cluster outlined in green and magnified), some will be much easier to find than others

In these training images you are forced to place markers, but you do not have to if you can't find a target in the experimental trials

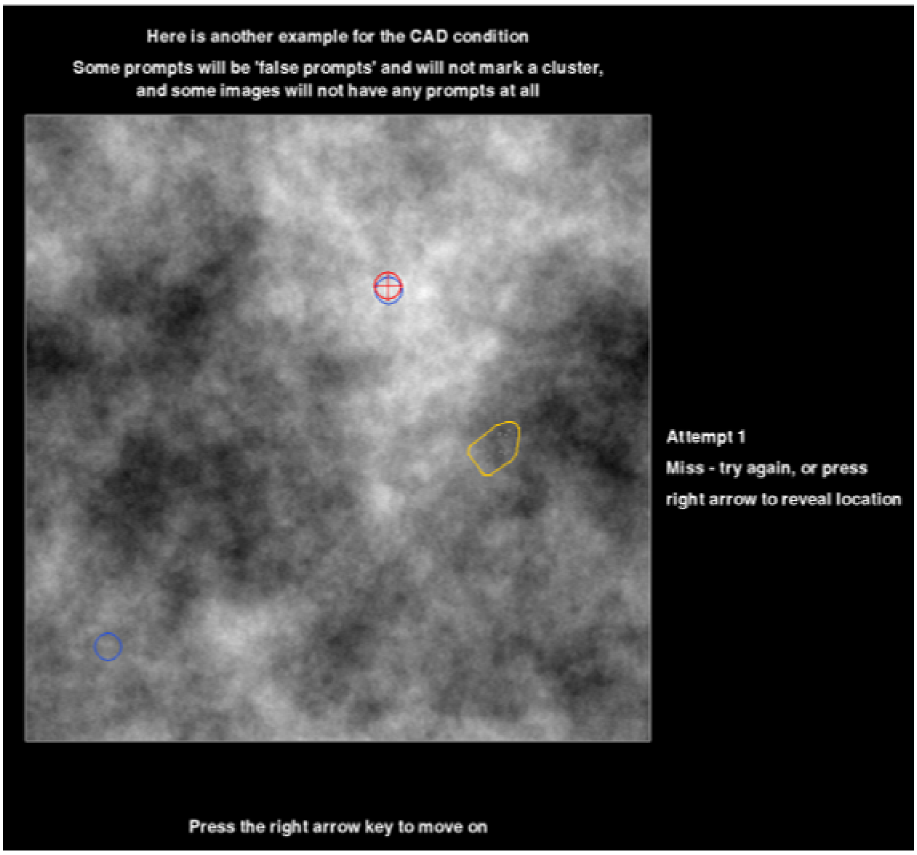
Press the right arrow key to move on

Figure 6.4: Training screen to introduce participants to microcalcification clusters as targets in the experiment. The screen features three example clusters and a different cluster inserted into a background, with a magnified view of the cluster ROI.

Following the initial training screen, the participants were shown a total of 10 training images. First, two images were displayed with a cluster outlined for participants to gain further familiarity with the targets.

Participants then began the interactive training images. The first four of these did not contain CAD prompts and just explained to participants how to interact with the images by placing markers on suspicious regions and removing markers if they changed their minds. When they finished placing markers, they pressed the right arrow key to move on. Each time they clicked in an incorrect location, they would receive a text prompt to inform them that this was an incorrect attempt (see Figure 6.5). If they correctly identified a target, they would receive a text prompt to say it was a ‘Hit’, and the cluster region would be outlined in orange. One of the no-CAD examples did not contain a target and contained a text prompt that overlaid the image, once they pressed the move on button, to inform them that fewer than 50% of images in the experiment contain a target.

The training also included three examples with CAD that explained how to properly operate it in second reader mode: search the image initially unaided, then press the up arrow key to display the prompts, and search the image again with the aid of prompts. One of the CAD examples is shown in Figure 6.5. The final training image featured a pre-placed marker with an instruction on how to remove it, to make sure participants were capable of removing markers they had placed if they needed to.



Here is another example for the CAD condition

Some prompts will be 'false prompts' and will not mark a cluster, and some images will not have any prompts at all

Attempt 1

Miss - try again, or press right arrow to reveal location

Press the right arrow key to move on

Figure 6.5: Screenshot of interactive training with CAD. This example contains two false CAD prompts (blue circles) and a user placed marker (red circle with cross hair), which has triggered the message on the right hand side. The user has pressed the right arrow key to reveal the target location (orange outline).

Once the training set had been completed, participants were given the opportunity to ask the researcher any questions they may have had before starting the experiment. All participants were given final instructions about the experiment to reiterate the key points. These included:

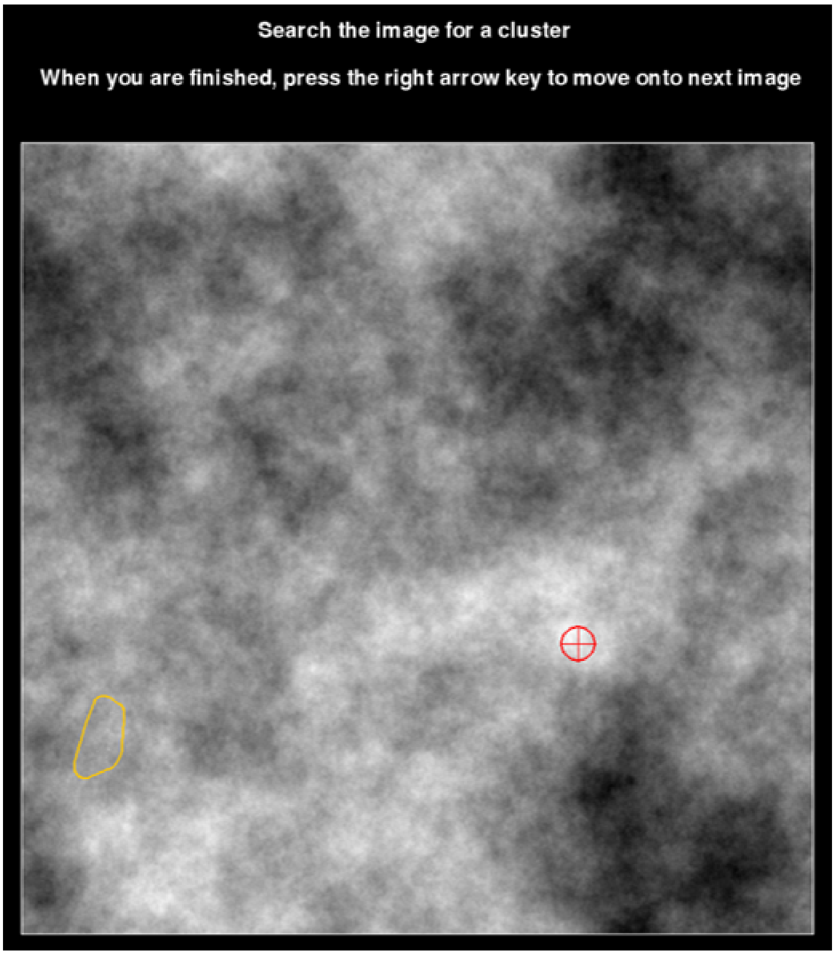
- Images contain no more than one cluster and between 1% and 50% of the images contain a cluster, so you do not have to place a marker if you cannot find one
- You can remove markers that you have placed
- In the prompted condition, there may be 0, 1 or 2 prompts

- You can toggle prompts on and off with the up arrow key
- You will not get feedback in the actual experiment images
- You do not have to place multiple markers on a target, one on any part of it will count as a hit

The experiment screen in the no-CAD condition (Figure 6.6) featured two instructions that remained visible throughout: search for a target and move on to the next image when you are finished. The experiment screen in the CAD condition (Figure 6.7) started with two instructions: search for a target and then turn on the prompts. Once the participants turned on the prompts, they received one of two further instructions below the image: either search again with the aid of the prompts or that there were no prompts available for this image.

The order of the 200 experiment images was randomised for each participant. We blocked our participants by CAD condition, half started with the no-CAD condition and the other with the CAD condition. For the no-CAD and CAD groups, half of the participants had images drawn from set A and half had images drawn from set B. We randomly allocated participants to these groups. No time limit was enforced during the experiment, but we advised participants not to spend longer than around 1 minute on any single image. A break screen would appear after every 25 trials during the experiment.

Participants completed a nine-point calibration and validation procedure at the beginning of each condition. This was repeated if accuracy of fixations exceeded 1 degree visual angle on any point in the validation sequence. The calibration procedure could also be repeated at any point during the experiment if calibration accuracy began to drift, detected by a ‘drift correction’ procedure where the participant looks at a central fixation point and the experimenter presses a button to measure the displacement from



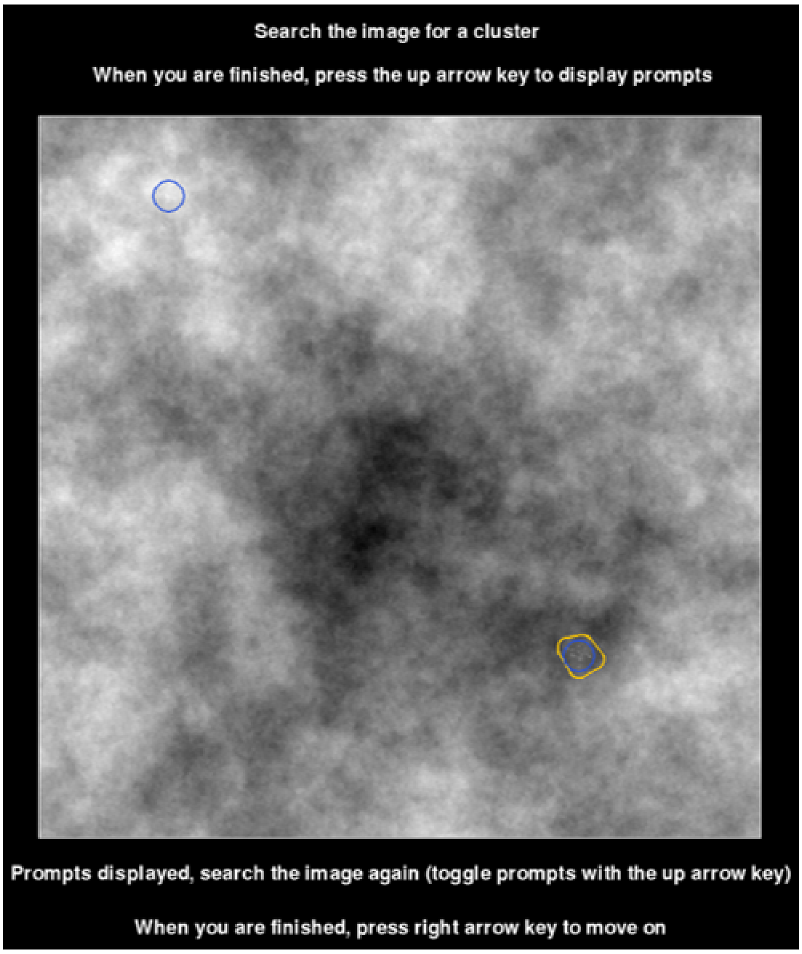
Search the image for a cluster

When you are finished, press the right arrow key to move onto next image

Figure 6.6: Screenshot of experiment in the no-CAD condition. The cluster is outlined (in orange) for illustrative purposes only. There is a user placed marker (red circle with cross-hair) on a background region. There was a progress counter in the bottom right corner to show image number out of 100.

that point. The drift correction was repeated mid-way between and immediately after the break screens. If an accurate calibration could not be achieved, the experiment was run without the eye tracking and behavioural data only was collected. For classification of fixations, the maximum drift in eye position coordinates between start point of fixation and end point of fixation should be less than 1.5 degrees visual angle, and the minimum fixation duration was set at 100 ms.

The target prevalence was 40%, and in the CAD condition 80% of the targets were prompted by CAD. From each of the image sets, we selected 8 cluster targets that would be unprompted. We matched the properties (listed in Table 6.2) using a Kolmogorov-Smirnov test with all p's>0.92. CAD prompts were manually placed on the images, with true prompts placed on the most central part of the cluster and false prompts typically placed on bright regions of the background. The false positive prompt rate was 0.5 per image, with false prompts placed in the same positions between set A and B, but rotated 180 degrees for set B to match the initial rotation of set B images. Out of the 8 unprompted targets in each image set, four contained false positive prompts.



Search the image for a cluster

When you are finished, press the up arrow key to display prompts

Prompts displayed, search the image again (toggle prompts with the up arrow key)

When you are finished, press right arrow key to move on

Figure 6.7: Screenshot of experiment in the CAD condition. There are two prompts (blue circles), a false prompt and a true prompt. The cluster is outlined (in orange) for illustrative purposes only. Before the CAD prompts were displayed there was no text below the image. If no prompts were available, the first line of text below the image would instead read 'No prompts to display'. There was a progress counter in the bottom right corner to show image number out of 100.

6.2.4 Analysis

To determine whether users successfully marked targets, we took the convex hull of the calcifications in each of the clusters and then extended this outwards by a tolerance of 10 pixels to create a border around the target (see Figure 6.2 for an example). Any marker placed within this was recorded as a hit (maximum of one hit per target/image). A preliminary test with 5 participants measured the accuracy of their clicking. These participants viewed 25 images with a grey background and a single white dot (5 pixel diameter) and clicked as closely as possible to the dot. Average distance from click and target dot was 2.7 pixels. Therefore, a 10 pixel border around clusters was assumed to be sufficient to capture all participant clicks on targets.

Sensitivity was defined as the fraction of targets correctly located. Participants were not limited by the number of markers they could place on an image, and false positive responses per image indicate the mean number of incorrect observer-placed marks per image.

We used a circle of radius 2.5 degrees visual angle centred around fixations, typically associated with the useful field-of-view in mammography (Kundel and Nodine, 2004), for calculating the percentage of the image fixated (image coverage). Any location inside that region was considered covered. For calculating the percentage of targets that were fixated, we used a function that detects the collisions between rectangles (bounding boxes around clusters) and circles (fixations). A similar process was done for the percentage of prompts fixated but with collisions between circles.

To determine whether the differences between conditions for trial time, image coverage, sensitivity, false positives per image, and percentage of targets and prompts fixated were significant, we used a bootstrap technique across participants and images (Efron and Tibshirani, 1993). We were then able to calculate the t-statistic and corresponding p-value, and the 95% confidence intervals on the differences. This treats participants and images as random rather than fixed effects. This method produces a more conservative estimate of an effect than a paired t-test.

To illustrate this approach, we can focus on a single variable as an example. We recorded the time spent viewing each image and wished to compare that between the no-CAD and CAD condition. We have N participants, each with M values for each condition, where M is the number of images per condition (100). We repeat the following process B times, where B=10,000.

1. Choose N readers with replacement, where N is the number of participants in our study
2. Choose M images with replacement, where M is the number of images in each condition
3. Calculate the differences in trial time between the no-CAD and CAD conditions within those participants for the chosen images
4. Calculate the mean within-participant difference

We then take the mean of the distribution of B differences to get the final mean difference between conditions. The distribution of B differences is used to calculate the 95% CIs. The t-value can be calculated by dividing the mean difference by the standard deviation of the distribution of B differences. Using the degrees of freedom (N-1), the p-value is calculated from the two-tailed t-distribution. It is important to note that M is not necessarily equal to 100. For instance, when comparing the sensitivity between conditions, we are of course only considering the 40 target present images in each condition. This is further reduced to 32 and 8 images when comparing prompted and unprompted targets, respectively.

6.3 Results

Results are given for the no-CAD condition and the CAD condition. We split the CAD condition further into the pre-CAD and the with-CAD conditions to distinguish between visual search before and after participants enabled prompts on the images. The error bars shown in the figures in this section are the within-subject confidence intervals, described by Cousineau (2005), which illustrate the variation between participants and conditions. However, these were not used for significance testing.

6.3.1 Eye tracking data

To illustrate the eye movement parameters we were able to extract for each experiment trial, we have displayed the fixations and saccades for a single participant in Figure 6.8. This was an image in the no-CAD condition, so no prompts were available. This participant started in the centre of the image before adopting a scanning technique back-and-forth across the image until they fixated the target cluster in the bottom right of the image. Overall, they spent 17.40s viewing the image, made 45 fixations, and it took them 9.00s to fixate the target. In total, their dwell time on the target was 5.69s, and they placed a marker on the target after 15.56s. The total image coverage was 84.1%.

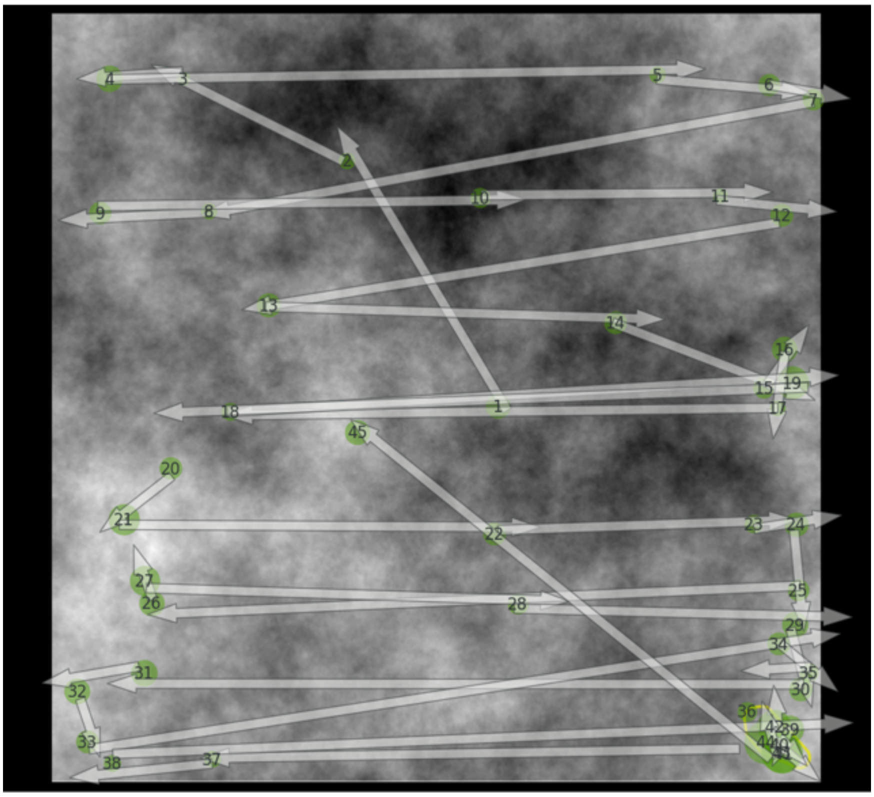
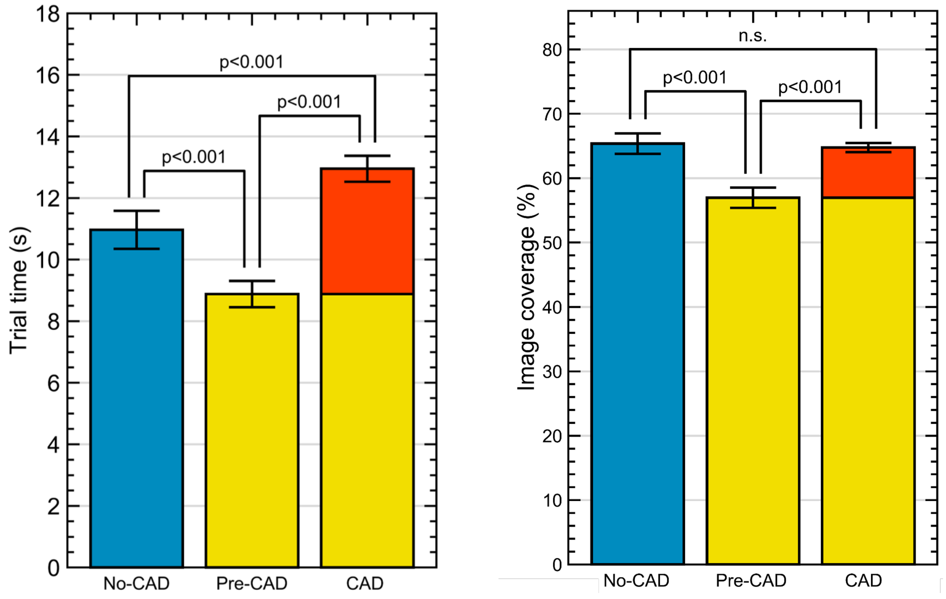


Figure 6.8: Eye movement data from a single participant on an image showing their fixations (green circles) and saccades (arrows). The cluster is in the bottom right corner. Fixations are numbered from 1 to 45 from where they started and finished their search on the image. The size of the fixation circles indicate the fixation duration – the longer the fixation, the larger the circle.

6.3.2 Trial time and coverage

Trial time and coverage was measured for each image, and the mean across participants is shown for each condition in Figure 6.9. Pre-CAD trial time was significantly lower than for the no-CAD condition by 2.09s (95% CI, 1.11−3.09s; t(51)=4.14, $p<0.001$). This was accompanied by a significant decrease in image coverage to 57.0% in pre-CAD from 65.4% in the no-CAD condition, a difference of 8.4% (95% CI, 5.4−11.7%; t(47)=5.29, $p<0.001$). The addition of CAD, i.e. pre-CAD + with-CAD (8.88s + 4.07s), led to a significant overall increase of 1.98s in trial time compared to no-CAD (95% CI, 0.99−3.05s; t(51)=3.82, $p<0.001$). Despite the greater trial time in the CAD condition, it was not reflected in the percentage coverage of the image, with a non-significant difference compared to no-CAD of 0.6% (95% CI, –1.4–2.7%; t(47)=0.58, p=0.568).

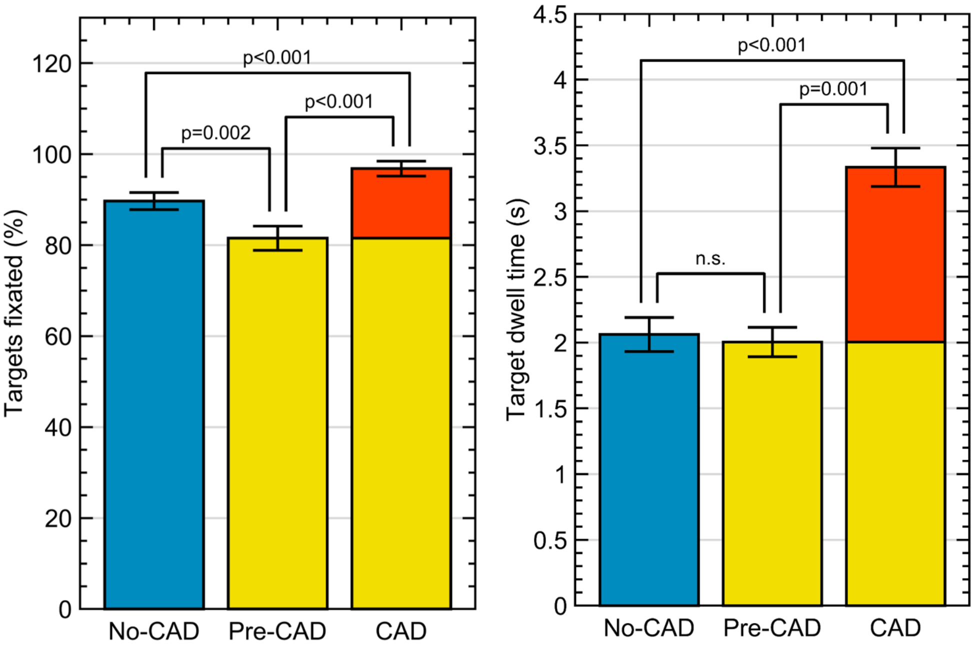


(a) (b)

Figure 6.9: (a) mean trial time and (b) mean image coverage for each condition. The red portion of the CAD bars is the additional trial time and coverage gained after participants turned on the prompts.

6.3.3 Targets and prompts fixated

Figure 6.10 shows the mean percentage of targets fixated and dwell time on targets for each condition. The percentage of targets fixated varied between conditions, with significantly more targets fixated in CAD (96.8%) compared to no-CAD (89.7%) as expected, a difference of 7.1% (95% CI, 3.8–10.8%; t(47)=3.96, p<0.001). In the pre-CAD condition, only 81.5% of targets were fixated, 8.2% less than in the no-CAD condition, which was a significant reduction (95% CI, 3.3–13.3%; t(47)=3.23, p=0.002). The mean dwell time on the targets was shorter in the with-CAD condition than in no-CAD (1.33s with-CAD vs 2.06 no-CAD). This difference of 0.73s was significant (95% CI, 0.39−1.06s; t(47)=4.26, p<0.001).



(a) (b)

Figure 6.10: (a) mean targets fixated percentage and (b) mean target dwell time for each condition. The red portion of the CAD bars is the additional targets fixated and target dwell time gained after participants turned on the prompts.

In the with-CAD condition, 76.2% of true prompts were fixated, a significantly lower fraction than the 95.8% of false prompts fixated (difference of 19.5%; 95% CI, 15.6−23.5%; t(47)=9.735, p<0.001) and participants spent significantly longer viewing individual false prompts than individual true prompts (2.57s vs 1.89s). This difference in dwelltime of 0.69s was significant (95% CI, 0.35−1.01s; t(47)=4.103, p<0.001). There were a total of 32 true positive prompts and 50 false positive prompts in the CAD condition, and the proportion of these prompts that were acted on by participants (marked as a target) is shown in Figure 6.11. On average, participants acted on (34.0 ± 28.5)% of true prompts and (25.1 ± 21.8)% of false prompts. The rate that participants acted on true prompts will have been in part reduced by the fact that, on average, 55% of true prompt regions were already marked in the pre-CAD condition. Therefore, in those situations they did not have to mark those true CAD prompts again.

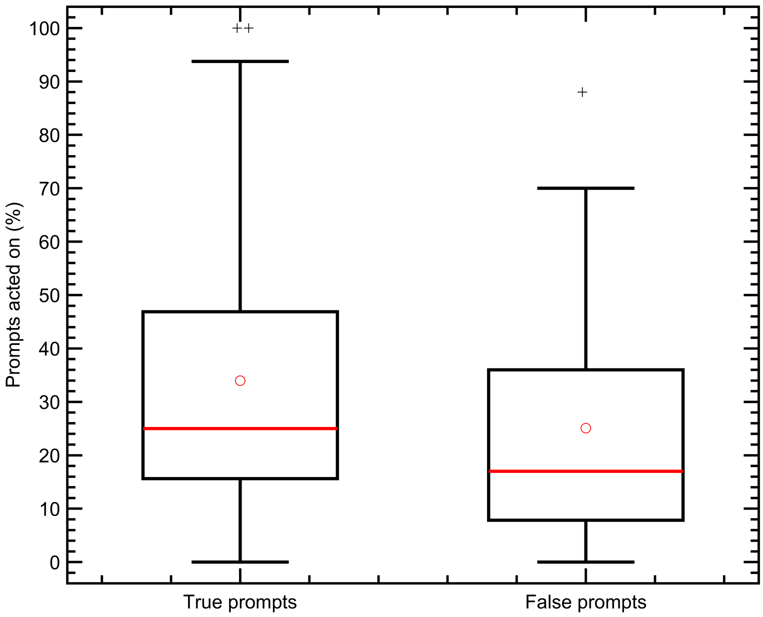


Figure 6.11: Percentage of true and false prompts marked as a target by participants in the CAD condition. The mean of each distribution is displayed as a red circle.

6.3.4 Observer performance

The mean observer sensitivity is shown in Figure 6.12 for all images and those with and without targets prompted by CAD. There were 32 targets in each image set that were prompted by CAD and 8 in each that were not prompted by CAD, also referred to as prompted and unprompted targets, respectively. Overall observer sensitivity significantly increased by 15.8% to 81.9% in the CAD condition from 66.1% in no-CAD (95% CI, 10.0−22.0%; t(51)=5.16, p<0.001), and by 20.4% for prompted target trials to 88.2% in CAD from 67.8% in no-CAD (95% CI, 14.5−26.8%; t(51)=6.56, p<0.001). However, there was no significant difference in sensitivity between no-CAD and CAD for unprompted target trials, a difference of 2.9% (95% CI, −7.2−14.2%; t(51)=0.54, p=0.59). Sensitivity in the pre-CAD condition was significantly lower than the no-CAD and CAD conditions across all trial types; in all cases t(51)>2.62, p<0.012.

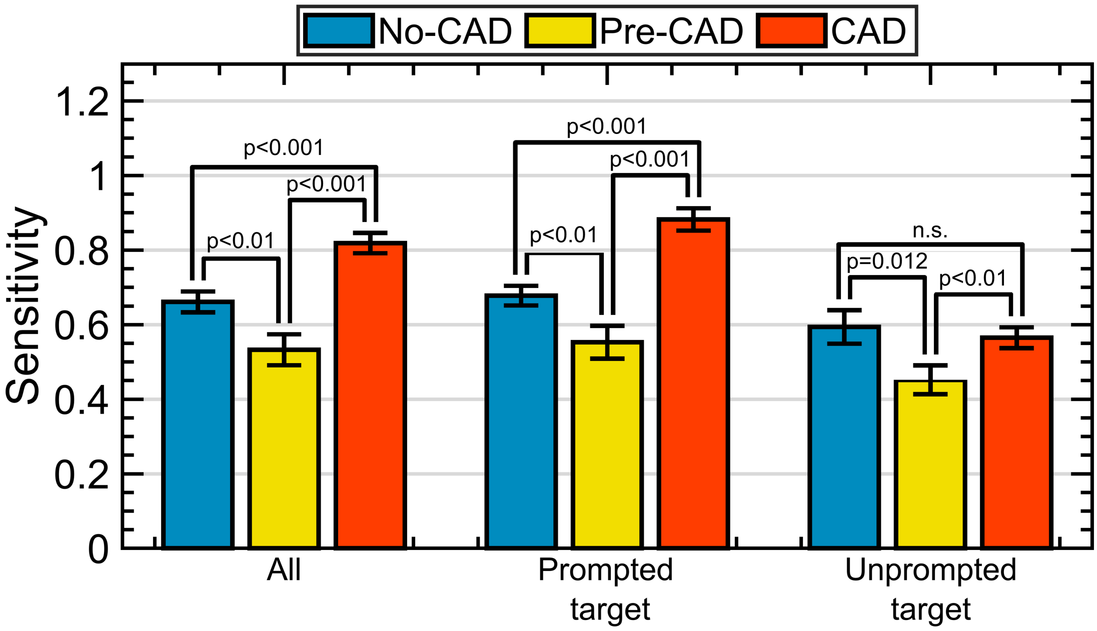


Figure 6.12: Mean observer sensitivity for all target present trials, and for trials with prompted and unprompted targets.

The mean rate of false positive responses per image are shown in Figure 6.13, given for all trials and for target absent trials with and without false prompts. There were a total of 60 target absent trials in each condition, and in the CAD condition 30 contained false positive prompts and 30 did not. There was a significant increase in the number of false positive responses per image from 0.28 in the no-CAD condition to 0.35 in the CAD condition for all trials, an increase of 0.07 (95% CI, 0.01−0.13; t(51)=2.27, p=0.028), and for target absent trials with false prompts up to 0.64 in CAD from 0.36 for those same images in no-CAD, a significant difference of 0.26 (95% CI, 0.17−0.40; t(51)=4.66, p<0.001). For target absent trials with no prompts, there was no significant difference between the no-CAD and CAD conditions, a difference of 0.02 (95% CI, –0.04−0.09; t(51)=0.67, p=0.51). False positive responses per image for the pre-CAD condition were significantly lower than the no-CAD and CAD conditions across all trial types; in all cases t(51)>2.16, all p<0.035.

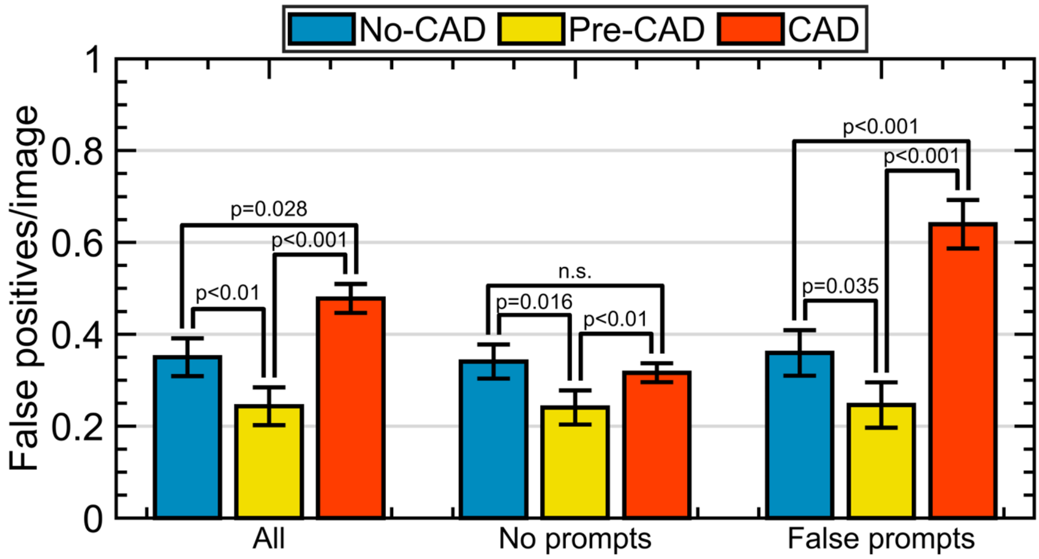


Figure 6.13: Mean number of false positive responses per image for all trials, and for all target absent trials without false prompts and with false prompts. For the ‘No prompts’ and ‘False prompts’ groups, we used the same corresponding set of images when comparing to the no-CAD condition.

6.3.5 Target detectability

Target detection percentage is defined as the percentage of participants that successfully locate a given target, shown in Figure 6.14 for no-CAD and CAD conditions. As expected, the majority of targets had a higher detection rate in the CAD condition, since 80% were prompted by CAD. For the targets that were not prompted in the CAD condition, overall, there was no significant difference in the detection rates between the conditions (p=0.50, Wilcoxon signed-rank test), with the majority of points falling on or near to the line of equality. There were 3 (out of 8) unprompted targets that had a much lower detectability in the CAD condition compared to no-CAD. There were no significant differences in the detectability of unprompted targets in images with and without false CAD prompts (p=0.50, Wilcoxon signed-rank test).

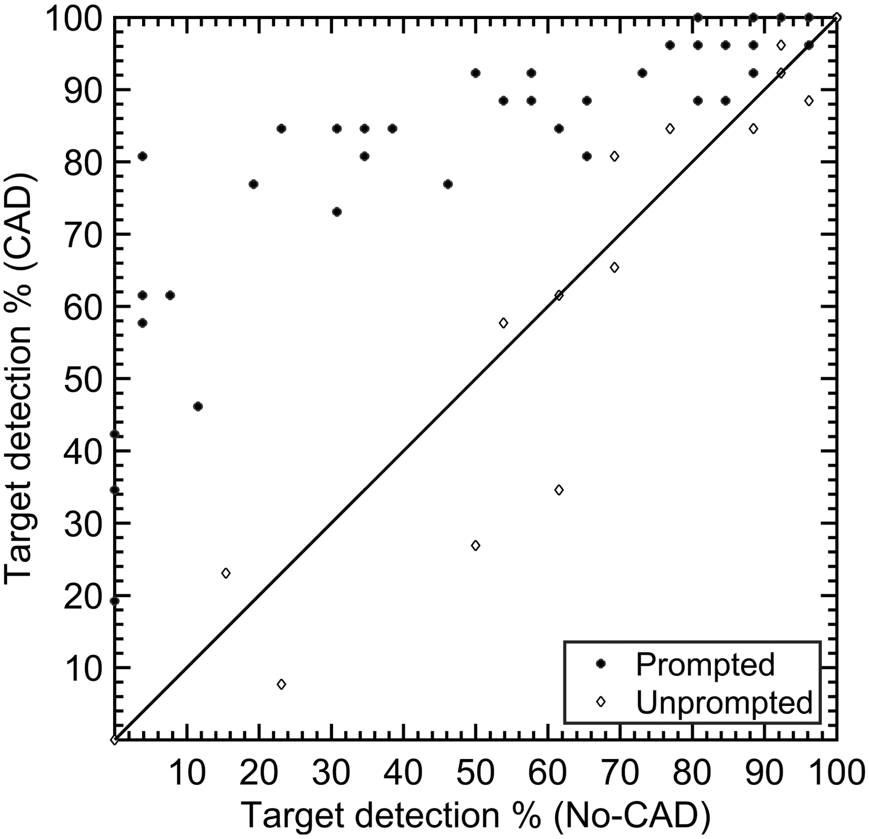


Figure 6.14: Percentage of participants that detected each target in the no-CAD and CAD conditions. It is indicated whether a target was prompted or unprompted. There were 2 targets that were not detected in either condition by a single participant (unprompted in CAD condition) and 8 targets that were detected by all participants in both conditions (all prompted in CAD condition). A line of equality is displayed for comparison purposes.

6.4 Discussion

Cross-sectional studies typically evaluate the efficacy of CAD in second reader mode, where readers search initially unaided and then once again with the aid of CAD. The initial search is taken as the no-CAD condition, but it was hypothesised (Astley and Gilbert, 2004) that the fact this initial search was preliminary to a further search with CAD means that it may become truncated. This was observed both in terms of less time spent viewing the images and a lower percentage image coverage for pre-CAD search compared to unaided search. As with the clinical studies (see Table 6.1), CAD improved detection rate at the cost of increased false positive responses.

Viewing time in the pre-CAD condition was reduced by 19% compared to the no-CAD condition, with a corresponding 8.4 percentage points reduction in image coverage. In addition, 8.2% fewer targets were fixated pre-CAD compared to the unprompted condition. A previous eye tracking study with non-experts has demonstrated that presenting CAD from image onset leads to a significant reduction in image coverage compared to unaided viewing (Drew et al., 2012). We have extended this to second reader CAD, demonstrating that the expectation of prompts is enough to truncate the initial search of the image. Our main focus is on cross-sectional studies, where the assumption is that pre-CAD search is similar to unaided search, since this is the most common trial type for evaluating CAD. However, our results suggest that this assumption does not hold, and it is not a valid comparison between unprompted and prompted behaviour.

To control for the safety-net effect, it may be appropriate to adopt an alternative study design. A fully-crossed design, such as the one used in this study, provides entirely separate no-CAD and CAD conditions for each reader and therefore any improvements in CAD can be measured between those. However, it should be noted that each reader completes different target sets for the no-CAD and CAD condition, with the same backgrounds used between sets. We matched image properties between conditions and with enough images and by bootstrapping the analysis, the variation across images is taken into account. Another alternative is a randomised control trial with separate CAD and no-CAD groups. Instead of image variability however, this may create an issue with reader variability between groups, which is known to be a factor in mammography interpretation (Elmore et al., 2009).

The increase in sensitivity from the no-CAD condition to the CAD condition was 15.8 percentage points, which is consistent with literature for both experts (Samulski et al., 2010; Alberdi et al., 2004; Taplin et al., 2006) and non-experts (Drew et al., 2012; Russell and Kunar, 2012; Kunar et al., 2017; Ionescu et al., 2018). These studies reported that for unprompted targets, sensitivity was significantly reduced. Out of these studies, only Taplin et al. (2006) operated CAD in second reader mode, reporting an approximate 7 percentage point drop in sensitivity for unprompted targets compared to unaided viewing. In our study, there was no difference in observer sensitivity between the no-CAD and CAD condition for unprompted targets. This discrepancy may be due to the relatively small number of unprompted targets (8 out of 40) in each image set, reducing the statistical power of the finding.

A study by Ionescu et al. (2018) reported that on unprompted target images, the presence of false prompts caused a significant reduction in detection rate compared to images without false prompts, when prompts where displayed from image onset. However, by operating CAD in second reader mode instead, readers are able to search the image initially without prompts present. This may reduce or eliminate the distraction of false prompts. This is indeed what we observed, with no difference between unprompted target sensitivity for images with and without false prompts present. However, it is again possible this finding was underpowered. Furthermore, while Drew et al. (2012) reported that CAD focuses visual attention around prompts, this was not observed in our study since readers viewed the images initially unaided in the CAD condition. With an initial unaided search, no significant difference was observed between the no-CAD and CAD image coverage.

In mammographic screening, the cancer prevalence is typically less than 1% (Evans et al., 2013), whereas the target prevalence used in this study was 40%. Gur et al. (2003) observed that target prevalence did not affect observers' AUC. However, miss rate of cancers has been shown to increase under low prevalence conditions (Evans et al., 2013). Furthermore, previous studies have shown that at low prevalence ($\leqslant$10%), the miss rate of unprompted targets is significantly increased compared to at high prevalence (50%) (Kunar et al., 2017; Drew et al., 2020). It is possible, therefore, that we may have observed similar effects as reported here had the prevalence been lower. In clinical settings where prevalence is $<$1% and readers are under greater pressure, it may be expected that initial pre-CAD search is truncated further, but this has yet to be demonstrated. The signal that observers get from CAD will also be affected by the target prevalence (Russell and Kunar, 2012), since a greater number of targets means that a higher proportion of the prompts that participants see are true positives, and therefore they will likely have a greater confidence in the CAD system. In turn, confidence in false prompts will also be greater and may increase false response rate. However, Kunar et al. (2017) reported no effect of target prevalence on the false response rate for images containing incorrectly placed CAD prompts.

While sensitivity was lower in the pre-CAD condition compared to the no-CAD condition, it is not clear whether this is due to the safety-net effect. Participants fixated significantly fewer targets pre-CAD (a consequence of the safety-net effect in itself) and so it is plausible that there would be a sensitivity drop. However, it seems clear that pre-CAD, participants were withholding many of their decisions until they had the aid of prompts: in images with no prompts available, there was still a significant increase in the false positive response rate. This reliance on the output of CAD to make a final decision was most detrimental for target absent images containing false prompts, where the false positive rate was almost double that of the unaided condition.

6.5 Summary

In this chapter, we described an experiment investigating the hypothesised safety-net effect with computer aided detection, where the expectation of prompts adversely affects the initial unaided search of the image. Non-expert participants searched for microcalcification clusters in 1/f noise distributions with their eye movements tracked throughout.

There were 100 images in the no-CAD condition and 100 in the CAD condition. In the CAD condition, participants viewed images initially unaided, and then once again with the aid of CAD. Overall, 80% of targets were prompted with 0.5 false prompts per image. The target prevalence was 40%. Eye movement data was used to calculate the proportion of images viewed (image coverage) and the targets and prompts fixated and their corresponding dwell time.

Experimental evidence for the safety-net effect was observed; before prompts were displayed in the CAD condition, image coverage and trial time were significantly lower than in the no-CAD condition. As a result, significantly fewer targets were fixated pre-CAD.

For prompted targets, sensitivity was significantly greater compared to the no-CAD condition, whereas there was no difference for unprompted targets. As with many clinical studies, the increase in sensitivity with CAD came with an associated cost of a significant increase in the false positive response rate.

Cross-sectional studies evaluating the efficacy of CAD should be aware of the safety-net effect if they make the assumption that the initial pre-CAD viewing of images is equivalent to unaided viewing. There exist alternative methods of CAD implementations that may have less of an influence of visual search behaviour, such as interactive CAD (discussed in Section 3.4). We will explore this type of CAD system further in Chapter 7.

Chapter 7

Interactive CAD

7.1 Introduction

Traditional CAD algorithms, where readers first search mammograms unaided and then once again with the aid of prompts, are often associated with an increase in sensitivity at the cost of a significant increase in false positives. This was observed with non-expert readers, as well as the safety-net effect (Chapter 6). The increase in the false positive response rate can easily be attributed to the high number of false prompts presented to readers with the method of prompting. Therefore, as discussed in Section 3.6, methods such as interactive CAD offer an alternative which may improve sensitivity compared to single reading alone without increasing the false positive rate by reducing the number of false positive prompts the reader will see.

An example of interactive CAD (ScreenPoint, 2021) is shown in Figure 7.1. The reader is free to review the image as in unaided viewing, but they may also wish to ‘query’ regions that raise suspicion and review the CAD output (if available) on that region. Deep learning convolutional neural networks are used to assign confidence values (between 1 and 100) to suspicious regions in the mammograms, displayed above each prompt. In addition to individual confidence values on prompts, an overall image score (between 1 and 10) is provided, denoting the probability that the image contains an abnormality. This score is based on the suspicious regions detected by the neural network.

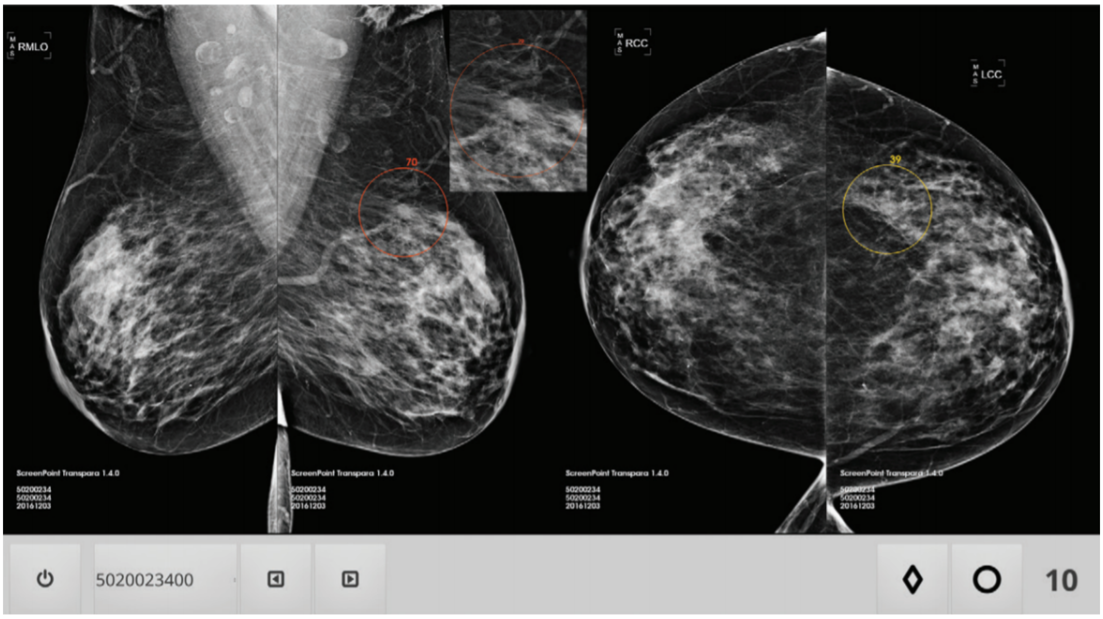


Figure 7.1: Clinical example of an interactive CAD system (ScreenPoint, 2021). MLO and CC-view mammogram with invasive ductal carcinoma, with the abnormality outlined by a CAD prompt. The prompt has been displayed after a query by a reader and is accompanied by a confidence value of 70/100 in the MLO-view and 39/100 in the CC-view. The colour of the prompt also reflects the confidence, changing from yellow (low confidence) to red (high confidence). The image score is 10/10, displayed in the bottom right of the display, denoting the overall likelihood of an abnormality being present in the case. Image from Rodriguez-Ruiz et al. (2019*a*).

Image scores give an indication to readers of the likelihood of finding a cancer in a case from the onset before they have queried any region. The distribution of image scores in a validation set of malignant cases is shown in Figure 7.2 for ScreenPoint's CAD algorithm. In a screening set of cases with a standard cancer prevalence, the score distribution is set such that there is an even distribution of cases amongst each score category (Rodríguez-Ruiz et al., 2019a).

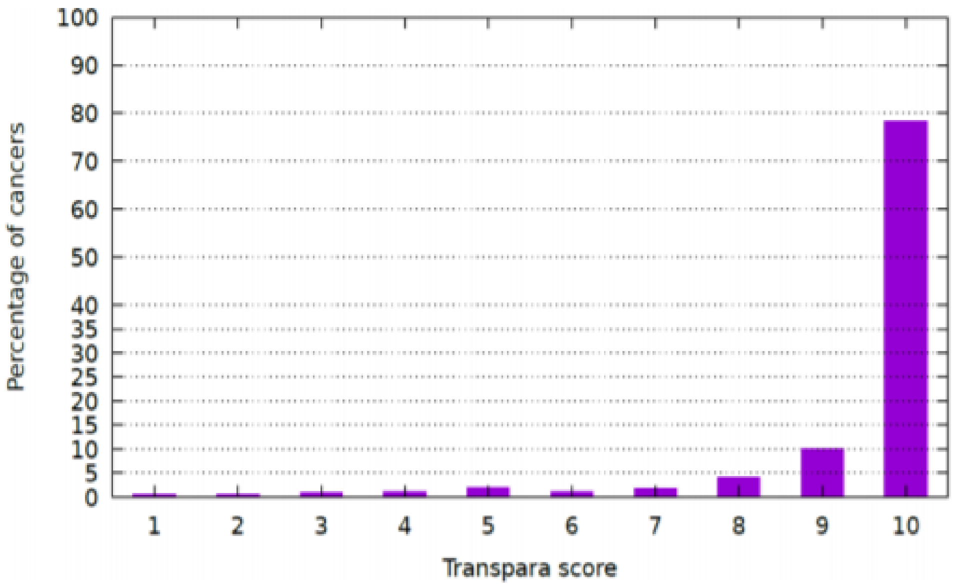


Figure 7.2: Distribution of image scores (referred to as Transpara score) amongst screen-detected cancers in a test set using ScreenPoint Medical CAD. The majority of cancer present cases are assigned a score of 10, with just 10% a score of 9, and just a few in the remaining score categories. Image from FDA (2018).

Clinical results with this method of CAD have shown that it can significantly improve pAUC compared to traditional CAD, from 0.57 to 0.62 (Hupse et al., 2013). Another study from the same group showed that AUC improved significantly from 0.87 with unaided reading to 0.89 with interactive CAD (Rodríguez-Ruiz et al., 2019a). Furthermore, reading time was significantly correlated with the image score, with readers spending more time on cases with a higher score (2% longer above a score of 5 and 11% shorter below a score of 6, compared to unaided reading) (Rodríguez-Ruiz et al., 2019a).

Drew et al. (2020) investigated how an interactive CAD method affected visual search behaviour in an eye tracking study with non-expert participants searching for Ts amongst Ls in synthetic backgrounds, in two prevalence conditions. Participants queried regions and were informed with a text prompt whether it was a ‘likely target’ or ‘not likely target’, rather than receiving a prompt with a confidence score. In the high prevalence setting (50%), d' did not change between the no-CAD and CAD conditions. However, in the low prevalence setting (10%) there was a small increase in d' (p=0.039). Furthermore, when CAD failed to provide a text prompt on targets, there was no associated miss cost with these compared to the no-CAD condition. This contrasted with the traditional CAD approach they used in the study, where a significant cost was incurred when CAD failed to prompt a target. This suggests that interactive CAD may eliminate the reduction in sensitivity associated with unprompted targets.

An interactive CAD approach, compared to traditional CAD, changes the way in which the reader receives information from the system. Local information, at the site of individual prompts, is more detailed and is only displayed when the reader actively clicks on that region rather than displaying every prompt for the whole image. The global information is in the form of an overall image score but does not provide spatial details. This study aims to investigate how this method of prompting will affect the overall performance of participants, and their behaviour on prompts and images according to the confidence level of associated CAD scores. The results discussed in this chapter have been published in Du-Crow et al. (2020).

7.2 Methods

This study comprised of two separate visual search tasks, Experiment 1 and Experiment 2. Both experiments used 1/f$^{1.5}$ noise distributions as in Chapter 6, with microcalcification clusters inserted into them as targets. In **Experiment 1**, there were two conditions in which non-expert participants searched images: a no-CAD condition and an interactive CAD condition. In **Experiment 2**, CAD was available in both conditions. However, in one condition, there was an additional overall image score provided below the image that denoted the likelihood that the image contained a target (but not its location). These are referred to as the CAD and CAD+Score conditions. In the CAD conditions, participants could query suspicious regions by clicking on them. If a prompt was available on that region, it would be displayed, along with a prompt confidence value (ranging between 1 and 100).

Forty-two participants (median age 22, age range 18-58, 34 female) and 43 participants (median age 21, age range 18-47, 35 female) were recruited for Experiments 1 and 2, respectively, and informed consent was obtained. All participants had not taken part in the study discussed in Chapter 6. The eye-movements of two participants in Experiment 1 and one participant in Experiment 2 could not be accurately calibrated so in those cases only behavioural data were collected. Sixteen participants in each experiment were undergraduate psychology students and received course credit for taking part. The rest (students, university staff, and members of the public) received £10 in exchange for their time. The study was approved by the University of Manchester Research Ethics Committee (2018-4586-6410).

7.2.1 Stimuli

The images used in this study were the same as those used in Chapter 6, referred to as image sets A and B. These images were created in MATLAB using open-source code (Methven and Qi, 2018), and are synthetic 1/f noise distributions. The targets used in this study were malignant microcalcification clusters, which were extracted from magnified images of slices of mastectomies (Warren et al., 2012). The images were sized at 800×800 pixels.

Since the study in Chapter 6 was completed prior to the experiments described in this chapter, we were able to use the target detection percentages in the no-CAD condition of that study to reorder the images into sets A' and B'. This was to ensure that any differences in the difficulties of detecting the targets between the two image sets were evened out for these experiments. To do this, we shuffled the targets that were prompted by CAD in sets A and B and compared the target detection percentage in the no-CAD condition, until the medians were within 0.5% and the distributions of the 40 targets were matched according to a Kolmogorov–Smirnov test ($p>0.98$). In total, 4 target images were swapped between sets A and B to form A' and B', with the non-target images remaining unchanged. The result of this is shown in Figure 7.3, where the distribution of target detection percentages of sets A' and B' used for Experiments 1 and 2 is shown. The median value of sets A' and B' were 78.9%, and the means were 65.2% and 67.0% for set A' and set B', respectively.

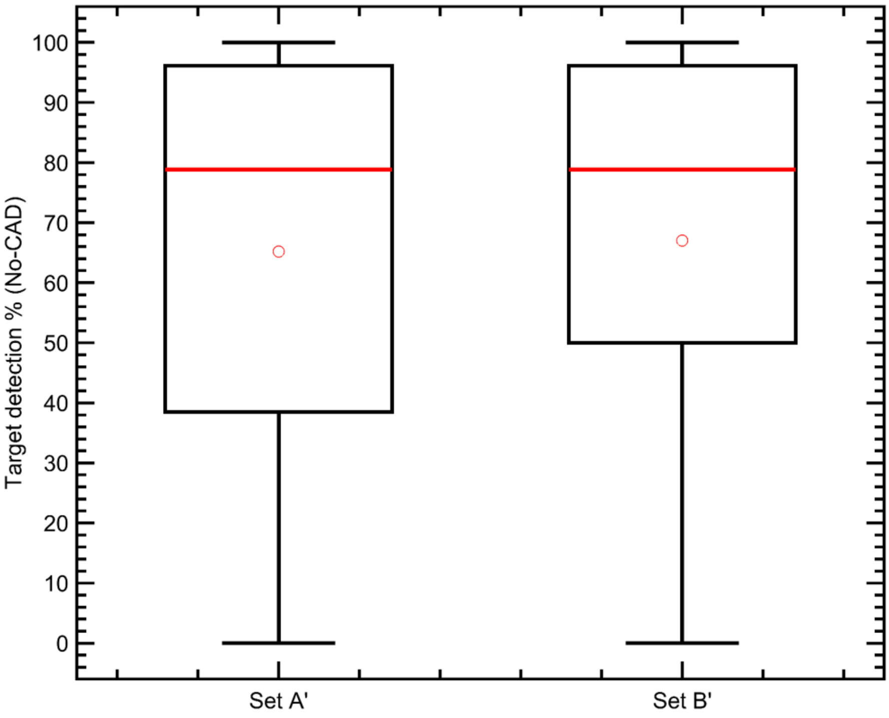


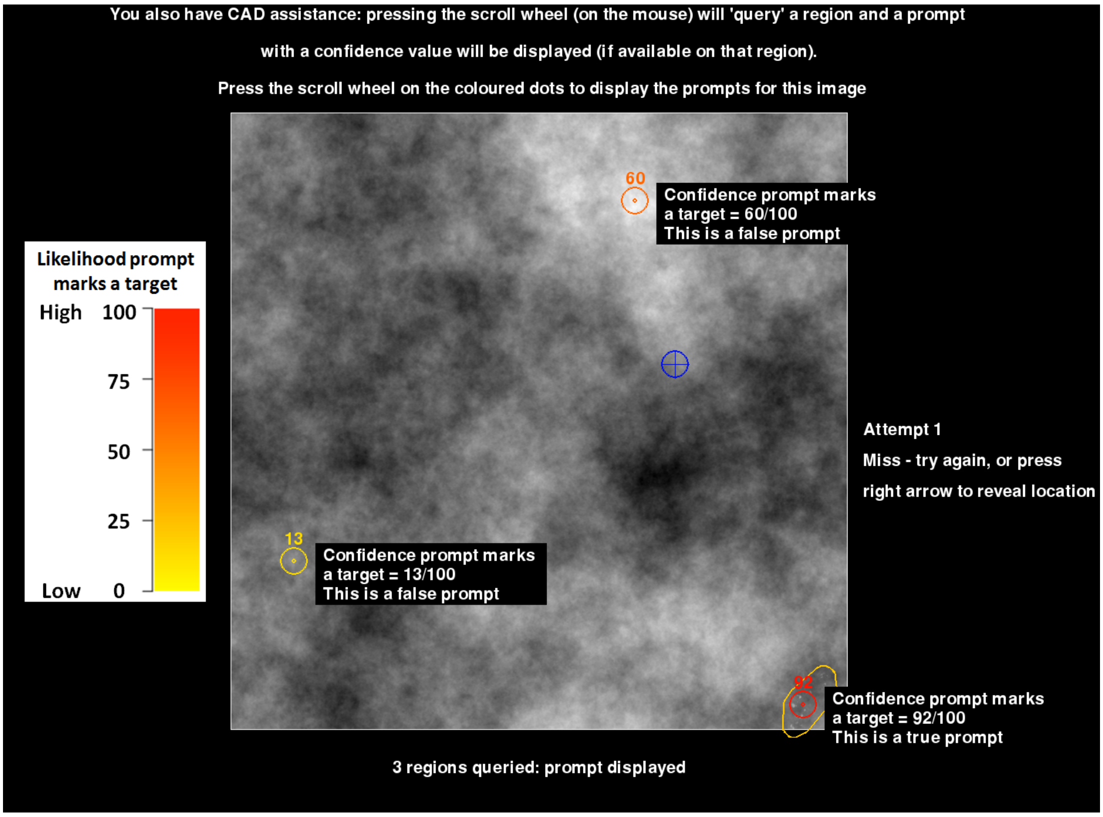
Figure 7.3: Target detection percentage in the no-CAD condition of the safety-net experiment described in Chapter 6. Targets were swapped between sets until medians and distributions were matched in terms of detection percentages.

7.2.2 Experimental setup and procedure

The experimental setup was the same as that shown in Figure 6.3, with participants at a viewing distance of 73 cm. Eye movements were again tracked with an EyeLink1000 desktop eye tracker and the experiment code was written using PyGaze (v0.6.0) (Dalmaijer et al., 2014). All participants were tested using the FrACT (Bach, 1996) software to ensure their vision was normal (glasses or contact lenses were allowed).

Participants underwent a training set of images to gain an understanding of the target and image appearance, the search task, and how to operate the interactive CAD system. Initially, participants were shown isolated examples of the microcalcification clusters as an introduction to the targets. This was accompanied by an example of one cluster inserted into a background. Following this, two further clusters were shown inserted into the backgrounds with their positions outlined. This provided participants with further familiarity with the targets before moving onto the interactive section of the training set.

The interactive training set started with four images without CAD to familiarise participants first with the controls, provide them with feedback on their input, and allowed them to gain further practice with the targets. They would left click to place a marker on suspicious regions and right click on those markers to remove them if they changed their mind. On-screen text prompts would inform them if they had successfully found targets. Once they had finished marking regions, they would press the right arrow key to move on. After those four images without CAD, they were introduced to how the interactive CAD system worked, which is shown in Figure 7.4. This first example indicated where the prompts were and instructed them to query them (by pressing down on the scroll wheel of the mouse) so they would receive the text prompts stating whether it was a true or false prompt and explained the associated confidence value.



You also have CAD assistance: pressing the scroll wheel (on the mouse) will 'query' a region and a prompt with a confidence value will be displayed (if available on that region).

Press the scroll wheel on the coloured dots to display the prompts for this image

Confidence prompt marks a target = 60/100
This is a false prompt

Attempt 1

Miss - try again, or press right arrow to reveal location

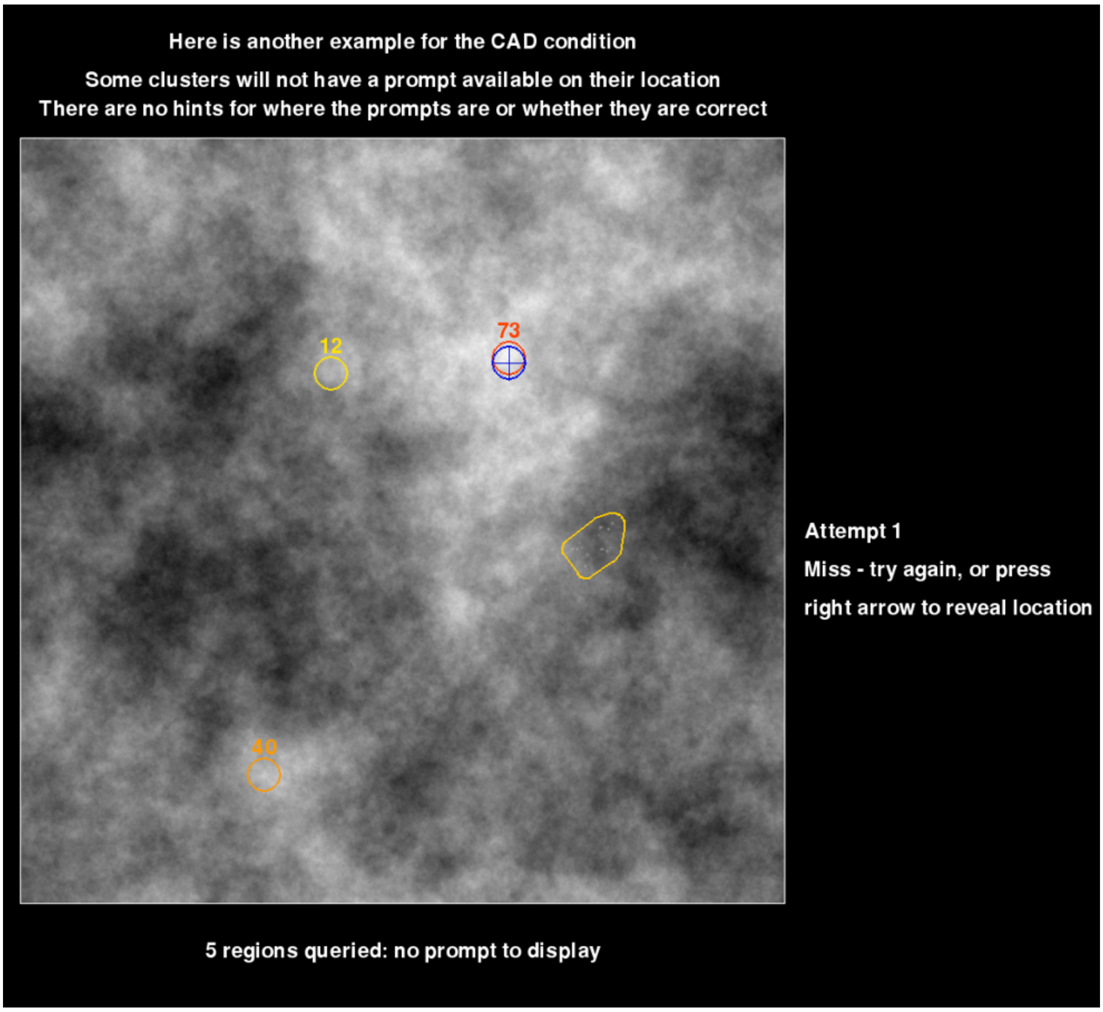
Confidence prompt marks a target = 13/100
This is a false prompt

Confidence prompt marks a target = 92/100
This is a true prompt

3 regions queried: prompt displayed

Figure 7.4: Screenshot of experiment training introducing interactive CAD used in Experiments 1 and 2. Participants are instructed to ‘query’ the small coloured dots by clicking on them, which brings up each prompt and the associated message informing them whether it was a true or false prompt. The cluster has been outlined in yellow in this example for illustrative purposes only since here the participant has only marked a non-target region and received a message informing them that this was an incorrect marker.

There were two further examples with CAD where there were no hints on where to query the images. A screenshot of the first of these is shown in Figure 7.5. In this example, there were three false prompts available, demonstrating that the CAD software did not always prompt targets. The other example had two prompts available, a true prompt and a false prompt.



Here is another example for the CAD condition

Some clusters will not have a prompt available on their location
There are no hints for where the prompts are or whether they are correct

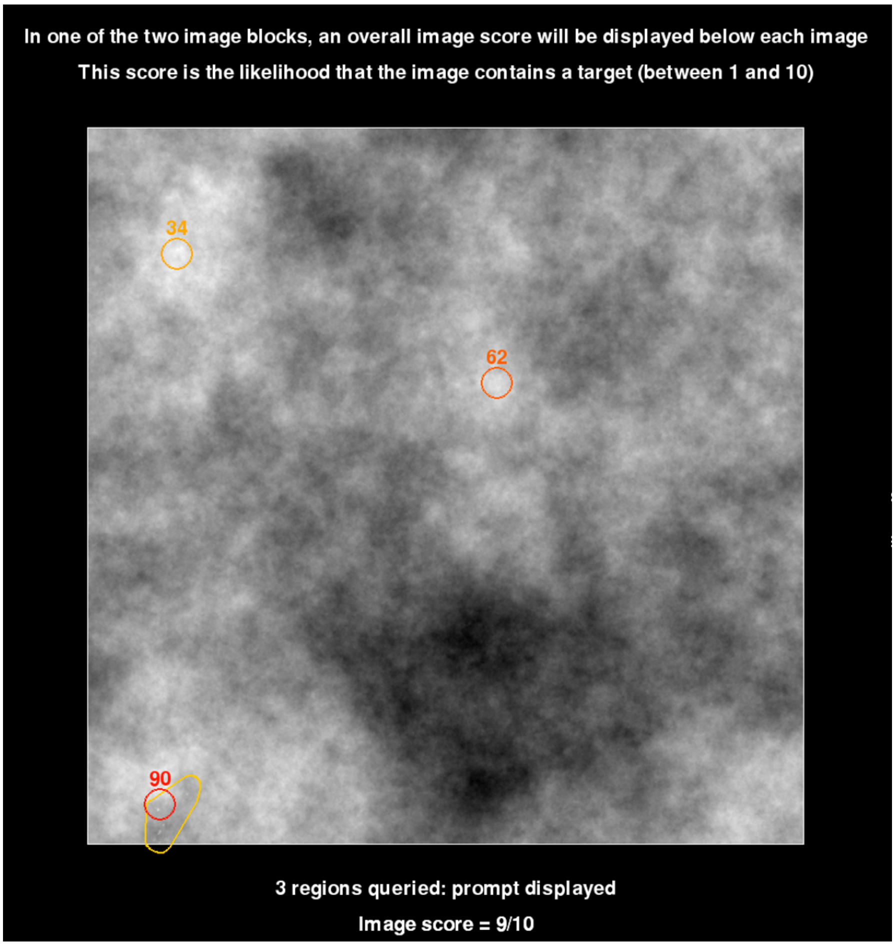
Attempt 1

Miss - try again, or press
right arrow to reveal location

5 regions queried: no prompt to display

Figure 7.5: Screenshot of experiment training with interactive CAD example showing false prompts used in Experiments 1 and 2. In this case, the participant has queries all 3 of the available prompts as well as 2 additional regions without prompts available. The cluster is outlined in yellow for illustrative purposes and the participant has marked a false prompt and received feedback on their click to the right of the image.

In the training set for Experiment 2, there were further examples provided to introduce participants to the overall image score. The first of these is shown in Figure 7.6. In total, there were three examples with an image score in the training set. The image score was presented below the image and was provided in addition to CAD. Finally, in the training set for both experiments, there was an example to remind them on how to remove markers once they had placed them.

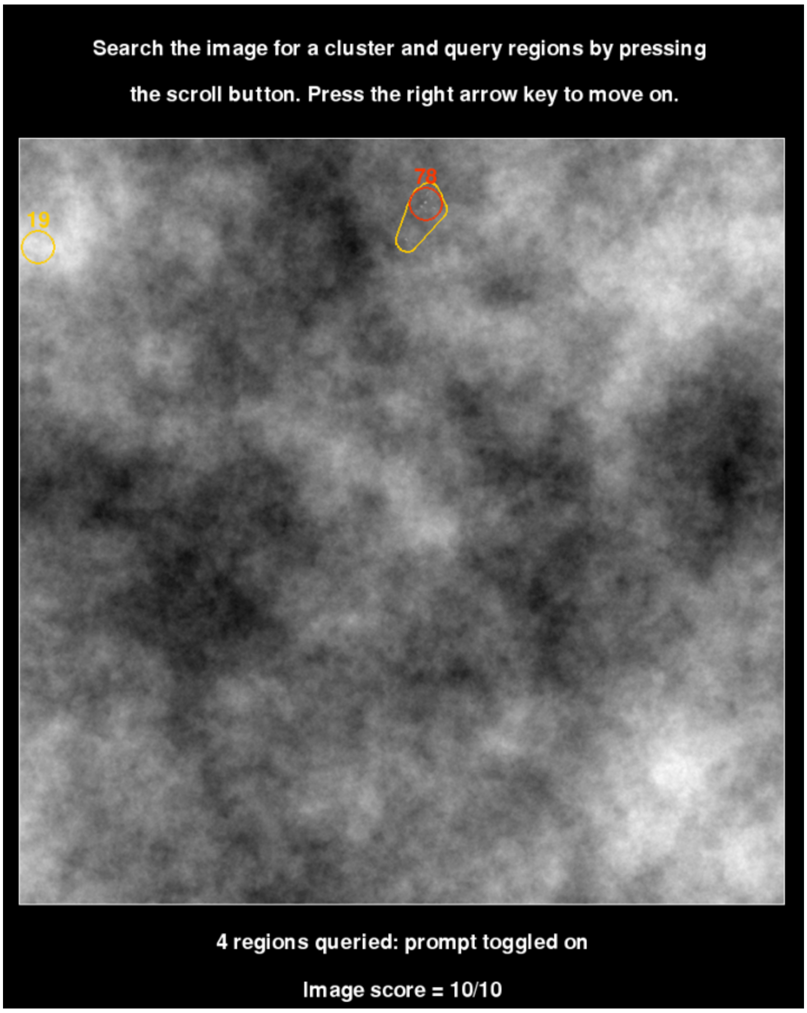


In one of the two image blocks, an overall image score will be displayed below each image
This score is the likelihood that the image contains a target (between 1 and 10)

3 regions queried: prompt displayed
Image score = 9/10

Figure 7.6: Screenshot of experiment training introducing overall image score used in Experiment 2. An image score of 9 indicates that there is a high probability that it contains a target. In this case, the participant has queried 3 regions, all of which had a prompt available at their location (2 false prompts and 1 true prompt). The cluster is outlined in yellow for illustrative purposes only.

Once the training was completed, participants were given the opportunity to ask the researcher questions before starting the experiment. They were again given final reminders (see Section 6.2.3), such as there being between 1 and 4 prompts available on each image in the CAD conditions, toggling prompts on and off with the scroll button instead, and the image score being provided in one of the conditions (in Experiment 2). An example experiment screen is given for the CAD+Score condition in Experiment 2 in Figure 7.7. This is the same as the CAD-only conditions in Experiments 1 and 2, just with the additional Score text below the image.



Search the image for a cluster and query regions by pressing the scroll button. Press the right arrow key to move on.

4 regions queried: prompt toggled on

Image score = 10/10

Figure 7.7: Screenshot of experiment image in the CAD+Score condition in Experiment 2. The text at the top is the same in both Experiment 1 and 2 for all conditions with CAD. The only addition for the CAD+Score condition is the image score provided below the image. The cluster is outlined in yellow for illustrative purposes only.

The order of the 200 images in image sets A$'$ and B$'$ were randomised for each participant. For both Experiments 1 and 2, we blocked the participants by conditions. Half of the participants started with condition 1 (no-CAD/CAD) and the other half with condition 2 (CAD/CAD+Score). For both conditions 1 and 2, half of the participants saw images drawn from set A$'$ and the other half saw images drawn from set B$'$. We randomly allocated participants into these groups.

Target prevalence in each image set was 40%, with 4 out of the 40 targets having no prompt available on its location (unprompted). Participants were not informed of the exact target prevalence in the experiment, but were informed that it was within the range of 1% to 50%. We did not enforce a time limit on the images but did advise participants not to spend longer than 1 minute on a single image. A break screen appeared after every 25 trials during the experiment.

In the CAD condition, 90% of targets had a prompt available on their location, with an average of 2 false prompts available per image – a similar sensitivity and rate of false prompts available as the interactive CAD system described in Hupse et al. (2013). To assign the prompt confidence values to the prompts, we used the distribution of target detection percentages measured previously in the no-CAD condition of the safety-net experiment, shown in Figure 7.8. For true prompts, the confidence value was equal to the target detection percentage in our previous experiment (Chapter 6). For false prompts, this distribution was used to randomly assign values to each prompt. For example, 9 out of 80 target images (11%) had a target detection percentage between 0% and 10%, therefore 11% of the 200 false prompts in each image set were randomly assigned confidence values between 1 and 10. The colour of the prompts reflected the confidence, ranging from yellow (low confidence) to red (high confidence).

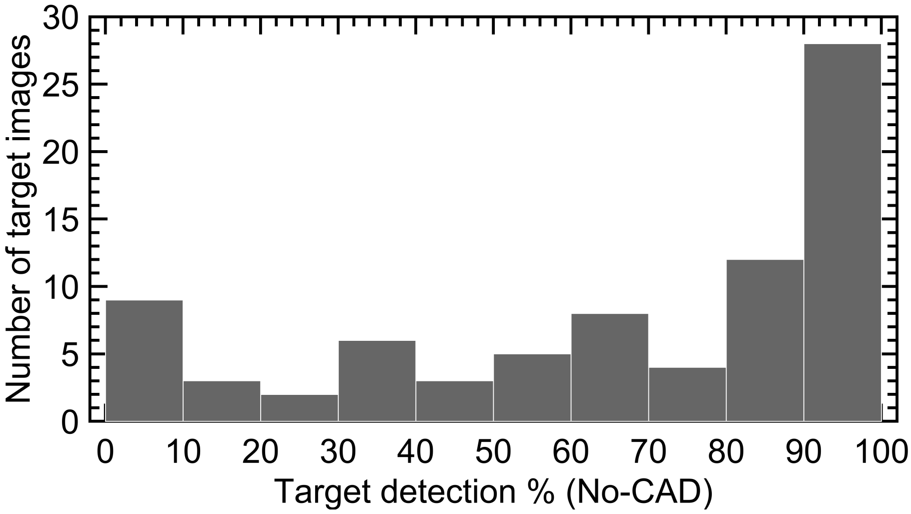
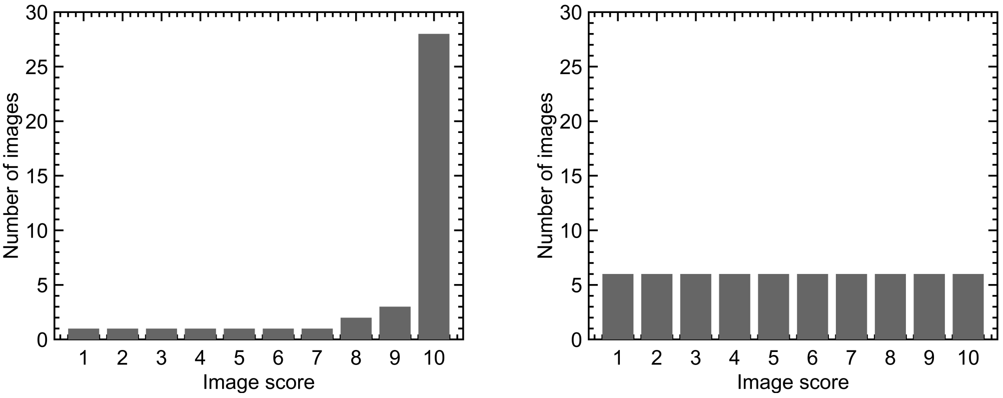


Figure 7.8: Distribution of target detection percentages in the no-CAD condition for all 80 targets in image sets A′ and B′. These were used to assign the prompt confidence values.

For Experiment 2, image scores were distributed for target present and target absent images differently, as shown in Figure 7.9. The majority (28/40) of target present images were assigned a score of 10, 3/40 a score of 9, 2/40 a score of 8, and 1 target image per score from 7 to 1. For target absent images, the images where equally split between the scores, with 6 images per score. These distributions were based on a clinical example of an interactive CAD system (see Figure 7.2).

To track participant eye movements, a nine-point calibration procedure was completed at the beginning of each condition and repeated if accuracy of fixations exceeded 1 degree visual angle on any point in the validation sequence. A 9-point validation sequence followed, and if any fixation deviated by greater than 1 degree visual angle then the calibration would be repeated. Drift correction procedures were completed midway between and immediately after break screens and the calibration procedure would be repeated if necessary. If an accurate calibration could not be achieved, participants completed the experiment without eye tracking, and behavioural data was collected.



(a) target present

(b) target absent

Figure 7.9: Number of images with a given image score for (a) target present and (b) target absent images.

As before, for classification of fixations, the maximum drift in eye position coordinates between start point of fixation and end point of fixation should be less than 1.5 degrees visual angle, and the minimum fixation duration was set at 100 ms.

7.2.3 Analysis

To determine whether participants clicked on targets, we used the same process that was outlined in Chapter 6, where the convex hull of calcifications was extended outwards by 10 pixels (approximately 0.25°) and any marker placed inside this was classified as a hit. To determine whether a participant had queried a prompt region, and therefore whether to display any prompt available for that region, we used a function that detected any clicks within a circle with a radius the size of a prompt extended by 20 pixels or 0.45° (total radius of 38 pixels or 0.84°). This larger tolerance was used to capture a larger area around the prompt since queries are likely to be less precise than marks on perceived targets. Participants could query as many times as they wanted, and were able to toggle prompts on and off.

Image coverage was calculated using a circle of radius of 2.5 degrees visual angle centred around fixations, with any point inside that region considered covered. For calculating dwelltime on queried regions, we used a function that detected the collision between circles (useful field-of-view around centre of fixations and the circular-defined queried region), and summed all fixations within this region. To compare the differences between conditions for trial time, image coverage, sensitivity, and false positive response between conditions, we used the bootstrap approach that was outlined in Chapter 6.

7.3 Results

Results are given for the no-CAD and CAD conditions in Experiment 1, and the CAD and CAD+Score conditions in Experiment 2. Firstly, participant data were looked at in terms of overall performance and thoroughness of visual search. Then we explored how participants queried the prompts and the effect of prompt confidence on how participants acted on those prompts. Finally, we analysed how the image score affected the trial time and false positive rate, as well as the interaction with prompt confidence.

7.3.1 Observer performance

As shown in Table 7.1, the mean trial time was not significantly different in the CAD condition compared to no-CAD in Experiment 1, although, numerically, the CAD condition took 1.35s longer. There was also no difference in the percentage image coverage. Overall sensitivity was numerically worse with CAD and went from 67.3% in no-CAD to 65.7% in CAD, but this difference of 1.55% was not significant. The mean number of false positive responses per image was 0.26 for no-CAD and 0.22 for CAD, a non-significant difference of 0.04. The proportion of targets fixated was 89.4% in the no-CAD condition and 90.4% in the CAD condition.

Table 7.1: Experiment 1 results for no-CAD and CAD conditions. The 95% CIs, *t*-statistic, and *p*-value were calculated using the bootstrap approach outlined in Section 6.2.4.

	No-CAD	**CAD**	**Diff.**	**[95% CIs]**	**t**	**p**
Trial time (s)	12.09	13.43	1.35	[− 0.17, 2.80]	1.78	0.083
Image coverage (%)	65.07	65.97	0.90	[− 1.85, 3.55]	0.65	0.517
Sensitivity (%)	67.30	65.70	1.55	[− 5.4 , 8.6]	0.45	0.658
FPs/image	0.26	0.22	0.04	[− 0.03, 0.13]	1.02	0.313

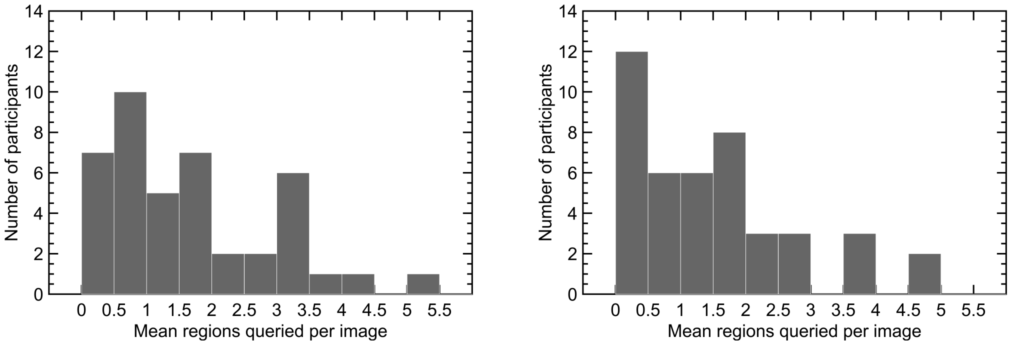
Table 7.2 shows the results for Experiment 2, where there was a significant increase in mean trial time with the introduction of the image score of 1.87s. The three other variables only differed numerically. The percentage image coverage was 2.52% larger in the CAD+Score condition, but this difference failed to break the .05 threshold. Overall sensitivity was 64.1% for CAD and 67.1% for CAD+Score, a non-significant difference of 3.0%. The false positive rate was 0.40 per image for the CAD+Score condition and 0.33 for CAD, a numerical difference of 0.06. The proportion of targets fixated was 88.2% in the CAD condition and 90.6% in the CAD+Score condition.

Table 7.2: Experiment 2 results for CAD and CAD+Score conditions. The 95% CIs, *t*-statistic, and *p*-value were calculated using the bootstrap approach outlined in Section 6.2.4. Statistically significant results are highlighted in bold.

	CAD	**CAD+ Score**	**Diff.**	**[95% CIs]**	**t**	**p**
Trial time (s)	13.86	15.73	1.87	[0.2 , 3.6]	2.13	**0.039**
Image coverage (%)	65.50	68.00	2.52	[− 0.19 , 5.21]	1.83	0.075
Sensitivity (%)	64.10	67.10	3.00	[− 2.5 , 9.1]	1.07	0.289
FPs/image	0.33	0.40	0.06	[− 0.004, 0.13]	1.79	0.081

7.3.2 Effect of prompts

The distributions of the mean number of regions queried for Experiments 1 and 2 are shown in Figure 7.10. The median number of regions queried (with the 25th and 75th percentiles) for Experiment 1 was 1.44 (0.65−2.51) and 1.31 (0.45−2.11) for Experiment 2. This demonstrates that participants were engaging with the CAD system in both experiments.



(a) Experiment 1

(b) Experiment 2

Figure 7.10: Distribution of the mean number of regions queried per image in each experiment. For Experiment 2, this is the mean across both conditions.

As shown in Table 7.3, participants were significantly more likely to place a marker in reaction to true prompts than false prompts. In Experiment 1, 23.3% more true prompts were acted on than false prompts (t(41)=4.018, p<0.001). In Experiment 2, this difference was 17.9% (t(42)=3.033, p=0.004) and 17.4% (t(42)=3.677, p<0.001) for the CAD and CAD+Score conditions, respectively. Participants were also significantly more likely to mark any prompt (either true or false) compared to a region where they had queried but no prompt was available (p<0.001 for all conditions).

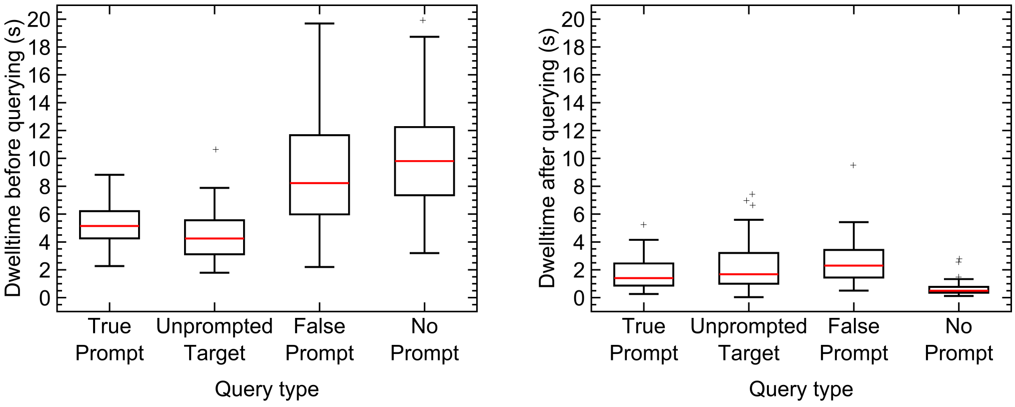
For each query, we measured the dwelltime on that region before and after they initiated the query (i.e. clicked on it). We averaged the dwelltime for each participant

Table 7.3: Observer behaviour by prompt type for Experiments 1 and 2: true prompts (TPs, where the prompt marks a target), false prompts (FPs, where the prompt does not mark a target), and no prompts (where the participant queries but no prompt is available for that location). For each prompt type, the median number of queries for that type is given, along with the mean percentage of queries of that type where a marker is placed by the participant.

		Median number regions queried per image (range)	TPs (N=36)		FPs (N=200)		No prompts	
			Queried (median)	Acted on (%)	Queried (median)	Acted on (%)	Queried (median)	Acted on (%)
Exp. 1	CAD	1.44 (0.29−5.31)	15	89.8	10	66.6	107.5	6.2
Exp. 2	CAD	1.18 (0.03−5.09)	11	74.3	11	56.4	96	13.4
	CAD+Score	1.39 (0.03−5.20)	10	77.7	11	60.3	106	16.4

and classified it by the type of query, i.e., whether it was on a true prompt, an unprompted target, false prompt, or no prompt available (non-target region). The results of this are shown in Figure 7.11, separated by dwelltime before and after the query was made. For comparing the dwelltimes, we were interested in target (true prompt vs unprompted target) and non-target regions (false prompt vs no prompt). All comparisons were made using participants that had dwelltime data for both query type using paired t-tests.

Before a participant queries a target region (true prompt or unprompted target), there is no difference between these two query types, and so the distributions should be the similar (the actual targets in each will be different so this may cause small variations). Before queries, the median time spent on unprompted targets was numerically lower, but there was no significant difference between the distribution of true prompt and unprompted targets (t(31)=1.842, p=0.07). Once the region is queried, despite the lack of prompt on the unprompted target regions, we expected participants to dwell on those regions for a period similar to prompted target regions, since they might have wanted to verify whether it was a target region or not. The median dwelltime was numerically higher for unprompted targets, but there was no significant difference in dwelltime after querying between true prompt and unprompted target regions (t(31)=1.772, p=0.08). Similarly, for non-target regions before the query was made, the distribution of dwelltimes should be similar since there is no difference between a false prompt region and a no prompt region. There was no significant difference between false prompts and no prompt regions (t(37)=0.812, p=0.42). After querying, participants were expected to move on quickly from non-target regions with no prompt present. This was what was observed, with a significant difference between false prompt and no prompt regions after querying (t(37)=6.914, p<0.001).



(a) (b)

Figure 7.11: Results for Experiment 1 only. Mean dwelltime (a) before and (b) after participants had queried each region. Regions are classified by the query type. Participants were only used if they had data available for both query types to be compared (True prompts vs Unprompted targets and False prompts vs No prompts).

We were also interested in the proportion of participants that queried each individual target in the context of how ‘difficult’ that target was to detect. Figure 7.12 shows the proportion of participants that queried each target versus the target detection percentage from the safety-net experiment (Chapter 6). There was a moderate positive correlation (r=0.53, p<0.001). However, this correlation was dependent on the 'difficult' targets, and when only targets with a detection rate of >40% are considered, the correlation was no longer significant (r=0.19, p=0.14).

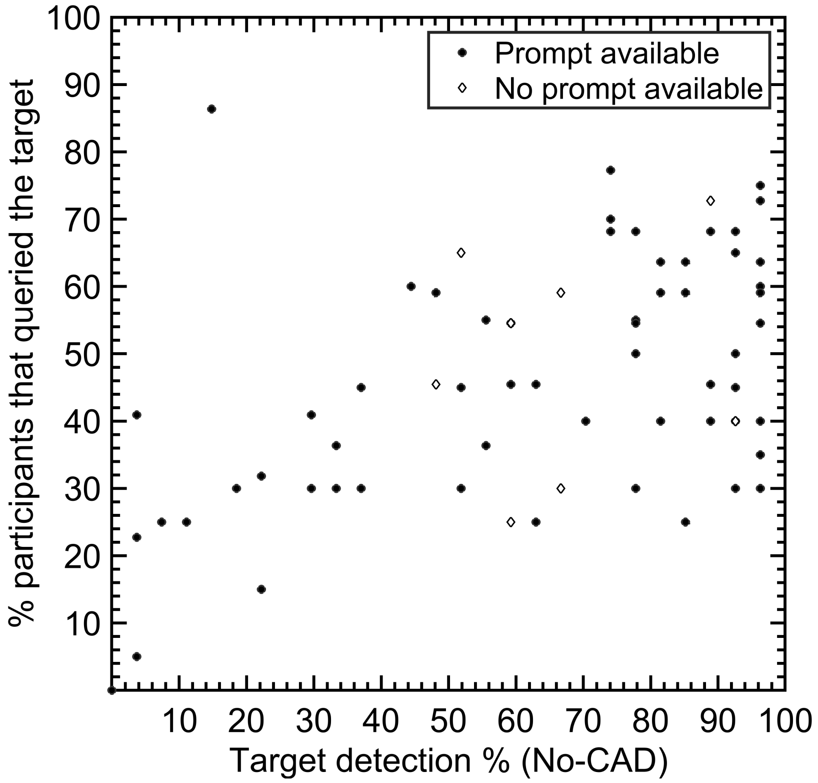


Figure 7.12: Percentage of targets that were queried by participants versus their detection rate in the safety-net experiment in Chapter 6.

In Figures 7.13 – 7.15 it can be seen how the confidence value affected how participants acted on prompts. The higher the prompt confidence, the more likely a participant was to believe that it was a target and subsequently place a marker on that region. True prompts were more likely to be acted upon than false prompts for all confidence values, although this was less distinguishable for confidence values below 60 in the CAD condition of Experiment 2 (Figure 7.14).

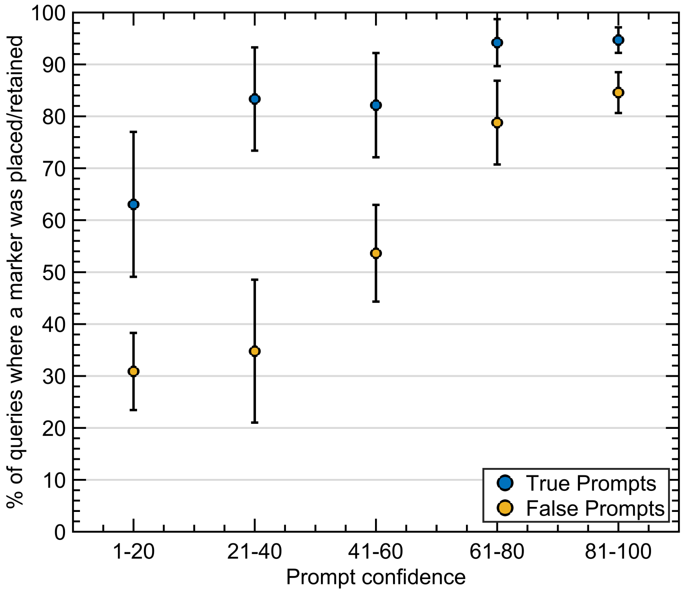


Figure 7.13: Results for Experiment 1. Mean percentage of queries where a participant subsequently clicked on that location or retained a marker they had already placed there versus the prompt confidence. The error bars are the standard errors across participants.

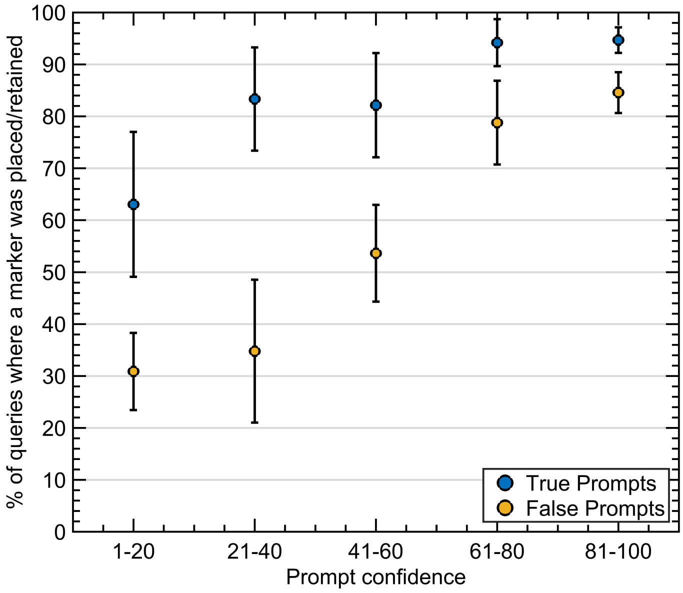


Figure 7.14: Results for the CAD condition in Experiment 2. Mean percentage of queries where a participant subsequently clicked on that location or retained a marker they had already placed there versus the prompt confidence. The error bars are the standard errors across participants.

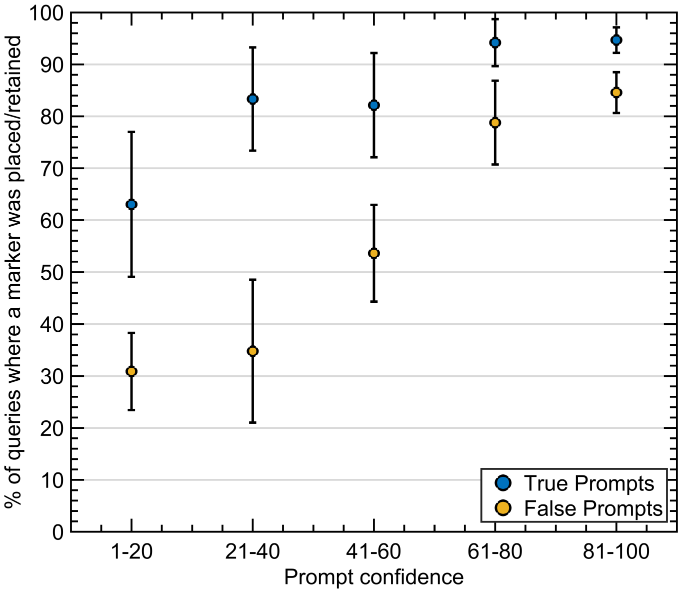


Figure 7.15: Results for the CAD+Score condition in Experiment 2. Mean percentage of queries where a participant subsequently clicked on that location or retained a marker they had already placed there versus the prompt confidence. The error bars are the standard errors across participants.

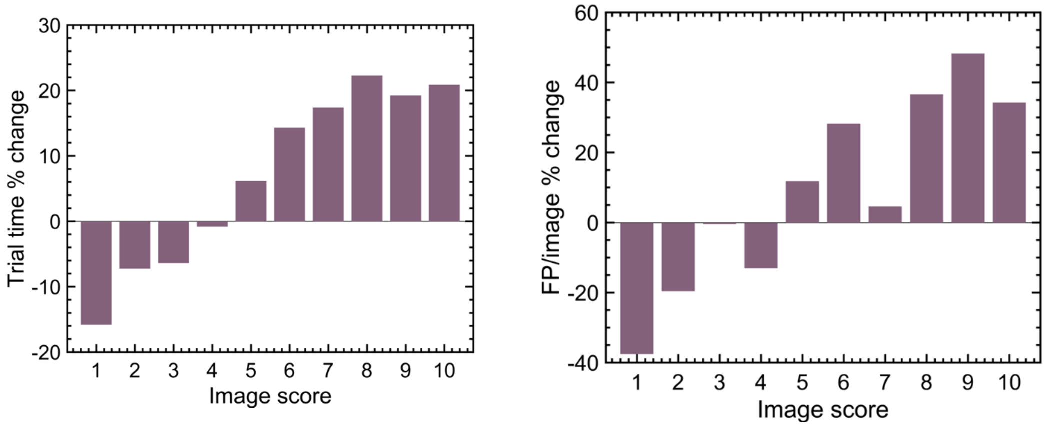
When participants fixate a target, if they recognise it as a potential target region, they may choose to query it. Of the targets that were fixated but not marked as a target by participants, we measured the proportion of those that were queried, given in Table 7.4. This was to gain an understanding of how CAD was used for detecting additional targets that they ultimately dismissed. Participants only queried an average of 17.1% of fixated but unmarked targets across the three CAD conditions. Even after fixating a target region, only in a small fraction of cases was that region then queried.

Table 7.4: The number of targets in each condition with CAD that were fixated but not marked by participants. We measured the proportion of these targets that were queried.

	Experiment 1	Experiment 2	
	CAD	CAD	CAD+Score
Median no. targets fixated but not marked	9.5	8.0	9.0
Median proportion of those targets queried	15.4% (2.8−27.5%)	12.5% (0−28.6%)	3.3% (0−18.2%)

7.3.3 Effect of image score

As shown in Figure 7.16a, in Experiment 2, the higher the image score in the CAD+ Score condition, the longer participants spent on those images compared to the same images in the CAD condition without a score. This correlation was significant (r(8)=0.96, $p<0.001$). Above a score of 5, there was a $>10\%$ increase in viewing time, and below 4 there was a reduction of $>5\%$. In Figure 7.16b, it can be seen that an image score $\geqslant$5 led to a greater number of false positive errors compared to those same images in the CAD condition, with a $>34\%$ increase above a score of 7. An image score of 1, 2 or 4 led to a decrease in the number of false positives, with a $>13\%$ reduction. This correlation between the change in false positive error rate and image score was again significant (r(8)=0.92, $p<0.001$).



(a) (b)

Figure 7.16: Results from Experiment 2. Percentage change in (a) trial time and (b) false positive responses per image when going from the CAD to CAD+Score condition as a function of image score. The same images were compared between the two conditions.

We also investigated how image score affected observer sensitivity in the CAD and CAD+Score condition. For the 28 images with a score of 10, the sensitivity was numerically larger by 4.6% in the CAD+Score condition compared to those same images in the CAD condition, but this was not significant (65.6% vs 61.0%, p=0.47). For the remaining score values for target-present images, there were too few images for each score (3 images or below) to reliably calculate sensitivity change.

In Figure 7.17, the interaction between prompt confidence and image score is shown. For a given prompt, the likelihood the participant would act on it by placing a marker was mostly invariant to overall image score, and appears to be primarily influenced by the confidence value of that prompt. So, when a prompt was available for the area queried, it almost completely overruled the general tendency of participants to rely on the image score (as shown in Figure 7.16b). However, it should be noted that most of the time the image score was all participants had in Experiment 2, since most areas queried did not contain a prompt.

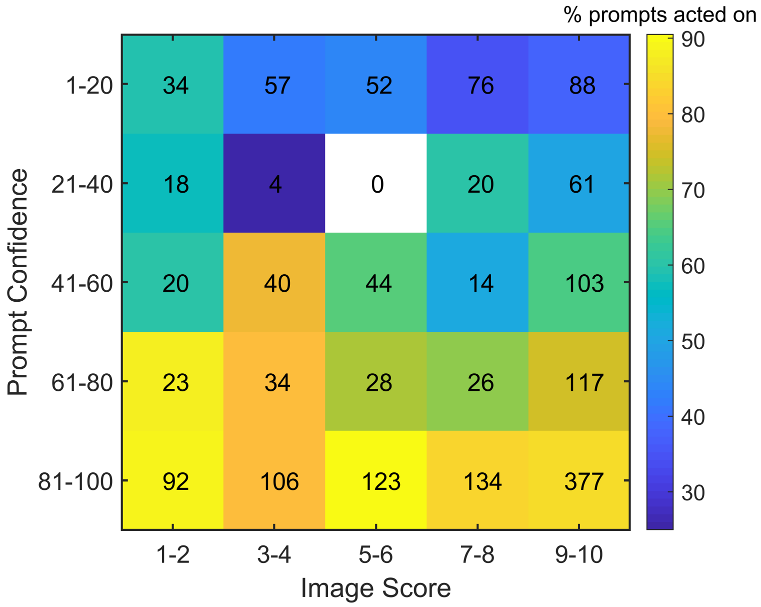


Figure 7.17: Results from the CAD+Score condition in Experiment 2. The percentage of true and false prompts acted on (marked by a participant) as a function of prompt confidence and image score. Both prompt confidence ($1-100$) and image score ($1-10$) have been divided into five bands. The percentage of prompts acted on ranges from 25.0% to 90.5%. The number in each square is the number of data points available for that combination of image score/prompt confidence.

We evaluated the trend between the proportion of prompts acted on and prompt confidence or image score. For prompt confidence, as with Figures 7.13 – 7.15, the proportion of prompts acted on was correlated with confidence. We plotted the proportion of prompts acted on for each prompt confidence, shown in Figure 7.18. The correlation between the proportion of prompts acted on and confidence was significant for both true prompts (r(23)=0.70, $p<0.001$) and false prompts (r(63)=0.66, $p<0.001$). In Figure 7.19, we compared the distribution of the proportion of prompts acted on for each image score band. A Kruskal-Wallis test showed that there was no relationship between the proportion of prompts acted on and image score (p=0.72).

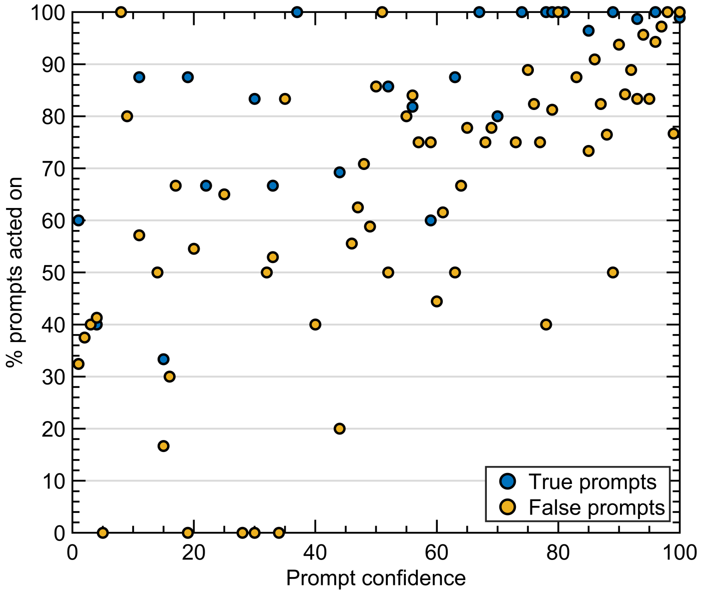


Figure 7.18: Percentage of prompts acted on versus prompt confidence, plotted for all confidence values.

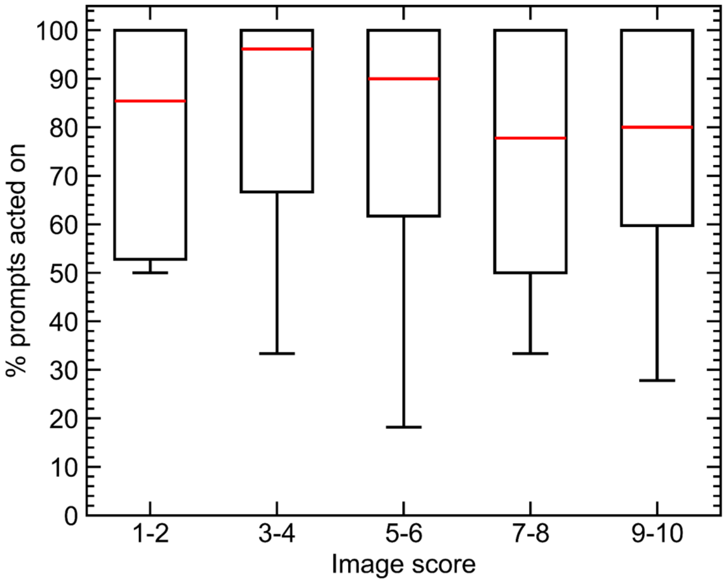


Figure 7.19: Percentage of prompts acted on versus image score.

7.4 Discussion

Interactive CAD systems offer an alternative approach to traditional CAD, where additional information is only provided to the reader when they request it. In our study, CAD did not benefit them in terms of their sensitivity and achieved a similar specificity compared to an unaided condition (Experiment 1). Furthermore, providing participants with an overall image score to indicate the probability of a target being present within the image also did not change their performance in terms of sensitivity or specificity (Experiment 2), but did significantly increase the time spent on images in the CAD+Score condition. Prompts were accompanied by a confidence value between 1 and 100, with participants more likely to mark prompts the higher the confidence value. True prompts were more likely to be marked than false prompts for all confidence values in Experiment 1, and for values above 60 in Experiment 2. Image score was correlated with both time spent on the images and the false positive response rate: images with a high overall score led to longer trial times and more false positive errors, with the opposite being true for low score images. When a prompt was available on images in the CAD+Score condition, we found that it was the confidence value of this prompt that drove participants' decision-making on whether or not to place a marker.

Clinical studies have reported improvements in sensitivity with interactive CAD systems with radiologists (Hupse et al., 2013; Rodríguez-Ruiz et al., 2019a), contrasting with the results of this study where the introduction of interactive CAD prompts and an overall image score failed to improve sensitivity above the level of unaided reading. The CAD system used in our study was modelled on those described in clinical studies, using a similar operating point, prompt appearance, and distribution of image scores. There are, however, key differences which may have contributed to the lack of improvement in sensitivity in our study. One important difference is that we used non-expert readers searching synthetic images, as opposed to radiologists searching mammograms. Although in the latter case motivation should be higher, given that the costs in missing a target are far greater, we do not expect this to be a factor since we are only interested in the within-participant differences between conditions. Furthermore, the motivation was assumed to remain consistent for the duration of the study, and we used a fully crossed study design (blocked by CAD condition).

There are more likely explanations for the lack of improvement with CAD in our study. If a participant did not find the target, they would not have queried it, would not have been presented with the prompt, and therefore they did not benefit from CAD. Also, the synthetic images used in this study did not have any particular region where a target was more likely to be found, which contrasts with mammograms where the likelihood of finding an abnormality depends on the underlying anatomy. Because of this, participants may have struggled to make an informed choice on where they should query the image. This was apparent in Table 7.3, where the majority of queried prompts were on regions without a prompt available. Compared to traditional CAD, participants saw far fewer true prompts, which in our case meant they did not improve their sensitivity, but they also saw fewer false prompts and so the false positive error rate did not change as with the previous study (see Chapter 6).

Furthermore, the number of targets that participants fixated was similar between the CAD and no-CAD conditions (90% vs 89%). Therefore, participants were not finding additional targets in the CAD condition and failed to recognise targets or dismissed them at a similar rate in both conditions. There was an approximate 25% difference in the number of targets fixated and detected. It was found that only a small fraction (around 17%) of these targets were actually queried in the CAD condition, which may also explain why CAD did not lead to an improvement in sensitivity. The low proportion of queries on these undetected targets may be a result of participants fixating them but not registering them as a potential target region.

This is a disadvantage of interactive CAD, where there is no warning that you have fixated a target even if there is a prompt available there. In traditional CAD, the prompt will be automatically displayed, but may be accompanied by one or more false prompts that you must distinguish between. A combination of interactive CAD and methods which utilise eye tracking data by indicating regions that were both fixated and prompted (see Chen and Gale (2010*a*)) could improve the use of interactive prompts. With the AI-CAD system described by Rodríguez-Ruiz et al. (2019a), traditional CAD is available in addition to the interactive prompts. Readers can choose to display the most suspicious prompts, with a false positive prompt rate of 0.02 and 0.2 per image for soft-tissue lesions and microcalcifications, respectively. This hybrid approach of interactive and traditional prompts will likely reduce the number of potential cancers that are prompted but overlooked, with a small increase in false prompts.

A similar result for sensitivity between no-CAD and CAD was observed in the high prevalence condition in the study by Drew et al. (2020). Since they did report a small benefit in sensitivity at low prevalence (10%), it may be that if we had run our experiment at this level then we would also have observed an effect. There were some differences between the two studies in terms of the stimuli that may not make this the case, particularly the fact they had clear distractors (Ls amongst target Ts), and so participants had clear potential regions to query, for example. Additionally, they did not use prompts with confidence scores, which is further from clinical examples of interactive CAD in mammography. We demonstrated in our study that the proportion of prompts acted on scaled with the confidence score. However, in the study by Drew et al. (2020), by making the CAD output ‘likely a target’ or ‘not likely a target’ it may force the participants decision either way: mark or not mark as a target, respectively. At a low prevalence, with a high number of the prompts being false prompts (90% had a CAD output of ‘not likely a target’) they will probably dismiss those and then when they come across the few true prompts (75% with a CAD output of ‘likely a target’) they are more likely to accept those and mark them. This may explain the benefit of using CAD at low prevalence.

In our experiments, participants were in control of how they were receiving the information from the CAD system, so despite there being a greater number of false prompts available than a traditional CAD system, they only saw a fraction of these. The CAD ‘signal’ (ratio of true to false prompts of the CAD system) in this study was stronger than the traditional CAD used in the study described in Chapter 6 at 1:1.36 versus 1:1.56. Consequently, participants were probably more likely to believe prompts when they came across them after querying a region with one available. This was evident when looking at the proportion of false prompts acted on across the three CAD conditions in Experiments 1 and 2, which was 68.0%, compared to just 14% in the previous study with traditional CAD. However, in Experiment 1, the specificity with CAD did not change compared to the no-CAD condition, which contrasts to the previous experiment where it was significantly reduced. Importantly, even if trust in prompts is higher and participants are marking a higher proportion of false prompts that they come across, they encountered a much lower number (a median of between 10 and 11 were fixated in this study versus 48 in the previous), therefore the overall false positive rate does not change.

Furthermore, the regions that participants queried that did not contain a prompt will also have been used by participants to make informed decisions. Across the three CAD conditions, participants queried a median of 107.5, 96, and 106 regions without a prompt available, marking only 3.0% of these on average. When a participant queried one of these regions, suspecting it might be a target region, they then received no prompt and may have used it as a reassurance that it was not a target. As a result, they may have decided not to place a marker on that region and therefore reduced their false positive response rate.

The most interesting results from this study are how participants interacted with the prompt confidence values and image scores, rather than the overall impact of the CAD on their performance. The relationship between the proportion of targets queried and the target detection rate is difficult to predict. You might expect that participants would query the harder targets more, since they would require help with these more, but of course they need to find those targets first. On the other hand, for the 'easier' targets (those with >70% detection rate), there was a large variation in how many participants were querying them. These are targets that participants are more likely to actually find and therefore more likely to query, however, they are also more likely to be confident enough not to need the help of CAD.

Experiment 2 investigated the effect of the addition of an overall image score. Since CAD-alone did not lead to an improvement in sensitivity but clinical examples of interactive CAD used an overall score in addition to prompts and confidence values, we hypothesised that the score may have been the missing factor. However, as discussed above, this was not found to be the case. Sensitivity also did not change significantly for images with a score of 10 (highest score) between the CAD+Score and CAD-only condition. We did not have sufficient data in the remaining image score categories for target present data to calculate sensitivity changes for those, so we will focus on images with a score of 10. A possible explanation for the lack of benefit of the image score in terms of sensitivity could be that in total there were 34/100 images with a score of 10 (28 target present and 6 target absent). With an average sensitivity of 65.6% in Experiment 2 for target present images with a score of 10, participants will have only believed there was a target in 54% of the 34 images with a score of 10, on average. Therefore, there may have been a sense that the scores were not as reliable as they actually were.

The image score did affect observer behaviour in terms of both time spent on images and the false positive response rate. Trial time was correlated with image score and was consistent with the results of a clinical study (Rodríguez-Ruiz et al., 2019a), in which, cases with a score of $\leqslant 5$ reduced reading time by $>5\%$ compared to unaided reading, and cases with a score of 9 or 10 increased reading time by $>5\%$. In our study, the prioritisation of high-scored images had a detrimental impact on performance in terms of the false positive rate, with a $>34\%$ increase for images with a score above 7. The higher scores made participants search for a longer period and in many cases may have given them a feeling that there must be a target in the image to find, resulting in them making false positive errors.

Prompt confidence and image score had a clear impact on observer behaviour individually, and we might have expected that the combination of the two in Experiment 2 would create a stronger CAD signal when a prompt was present. For example, an image score of $\leqslant 2$ and a prompt confidence $\leqslant 20$ would have the lowest proportion of participants acting on those regions, and an image score $\geqslant 9$ and prompt confidence $\geqslant 81$ would have the highest. However, this was not what was observed. When a prompt was available, it was the prompt confidence was the deciding factor in how participants chose to act on prompts, with no relation to image score for a given prompt confidence band. In the CAD+Score condition, image score played a key role in guiding overall behaviour, allowing participants to choose which images to pay more attention to. However, when prompts were available, image score was no longer a factor in decision making. For example, $>88\%$ of prompts were acted on in images with a score of 1 or 2.

7.5 Summary

This chapter described two experiments replicating interactive CAD to investigate how, if at all, it benefited readers in detecting targets in a visual search task. Our other aim was to measure how prompt confidence values and image score affected observer behaviour in terms of how they reacted to prompts. Both experiments involved participants searching for microcalcification clusters in 1/f noise distributions while their eye movements were tracked throughout.

Experiment 1 had a no-CAD and a CAD condition, and Experiment 2 had a CAD and a CAD+Score condition. In both experiments, there were 100 images in each condition. In the CAD conditions, participants would query regions by clicking areas they were thought were potential targets, and if available, a prompt would be displayed. Target prevalence was 40%, the same as in Chapter 6, 90% of targets having a prompt available, and 2 false prompts per image available. In the CAD+Score condition of Experiment 2, the majority (70%) of target present images had an image score of 10, while for target absent images the scores were evenly distributed amongst the images.

CAD did not improve sensitivity in either experiment. This is likely due to participants not seeing a large proportion of the true CAD prompts. This also applied to the false prompts, meaning that the false positive rate remained unchanged between the unaided and CAD condition in Experiment 1, in contrast to the results obtained with a traditional CAD approach in Chapter 6. Additionally, only a small proportion of targets that were fixated but not detected were queried. Therefore, the fact that these targets were not detected was not a result of participants dismissing true prompts but more likely that they did not recognise them as a potential target.

The higher the prompt confidence value, the more likely a participant was to act on it. This tendency was stronger for true prompts than for false prompts. In Experiment

2, image score was correlated with time spent viewing the images and the false positive error rate. However, when a prompt was available, it was the prompt confidence that was the key factor influencing the participant to act on the prompt rather than the image score.

Interactive CAD switches the dynamic between the reader and the computer output. Information is only given to the reader when they ask for it, and they are not overloaded by false positives from the system. Medical images with clear anatomical structures operated by expert readers, as opposed to synthetic images and non-expert readers, may be better suited to measuring the efficacy of interactive CAD since they will have an understanding of potentially suspicious regions in the images and thus where prompts are more likely to be. The image score appears to be a useful tool in guiding readers to pay more attention to certain cases, but in our study did not affect their choices on individual prompts.

To study the effects of CAD systems on the performance and visual search of expert medical readers, an eye tracking setup must be extended to a full radiology workstation with dual-monitors. This will be discussed further in Chapter 8.

Chapter 8

Dual-screen eye tracking

8.1 Introduction

Eye tracking is a useful tool to gain a deeper understanding of the search of mammograms and to investigate the effects of various interventions. Many studies are limited to using a single monitor with a single image view and therefore do not mimic realistic clinical conditions. Moreover, eye tracking can be expensive, typically costing more than £10,000 for commercial hardware and software packages. There is therefore a need for a relatively cheap and easy-to-use solution that allows dual screen eye tracking.

8.1.1 Eye tracking in mammography

Eye tracking has been used extensively for a variety of visual search experiments in the field of mammography (see Section 5.4.1). For instance, Kundel and Nodine (2004) recorded gaze positions of readers to allow them to formulate a model of the search of mammograms: a global analysis of the entire image followed by a closer examination of regions identified as suspicious, and this global approach was shown to be more proficient than a 'search-to-find' technique where reader gaze jumps around the image to inspect potentially abnormal regions (Kundel et al., 2007). Fixating near an abnormality has been shown not to guarantee that readers will fixate it after; when initially fixating 2° away, only in around 33% of cases was it fixated on the next saccade and in 50% of cases within the next three saccades (Wolfe et al., 2021b). Radiologists were found to differ in their time to hit the first lesion and their median dwell time on cancers between digital mammograms and digitised screen-film mammograms (Mello-Thoms, 2010). They also tend to make longer fixations and cover a smaller area of the image with digital breast tomosynthesis vs full-field digital mammography (Aizenman et al., 2017).

Many studies have investigated the use of eye tracking data to influence gaze behaviour either through training or interactively. Nodine et al. (2001) fed back to radiologists the areas of the image that they fixated for $>$1s but had not acted upon initially, and reported an improvement in performance. There have been interesting methods of combining radiologist gaze data with both their decisions and CAD output (Tourassi et al., 2010) and with image texture characteristics to predict diagnostic error with a machine learning model (Voisin et al., 2013a). Gaze data has also been used as an input for a mass segmentation approach (Ke et al., 2012) and as a potential tool to train less-experienced readers (Chen and Gale, 2010*b*). These studies were all constrained by the use of a single monitor where readers are limited to a single mammographic image. These single screens are half the size of what would usually make the full display and not a large single screen as described by Krupinski (2016). This is not reflective of clinical practice, where large dual screens are used. It is therefore desirable to work with a more realistic dual-monitor setup.

8.1.2 Eye tracking with dual screens

Mello-Thoms and Gur (2007) compared remote and head-mounted eye trackers for observer studies with radiologists and found that restricted head movements (from the use of a chin rest) with the remote system made them less natural than wearing a head-mounted system. Although head-mounted systems can cause fatigue when used for long periods of time, recent developments mean trackers are becoming less restrictive and more lightweight, so this factor is now less of an issue. In fact, there have been a number of observer studies that tracked gaze across dual clinical radiology monitors via head-mounted trackers with scene cameras to compare search patterns of radiologists (Mello-Thoms, 2008), investigate the effect of dual screens versus single screens on reader search (Krupinski, 2016), explore the use of radiologists gaze to inform computer-aided diagnosis software (Gandomkar et al., 2017), and measure the impact of interruptions on the visual search and accuracy of radiologists reading chest CTs and radiographs (Drew et al., 2018). These studies used eye trackers that were widely commercially available, but the authors did not give a detailed description of the accuracy of the systems.

Another method for dual-screen eye tracking was presented by Dong et al. (2016). They used three remote eye trackers: one placed under each screen and another in the middle. These trackers allowed free movement of the head. The study also used a screen capture tool to monitor zooming and panning. However, the trackers used were not capable of direct on-screen tracking, and so a virtual model of the two monitors had to be created. Furthermore, the screen capture tool used to acquire the zoom and pan information reduced the overall frequency of the tracking system from 60Hz to 10Hz.

In order to choose the eye tracker for dual-screen tracking, we surveyed various head-mounted eye trackers, shown in Table 8.1. The typical accuracies for these trackers range between 0.25° and 1.0°. Though, this is accuracy under controlled tests and not necessarily the accuracy that would be achieved across two large screens. Each tracker offers different solutions for obtaining the coordinates on screens rather than from the reference frame of the scene camera, some of which are discussed further in Section 8.1.3. The Pupil Core glasses were chosen to be used for tracking on our dual-screen setup since the specifications were comparable to other trackers and offered open-source analysis software with useful plugins for on-screen tracking.

8.1.3 Pupil Labs glasses

We investigated the potential of Pupil Labs eye tracking glasses, referred to as Pupil Core and discussed in Section 5.3.2, as a relatively inexpensive solution to dual-screen eye tracking (Kassner et al., 2014). Similar to other eye tracking glasses, they are lightweight (22.75g), and therefore are unlikely to cause fatigue for long reading sessions. Eye tracking glasses allow observers to move as they would in normal viewing conditions, which is particularly important for viewing large clinical displays.

With screen-based experiments, eye tracking glasses pose the problem that the gaze coordinate system is given in reference to the coordinate system of the scene camera. As the participant is free to move, the position of the screen is ever-changing, often referred to as a dynamic region-of-interest (Ooms et al., 2014). There have been a number of proposed methods to map gaze coordinates onto these dynamic ROIs, often involving extensive manual work such as correcting the position of objects in the recording frame-by-frame (Tobii Pro, n.d.). A similar approach is used with SMI's BeGaze software, where the position and size of ROIs needs to be altered manually

Table 8.1: Specifications of commercially available eye trackers. N/A = not available.

Eye tracker	Monocular or binocular	Sampling frequency	Range	Typical Accuracy	Scene camera quality
Pro Glasses 2 (Tobii Pro, 2018)	Binocular with auto parallax correct	50 or 100Hz	82° horiz 52° vert	N/A	HD camera Resolution 1920×1080 25 fps
Eye Tracking Glasses 2 (SMI, 2017*b*)	Binocular with auto parallax correct	60 or 120Hz	80° horiz 60° vert	0.5°	HD camera Resolution 1280×960 24 fps
ViewPoint EyeFrame Scene Camera (Arrington Research, n.d.)	Binocular with auto parallax correct	30 or 60Hz	56° horiz 42° vert	0.25 – 1.0°	N/A
Mobile Eye-XG Glasses (ASL, n.d.)	Monocular, right eye	30 or 60Hz	60° horiz 40° vert	0.5-1.0°	N/A
ETMobile Glasses (Argus Science, n.d.)	Monocular, right eye	30 or 60Hz	60° horiz 40° vert	0.5°	Resolution 1600×1200
EyeLink II (SR Research, 2021)	Binocular	250 or 500Hz	40° horiz 36° vert	0.25 – 0.5°	250Hz sampling rate
Pupil Core (Pupil Labs, 2021*a*)	Monocular or binocular	200Hz	100° horiz 74° vert	0.6°	30Hz@1080p 60Hz@720p 120Hz@480p

throughout the video recording when necessary (SMI, 2017*a*). BeGaze also uses 'semantic gaze mapping', which is able to detect when a reference image of the scene is visible in the recording, and subsequently map the gaze data onto that reference image. However, this process is not automated and requires each fixation to be manually mapped by selecting its position on the reference image according to where the fixation appears in the recording (SMI, 2017*a*).

There are other, more automated solutions, such as DynAOI (Papenmeier and Huff, 2010), but as with the method proposed by Dong et al. (2016) this requires the creation of a 3D model of the experimental setup. Pupil Labs instead use a surface tracking plugin that allows the user to define 2D surfaces within recording by displaying QR-like markers, e.g. around the edges of displays, which act as reference points for the surfaces. These surface definitions can be loaded into any recordings which feature that same marker setup and automatically track that surface in the scene video, but they may require some level of manual correction to make sure they are correctly defined within each recording.

A recent study by Ehinger et al. (2019) performed a detailed comparison between the EyeLink1000 desktop eye tracker (see Section 5.3.1) and Pupil Core, and set out a comprehensive testing framework for eye tracking devices. Overall, they found Pupil Core to have a worse spatial accuracy compared to the EyeLink1000, 0.87° vs 0.52°. They also reported a strong initial decay in the calibration accuracy of Pupil Core of 0.25° after 4:42 mins, although this did stabilise after that point where the accuracy only dropped a further 0.04° at 6:16 mins. Their analysis was done on a single 24-inch monitor and only compared the accuracy at 2 time points after the initial calibration. While this provides us with a good reference, we required greater temporal resolution for our analysis, as well as an assessment of performance on a setup with a larger FOV.

In this study, we explored using Pupil Core for a search task on a clinical radiology setup with non-expert readers. We were primarily interested in how the spatial accuracy of the system varied over the duration of the experiment and across the display, to allow us to make recommendations for future experimentation with similar setups. This setup was planned to be used for an experiment with expert readers (see Chapter 10), so it was important to measure the performance of the device and recommend the time period for recalibrations, for example. This methodology can be adapted to a variety of large display setups and is therefore of interest to other research groups both inside and outside of radiology. Our secondary aim was to determine the detectability of our simulated masses in synthetic backgrounds. Firstly, this task is similar in nature to a radiology search task, which makes it suitable for testing performance of the device for future work. Secondly, these mass images were to be used for the experiment in Chapter 9, therefore it was desirable to get a reliable estimate for the detectability of various mass targets that could be used to select appropriate images to use.

8.2 Methods

8.2.1 Materials

We used the Pupil Labs – Pupil Core Headset with 200Hz eye cameras and a 120Hz world camera. A full description of Pupil Core and how it operates is provided in Section 5.3.2. The experimental setup is shown in Figure 8.1, with two side-by-side 21.3-inch Dome E5 displays with resolution of 2048×2560 pixels each. As can be seen in Figure 8.1, QR-type markers were placed around the edges of the monitors. Eight markers were displayed on each screen, separated by 13.4cm (or 18.1cm across the bezel) horizontally and 14.0cm vertically. These markers were used to define surfaces within the Pupil Labs analysis software (Pupil Player) onto which fixations and gaze data are mapped. We defined 4 sub-surfaces that were later combined in the analysis to make one surface for the whole display. This was necessary because the fish-eye lens of the scene camera means that using one large surface covering both displays cuts off large portions of the screens. By using four surfaces this effect can be minimised.

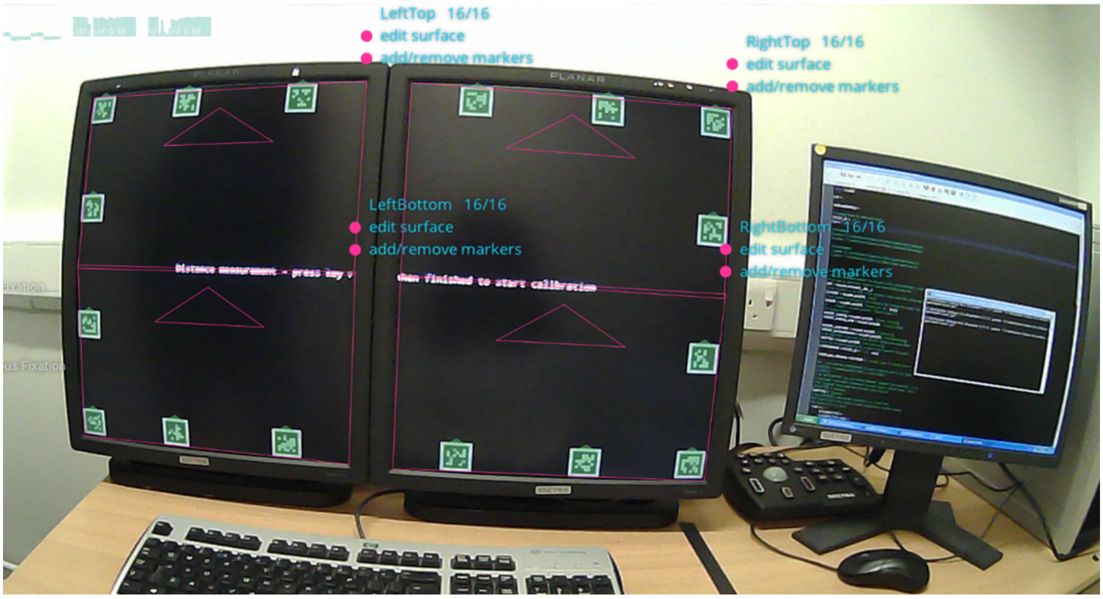
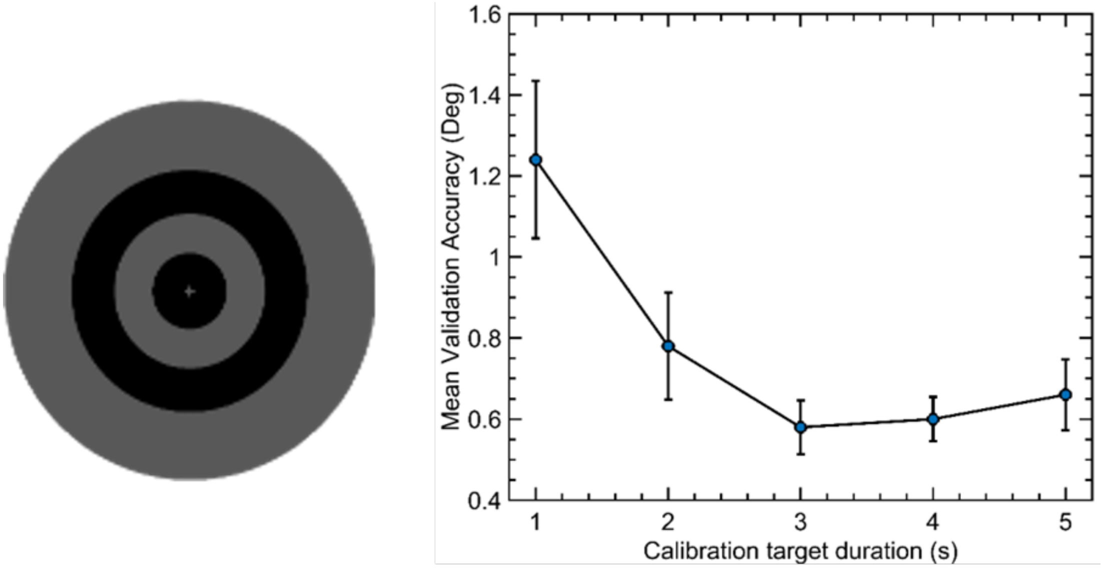


Figure 8.1: Experimental setup as seen from the scene camera of Pupil Core. Four surfaces were defined, each covering one quadrant of the dual display. The scene camera uses a fish-eye lens meaning straight lines can appear curved, so using four surfaces instead of one reduced the amount of screen that was cut off by non-curved edges of the surfaces. The lights were switched off during the experiment to mimic the low lighting conditions used by radiologists reading mammograms.

8.2.2 Calibration procedure

Calibration involved displaying the calibration target (Figure 8.2a) sequentially at 14 different positions, where the start and end point were set at the top left of the displays with other positions randomised. Preliminary tests were completed to determine the appropriate length of time to display the calibration target at each position during the calibration procedure. This is important as it needs to be long enough for the analysis software to detect where the target is. At the same time, it should be as short as possible so participants remain focused for the whole calibration process. Figure 8.2b shows the result of changing the presentation time in 1 second increments. This test was completed by a single researcher, five times for each target duration. Since we wanted to minimise the time taken to calibrate whilst maintaining accuracy, 3 seconds was chosen as the most suitable target duration time, since this is where the validation accuracy plateaued, measured using Pupil Labs' 'Pupil Player' analysis software.



(a) (b)

Figure 8.2: (a) Target used for calibration and validations (on-screen radius = 1.9cm) and (b) validation accuracy for different display lengths of time for calibration targets (time per target).

The monitors used are high contrast (typical brightness of 800 cd/m^2) and as a result, when displaying the surface markers and calibration targets, the white portions of those became too bright to be reliably detected by the software. This was worsened further when the lights were switched off to mimic the low lighting conditions used by radiologists reading mammograms. To compensate for this, the white parts (RGB = [255, 255, 255]) were made grey (RGB = [127,127,127]) so that they could be accurately detected.

8.2.3 Calculating the useful field-of-view

The useful field-of-view is an important factor in eye tracking, and can be used to calculate, for example, the proportion of an image that has been covered or whether certain features have been examined. Since the size of the UFOV in pixels on the screen depends on the distance of the participant from the screen, this has to be constantly measured in experiments with wearable eye trackers where the participant can freely move towards and away from the screen. Pupil Labs software has a head-pose-tracking plugin, meaning the position of the observer is tracked relative to reference markers – the same ones that are used for the surface tracking. At the start of the experiment, the participant's distance from the screen was measured with a tape measure and compared to the z-coordinate for that time measured by the software. Using this reference, the z-coordinates are converted to cm for calculation of the UFOV for the remainder of the experiment using the formula $UFOV$ (in cm) $= 2d\tan(\frac{VA}{2})$, where d is the distance to the screen in cm, and VA is the fixed visual angle of 5° diameter, commonly used for mammography visual search studies (Kundel and Nodine, 2004). The UFOV was then converted to pixels using the pixels per cm of the displays.

To check the validity of this approach, a test was conducted by placing the glasses on a chin rest at 3 different distances. To test the variability of the head-pose tracking, at each distance, the glasses were picked up and moved around before placing them back at the same distance from the screen. The results from this manipulation are given in Table 8.2. The variation in the measurements of 1 unit of z per cm results in an error for the size of the UFOV of around 8.5%. This is a reasonable size of error since it provides a much more accurate estimation of the UFOV than if we were to estimate a fixed distance for each participant.

Table 8.2: Results from the head pose tracking technique to measure distance to the screen. Each z measurement was repeated 5 times and the mean and standard error is given. The units of z per cm were used to calculate the size of the UFOV in pixels on the screen throughout the experiment.

Distance (cm)	**Mean z**	**1 unit of z in cm**
36.0±0.1	9.49±0.09	3.79±0.04
82.5±0.1	20.26±0.01	4.07±0.003
113.5±0.1	30.25±0.04	3.75±0.01

8.2.4 Stimuli

The images used in this study were similar to those used in Chapters 6 & 7, but with different targets. Synthetic 1/f noise images were created in MATLAB using open-source code (Methven and Qi, 2018), with a spectral roll-off factor (image roughness) of 1.5. These backgrounds resemble the glandular component of mammograms, although they lack linear structures. The targets used in the experiment were Gaussian blobs, which resemble masses in mammograms. An example experimental image is shown in Figure 8.3. As shown in Table 8.3, the masses varied in size, shape (either circular or elliptical), and contrast (high or low) with their surroundings. Images were sized at 2049 × 2049 pixels, centred between the two displays. Care was taken to ensure no mass targets were split across the displays. The 120 experimental images were split into 3 sets of 40, each containing 20 simulated masses - a prevalence of 50%. The experiment code was written using PyGaze v0.6.0 (Dalmaijer et al., 2014), with a PyGame back-end (v1.9.2).

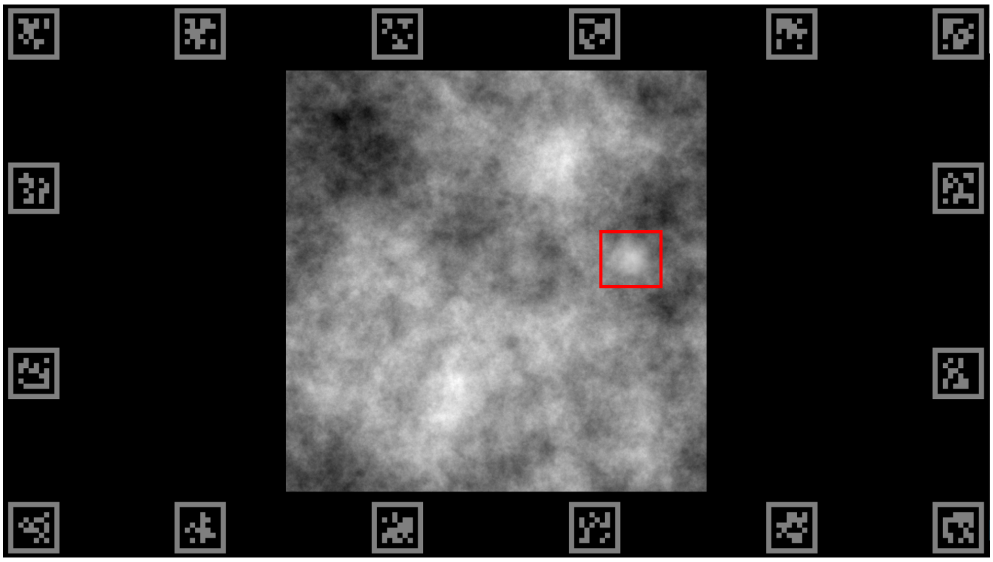


Figure 8.3: Example experimental image positioned on display, the mass signal is outlined by the red square for illustrative purposes.

Table 8.3: Size and brightness factors, and number of each of the eight target categories. For the mass targets, the size and brightness was selected randomly using the factors given in the table for each category. These factors have been normalised using the maximum value. The size refers to the diameter of the circles or the size of the major axis for the ellipses with aspect ratios ranging between 1.5 and 3. The number of targets of each category used in the experiment is given in the bottom row.

	Circular High Contrast			**Circular Low Contrast**			**Elliptical**	
	Small	**Medium**	**Large**	**Small**	**Medium**	**Large**	**Small**	**Large**
Brightness	0.78−1	0.78−1	0.78− 1	0.56−0.78	0.56−0.78	0.56−0.78	0.78−0.89	0.78−0.89
Size	0.13−0.31	0.36−0.51	0.56−0.75	0.13−0.31	0.36−0.51	0.56−0.75	0.25−0.50	0.50−1
No. targets	8	8	8	8	8	8	6	6

8.2.5 Experimental procedure

Each participant first completed 16 training images which familiarised them with the experimental interface – clicking to place a marker where they believed there was a target and right-clicking to remove a marker if they changed their mind. The training set contained 15 target-present images and 1 target-absent image. When viewing the target-absent image, participants were informed that only 50% of the experiment images would contain a target. During training, participants were given feedback on their clicks which informed them if they had found a simulated mass lesion and if so, highlighted the whole mass. Once they had completed the training set, the glasses were set up by adjusting the cameras to their eyes and setting the ROI window in the Pupil Capture software to cover the whole eye as described by Pupil Labs (2021*b*) and in Section 5.3.2. This took around 1 minute to setup. At the start of each run, the distance from the participant to the screen was measured and registered for that time point by pressing a button. These measurements ranged from 51–76cm (median 61cm), but participants were free to move, so this is only an estimate of the distance during the experiment.

The experiment was split into 3 runs. Each run started with a calibration sequence of 15 targets, where each appeared for 3 seconds. The positions of the calibration targets are shown in Figure 8.4. The position of the first and last target to appear was always the top-left, which is the same as the Pupil Labs default calibration procedure. The order of the other target positions was randomised. After this, an initial validation sequence was completed, which was the same as the calibration, but targets only appeared for 2 seconds each. A shorter time was desirable to minimise time spent validating but still allowing sufficient time for multiple fixations on the target after initial correction saccades for calculation of accuracy.

It is important to note why we are able to use a shorter time for validations compared to calibrations: calibrations were done using Pupil Player software which needs to detect the location of the targets using the scene camera – the time needed for calibration targets to be displayed for accurate results is 3s, as shown in Figure 8.2b. However, for validations we calculated the accuracy with our own software post-hoc where we knew exactly where the targets were displayed and for how long. We used the fixation data to calculate validation accuracy, which was possible to do with a target duration of 2s. Search of the experimental images was conducted with the lights turned off to more closely replicate conditions in the clinic and thus test performance of the glasses in low light conditions. During the experiment, after 1 minute had passed from the end of the previous validation sequence, the validation was repeated. The order of the experimental images was randomised for each run.

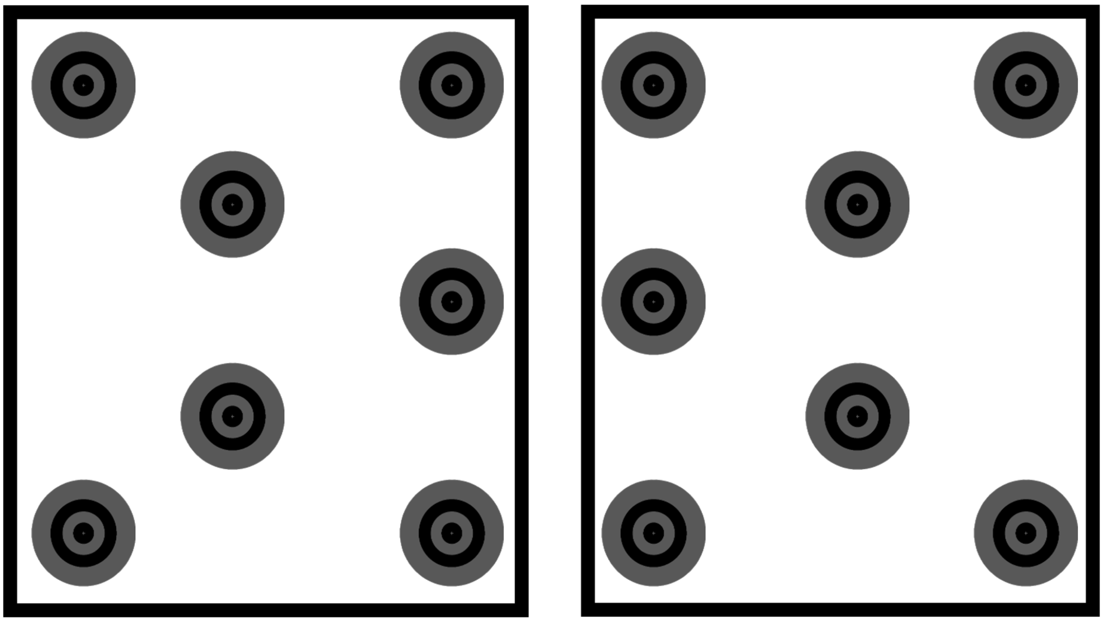


Figure 8.4: Schematic indicating the positions of the targets used for calibration. The rectangles represent the screen edges (not to scale).

8.2.6 Data analysis

There were two stages to our data analysis. The first stage is outlined in Figure 8.5, which details how the raw data from Pupil Capture was analysed in Pupil Player. The first step was calibration, referred to as 'offline calibration' since it is undertaken post-hoc, using the reference calibration targets in the video. This requires time points to be selected where the calibration has taken place and where the gaze should be mapped to. Marker detection is an automatic process which identifies and locates the surface markers displayed in the video footage; these are then used for surface and head pose tracking. Surface definitions were performed in Pupil Capture and were loaded into Pupil Player. If the corners of a surface did not align with the edge of the display or at the halfway point of the screen (a horizontal line was displayed across the screen to measure this), they were adjusted in Pupil Player. The four surface definitions used in this study are shown in Figure 8.1, where each covered a quadrant of the dual-screen display. The size of the surfaces were set in Pupil Player to be half of the size of a display, in pixels. Therefore, calculated gaze coordinates on the surfaces would match the screen coordinates in pixels. After they were defined, they remain fixed in place, around the markers used to define them, while the FOV changes according to head position. On export, the fixations and gaze positions on each surface are automatically generated. Additional files include the positions of each surface in the world coordinate system for every frame and the time points that any surface entered or exited the FOV.

The head pose tracking and fixation detection are also automatic processes, which require the researcher to set a reference marker for the head pose tracking, and a maximum dispersion threshold (1.5° visual angle) and minimum and maximum duration thresholds (100ms and 800ms, respectively) for fixation detection. The annotations were added to mark the time point where the distance measurement occurred, as well as the start of each validation sequence. The data were then exported to be analysed in Python.

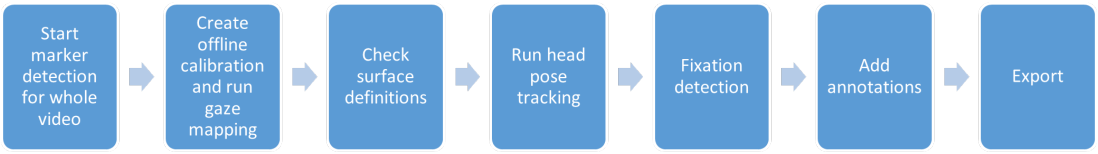

Start marker detection for whole video
Create offline calibration and run gaze mapping
Check surface definitions
Run head pose tracking
Fixation detection
Add annotations
Export

Figure 8.5: Flowchart of the data analysis in Pupil Player software. These processes require manual input to select appropriate time points or selection of areas for surface tracking.

The analysis in Python and Matlab is outlined in Figure 8.6. Fixations on each of the four surfaces (shown in Figure 8.1) were first mapped onto a single joint surface covering the whole screen area. Since the surface sizes were defined to match the pixel size of the display, coordinates on the surfaces could be directly mapped to screen coordinates. It was then necessary to synchronise the times between the data saved on the experimental computer and the data exported from Pupil Player. The experiment data includes the time and positions of user clicks, the time of the distance measurement, when validation sequences occurred, and the start time and duration of different images. To synchronise the times from the experiment computer with the time recorded by Pupil Capture, we compared the time of the onset of the first validation sequence recorded on both computers. Once the times were synchronised, fixations were assigned to either a specific image or validation sequence. From here, for individual images, the image coverage, number of fixations and saccades, and target dwell time were calculated. For validation sequences, the median distance from validation targets was calculated, in addition to individual accuracies for each target. This was done by taking the median distance between the fixation and the validation target across all fixations on that validation target. To calculate the distance between the fixations and validation target, we used Equation 8.1, outlined in Ehinger et al. (2019):

$$\theta = \arccos\left(\frac{f \cdot t}{\|f\|\,\|t\|}\right) \tag{8.1}$$

where θ is the angular distance in radians (converted to degrees), the fixation position $f = \binom{f_x}{f_y}$, and the validation target position $t = \binom{t_x}{t_y}$. In addition to accuracy, we also calculated the angular precision of the system using Equation 8.2, which calculates the standard deviation of fixation positions on a validation target:

$$\theta_{sd} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d\left(\binom{x_i}{y_i}, \binom{\bar{x}}{\bar{y}}\right)^2} \tag{8.2}$$

where n is the number of fixations on a validation target, $d\left(\binom{x_i}{y_i}, \binom{\bar{x}}{\bar{y}}\right)$ is the distance between a single fixation and the mean location of fixations on the validation target (Ehinger et al., 2019). Finally, data were analysed in Matlab to produce figures and overviews of data across participants.

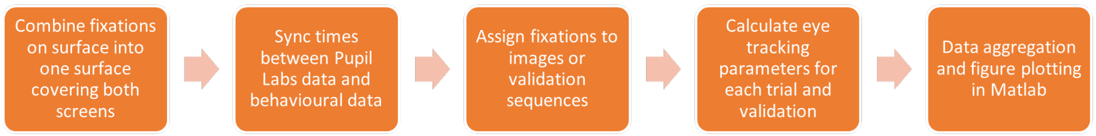

Combine fixations on surface into one surface covering both screens
Sync times between Pupil Labs data and behavioural data
Assign fixations to images or validation sequences
Calculate eye tracking parameters for each trial and validation
Data aggregation and figure plotting in Matlab

Figure 8.6: Flowchart detailing the analysis of the data that is exported from Pupil Player and analysed in Python and Matlab.

8.2.7 Participants

Twenty-six participants (median age 19, age range 18-23, 23 female) were recruited, and informed consent was obtained. Twenty-five of the participants were undergraduate psychology students and received course credit for taking part, and the other participant was a member of university staff who received no compensation for participation. The study was approved by the University of Manchester Research Ethics Committee (2018-4586-6410).

8.3 Results

The data presented comes from 26 participants, with a combined number of 67 ‘runs’, i.e., sets of 40 images completed (20 target present and 20 target absent). The missing data (six runs from set 1, three from set 2, and two from set 3) were caused by experimental issues, such as improper positioning of the scene camera. These missing data were from individual runs across 8 participants, and all runs for 1 participant whose behavioural data (clicks) was still used in the analysis. We will be referring to the spatial accuracy of the glasses as measured by repeated validation sequences as the ‘validation accuracy’. The spatial precision calculated as the standard deviation of fixation positions (Equation 8.2) will be referred to as the ‘validation precision’. We also looked at the reduction in measured accuracy over time and will refer to that as the ‘calibration decay’ to be consistent with literature on this topic (Ehinger et al., 2019).

8.3.1 Target detectability

One of the aims of this study was to test the detectability of our mass signals in the simulated images. The masses used were categorised based on their shape (circular or elliptical), brightness, either high contrast (HC) or low contrast (LC), and size (small, medium, or large), as shown in Table 8.3. The mean detection rate, i.e., the percentage of participants that successfully located the mass, is given for these 8 categories in Figure 8.7. All but one of the differences in detection rate between the categories were purely numerical, with the exception of the significantly higher detection rate of small circular HC masses compared to large circular HC masses (t(14)=2.73, p=0.02).

The detection rates of individual masses are plotted against their size (diameter of circular masses and size of major axis for ellipses) in Figure 8.8. We tested the negative correlation between size and detection rate. When correcting for multiple comparisons

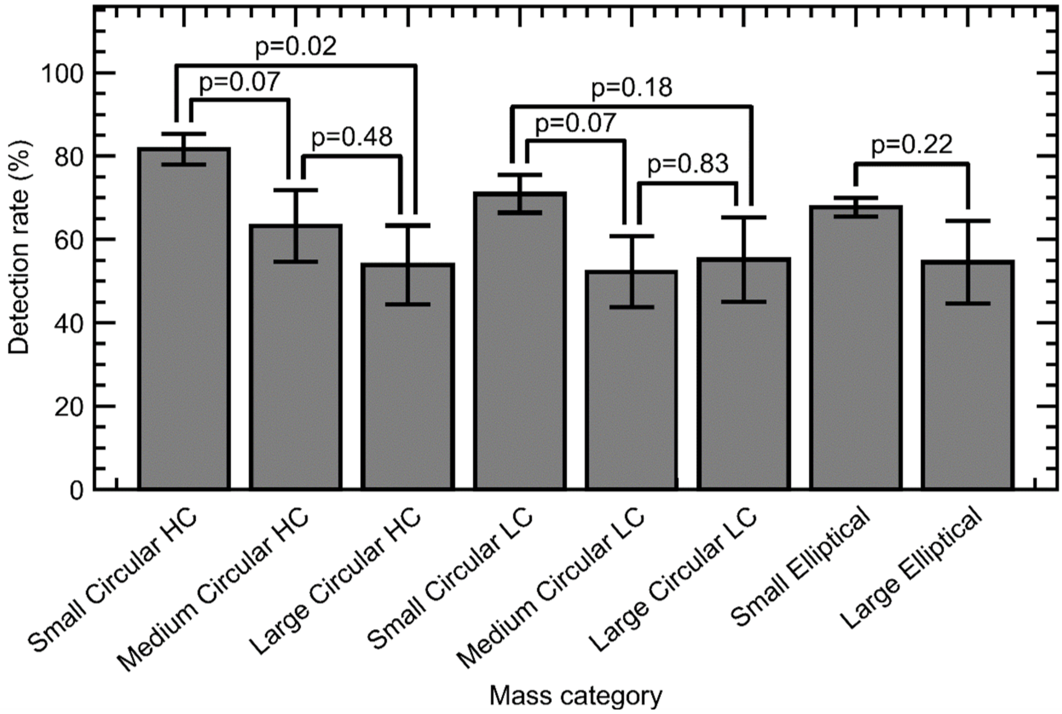


Figure 8.7: Detection rate for different mass categories. There were 8 targets of the first 6 categories (circular masses) and 6 each for the last two categories (elliptical masses).

using a Bonferroni correction (corrected p-value of 0.017), this was not significant for HC masses (r(22)= −0.41, p=0.04), LC circular masses (r(22)= −0.28, p=0.19), or elliptical masses (r(10)= −0.27, p=0.41). Finally, we also measured the overall false positives per image rate, which was 0.7. This suggested that there were features within the image that raised suspicion in addition to masses.

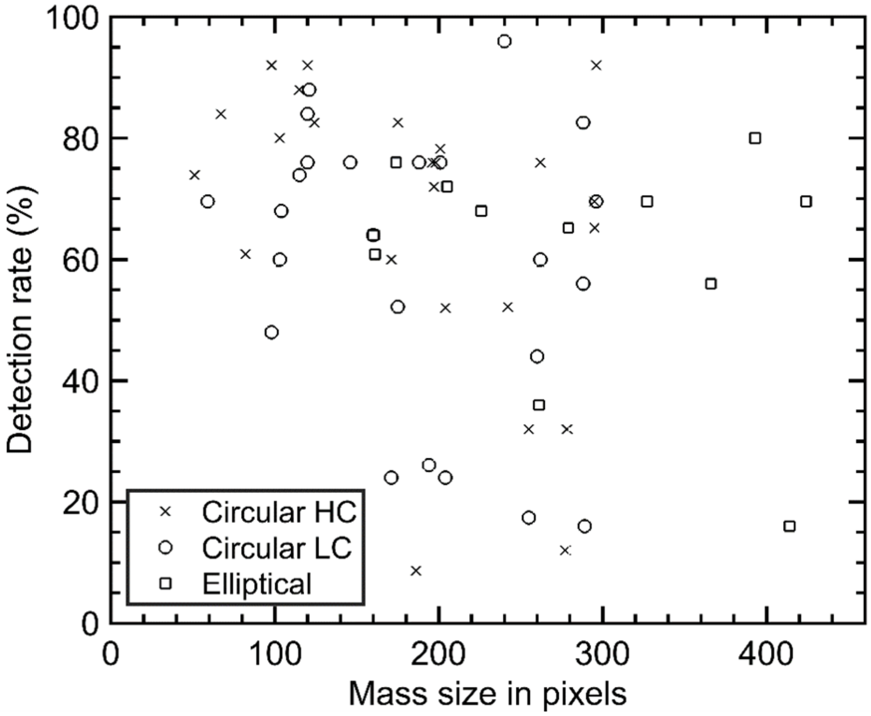


Figure 8.8: Detection rate for individual masses versus their size in pixels. The data is further distinguished by whether they were circular HC, circular LC or elliptical. The size refers to the diameter of the circular masses and the size of the major axis for elliptical masses.

8.3.2 Calibration decay

The calibration decay in the experiment is shown in Figure 8.9. Validation numbers 13 to 15 were excluded from analysis since we only had one participant who completed these. Validation number 1 was done immediately after the initial calibration and each validation thereafter started exactly 1 minute after the previous had ended. Each validation took approximately 33 seconds to complete (including a 3-second countdown), with the corresponding start times shown on the top axis. The median validation accuracy started at 0.67° and plateaued between validation number 4 and 8, where it remained between $1.22°-1.36°$, before it steadily increased from validation number 9 onwards.

The initial median validation precision, shown in Figure 8.10, was 0.37° and remained between 0.33°−0.38° until validation 7 where it increased to 0.46°. The precision did not go above 0.5° for all validations.

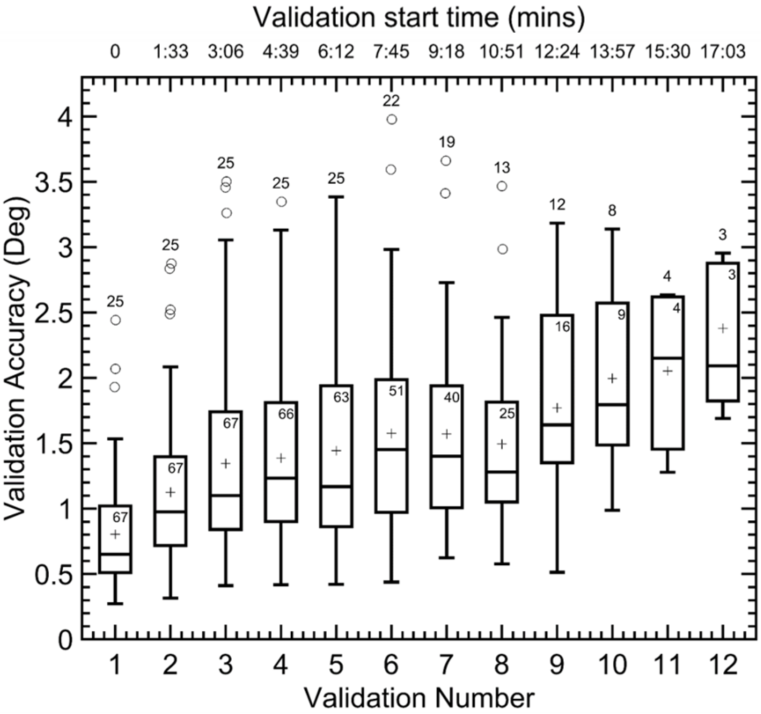


Figure 8.9: Calibration Decay: the validation accuracy across participants as a function of the validation number. The +'s indicate the mean for that validation number. The values in each box are the number of data points (runs) for that validation and above each bar is the number of participants those runs come from.

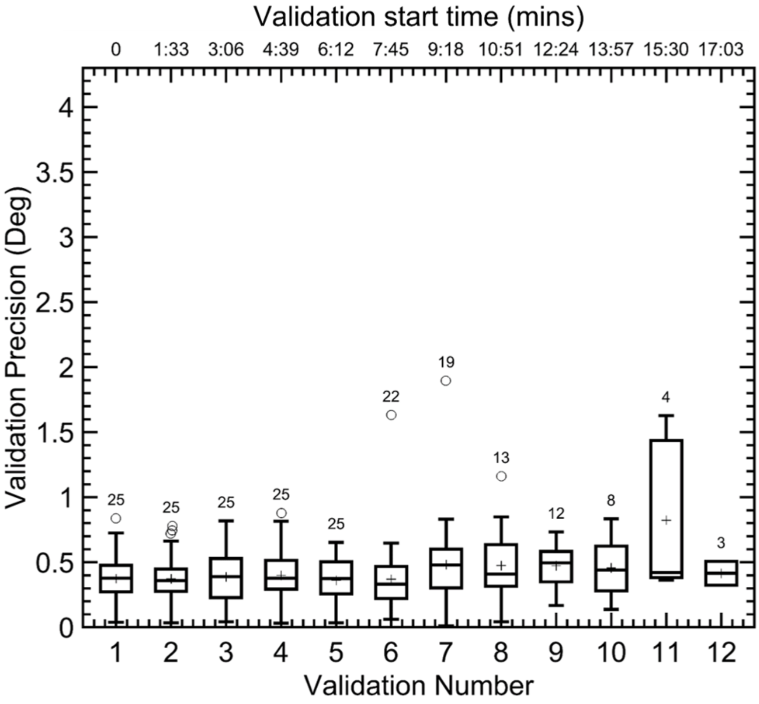


Figure 8.10: The validation precision across participants as a function of the validation number. The +'s indicate the mean for that validation number. The values in each box are the number of data points (runs) for that validation and above each bar is the number of participants those runs come from.

To determine an appropriate point to recalibrate the glasses, we first compared the validation accuracy distributions shown in Figure 8.9 to various reference accuracies (in steps of 0.05°), shown in Table 8.4. For each reference accuracy we found the earliest validation with an accuracy significantly worse than that reference. None of the validations were significantly worse than reference accuracies greater than 1.4°, so we did not include them in the table. Table 8.4 suggests that recalibration should be done around the time validation 4 occurred (4:39 mins to 5:22 mins) to maintain accuracy below 1.25°. This is the point around which the median accuracy remains before a sharp decay at validation number 6 in Figure 8.9. At 6:12 mins (validation 5) the accuracy was significantly worse than 1.25°.

Table 8.4: Comparisons of accuracy distributions to reference accuracies. For each reference accuracy, the lowest validation number with a mean accuracy significantly worse than the reference is stated. The statistics are derived from a bootstrap approach across runs.

Reference accuracy (degrees)	Validation with worse accuracy (start time in mins)	Mean difference [95% CIs]	t statistic	p value	Degrees of freedom
1.00	3 (3:06)	0.33 [0.19−0.48]	3.746	$<$0.001	66
1.05	3 (3:06)	0.27 [0.13−0.43]	3.110	0.003	66
1.10	3 (3:06)	0.23 [0.09−0.38]	2.557	0.013	66
1.15	4 (4:39)	0.23 [0.10−0.36]	2.902	0.005	65
1.20	4 (4:39)	0.18 [0.05−0.32]	2.272	0.026	65
1.25	5 (6:12)	0.20 [0.05−0.36]	2.133	0.037	62
1.30	6 (7:45)	0.28 [0.10−0.47]	2.541	0.014	50
1.35	6 (7:45)	0.23 [0.05−0.42]	2.083	0.042	50
1.40	10 (13:57)	0.54 [0.17−0.92]	2.378	0.045	8

We also looked at decay of calibration using a bootstrap approach where participant runs were treated as a random effect. For each statistical test, the runs were chosen such that we could make pair-wise comparisons between the validations within runs. Figure 8.11a shows that there was an immediate significant calibration decay from validation 1 to 2 by 0.41° (95% CI, 0.24°−0.60°; t(24)=3.684, p=0.001), but there were no further significant drops in accuracy for consecutive validations. Figure 8.11b shows the calibration decay relative to validation 1 for all further validations. The decay is significant across validations 2 to 12, except for validation 11 (a consequence of only having 4 runs at that point).

Although there is an initial significant calibration decay (between validations 1 and 2), it would not be practical to recalibrate after just 1.5 minutes, especially for experiments lasting around 1 hour in total. Therefore, there is a trade-off between the frequency of recalibrations and the spatial accuracy of the glasses. Recalibrating every 5 minutes during an experiment (the point during which validation 4 took place) maintains the mean accuracy below 1.25° and does not severely interrupt the flow of the experiment. The variance of accuracies was greatest at validation 5, so it is also desirable to recalibrate before that point.

The calibration decay varied depending on the position of the validation target. Generally, the validation accuracy was worse for the targets at the edge of the displays, with a median calibration decay across the first 5 validations of 1.06° for the 4 outer corners versus a median of 0.77° for the other positions. The trend in the validation accuracy was similar across all target positions, explored in Figure 8.12 and Figure 8.13, where the mean validation accuracy and precision are plotted for all validation targets, all except the corner targets, and corner targets only. The mean validation accuracy was consistently worse for the corners for the first 10 validations, compared to the other targets. The time which we determined to be the appropriate recalibration point is indicated by the red dashed line. Precision was also worse at the corners of the display but remains around 0.5° before the recalibration point. The median precision across the first four validations for all target positions was 0.37°.

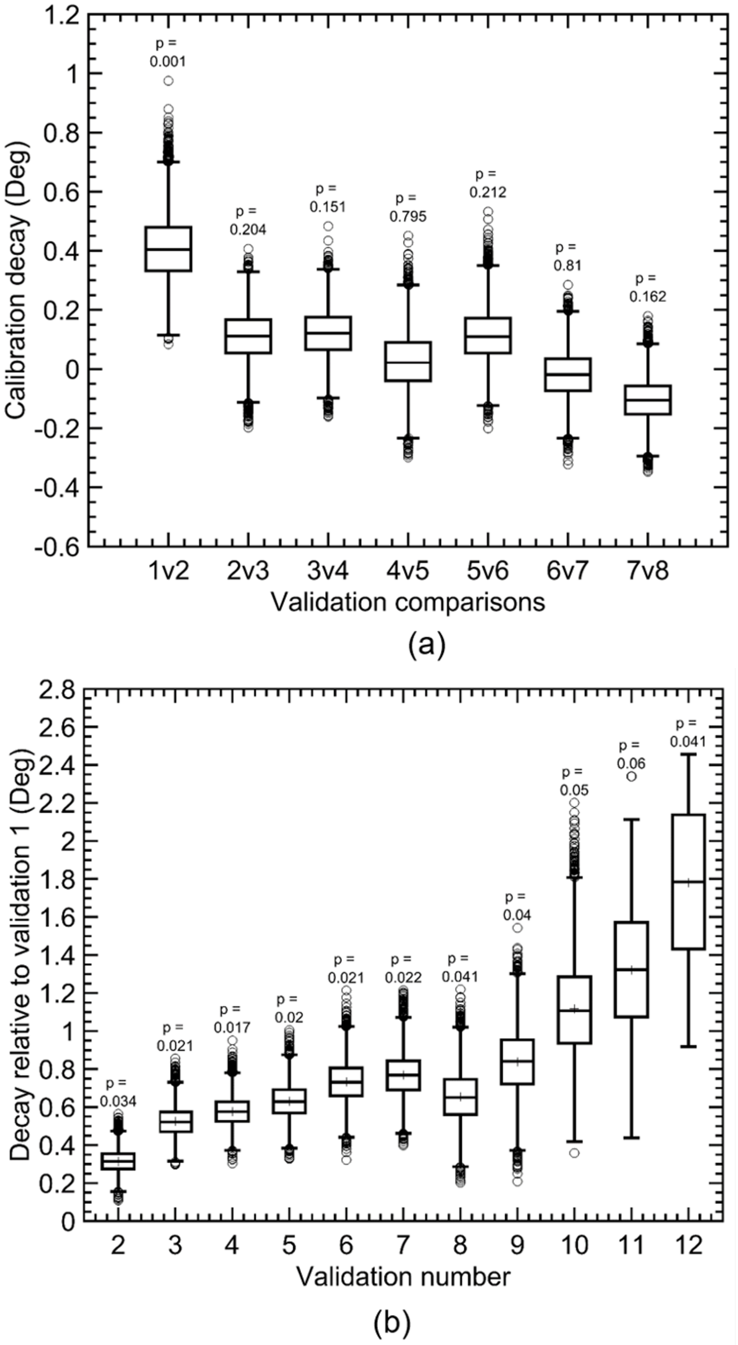


Figure 8.11: (a) Comparisons between adjacent validations. For each comparison, only the 25 runs from 13 participants that had data available for validation 8 were used. (b) Comparisons between validation 1 and validations 2 to 12. For each comparison, only runs from participants that also had data for the later validation were used (i.e. from 25 participants (67 runs) for 1v2 to 3 participants (3 runs) for 1v12). In (a) and (b) displayed above each bar is the p value for each comparison, calculated from a bootstrap approach.

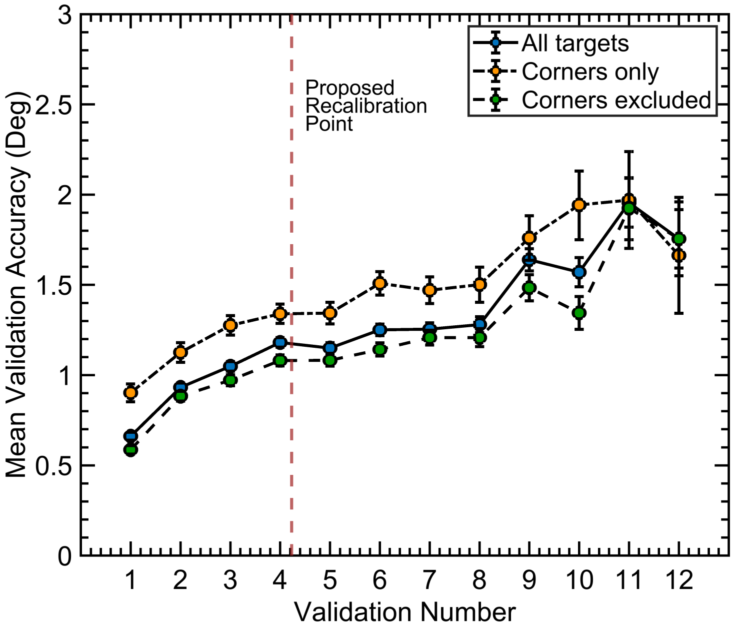


Figure 8.12: Mean validation accuracy across participants' as a function of validation number for all validation targets, corner targets only, and all except the corner targets. The error bars are the standard errors across targets and participants.

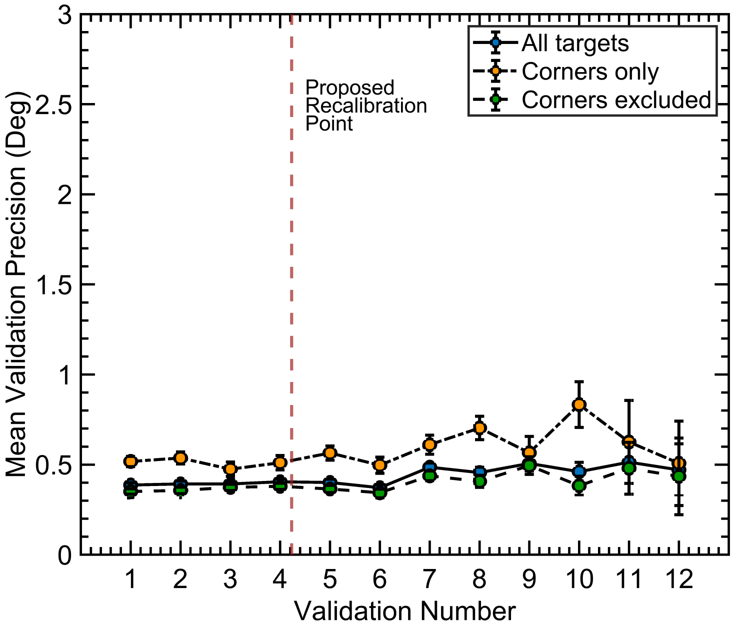


Figure 8.13: Mean validation precision across participants' as a function of validation number for all validation targets, corner targets only, and all except the corner targets. The error bars are the standard errors across targets and participants.

8.4 Discussion

Our investigation of using Pupil Core for gaze-tracking across a clinical radiology workstation has demonstrated the feasibility of this approach. The glasses were able to operate in typical clinical lighting conditions and cope with the high contrast monitors. Validation accuracy remained below 1.4° visual angle for the first 11 minutes of the experiment, after which the accuracy steadily decreased. The calibration decay was significant between the first and second validations, but did not increase significantly for further consecutive validations. Precision remained consistent across validations, starting at 0.37° with no major worsening throughout. Both the accuracy and precision were worse at the outer corners of the displays when compared to the central points, although these points would generally not contain any of the mammograms and are therefore less important for our purposes.

The calibration decay measured in our experiment was consistent with another study using Pupil Core (Ehinger et al., 2019), where there also was a significant early decay before stabilising. We had a higher temporal resolution and also measured the decay over a longer period; we showed there was a point where the calibration accuracy decreases rapidly - after around 12:24 minutes. We also observed a higher initial decay of 72.4% after 4:39 minutes compared to the reported 29.4% in Ehinger et al. (2019) after a similar time period (4:42 minutes). In terms of precision, this was similar in both studies, with 0.37° reported in this study versus 0.31° in Ehinger et al. (2019). Unsurprisingly, the accuracy was worse at the outer corners of the display, which is commonplace even for higher-accuracy desktop eye trackers, where some even weight the overall accuracy towards central points. We did not weight the accuracy and took the median across all targets. This result was due to the fact that during calibration and validation procedures, the participants were instructed to look directly forward and only move their eyes towards the target. Therefore, we were measuring the accuracy as the gaze moved away from the line-of-sight of the environment camera. In normal use, outside of calibration, participants move their heads to look at the outer edges of the display, and so the accuracy would resemble those at the centre of the screen.

There are a number of key advantages to using Pupil Core for our setup. They are lightweight and were comfortable for experiments which lasted as much as 1 hour of wearing. Compared to similar eye tracking devices, Pupil Core is relatively inexpensive in terms of hardware and offers free open-source software for capture and analysis of data. The software offers a variety of useful plugins, such as surface and head-pose tracking that works out of the box when used with the provided digital markers to display on or around regions of interest. The open-source nature also means there are a range of third-party modules developed by the community of users to expand the eye tracking capabilities. Another useful feature is the range of options for calibrating the headset. This can be done using on-screen or printed markers, either before recording or post-hoc by including a calibration procedure within the experimental recording. Through experimentation, we found the offline calibration most suitable for our setup as it produced the most reliable calibrations. There is also an option to adjust eye camera settings post-hoc, including the ROIs around the eyes and the algorithm settings.

The ease of use and freedom of movement do come at a cost, though. The accuracy of the headset, even well within the limits of display size for desktop trackers, is lower than other trackers at $\approx$1.0° vs $\approx$0.5°. Although this is a result of the trade-off between accuracy and display size, it is worth considering when designing experiments. Furthermore, the world camera, which captures the environment, uses a fish-eye lens. This complicates the surface tracking somewhat, by making the edges of the display appear curved, which can cause parts of the displays to be cut off when defining the straight edges of the surfaces across the displays. We were able to mitigate this by having a border of surface markers around the edge of the displays, and so all of the areas we were interested in were well within the borders of the surfaces they defined. Furthermore, we used four sub-surfaces to define the displays, instead of a single surface, to prevent the surface edges being warped.

Finally, the analysis in Pupil Labs software involves a lot of manual work in setting calibration and gaze mapping time points, alterations to surface definitions, and the addition of annotations. However, the open-source software means that some of these processes can be automated. Furthermore, the analysis method is simpler compared to similar eye tracking headsets, such as Tobii Pro (n.d.) and SMI (SMI, 2017*a*), in terms of mapping gaze coordinates onto the display. Both of the alternative methods involve frame-by-frame alterations to the ROI, compared to automated mapping and tracking of the ROI with Pupil Labs that may only require one or two alterations for each subject.

From our results, we would advise that recalibrations should be done every 5 minutes. This should maintain the median accuracy below 1.25° and precision below 0.4°, while not interrupting the flow of the experiment drastically with frequent recalibrations. For future experiments with this setup, the recalibration point should be chosen for the specific needs of that experiment. It may even be suitable to wait until 10 minutes have elapsed to recalibrate, since for the first 10 minutes of our experiment the mean accuracy was not significantly worse than 1.4°. However, the longer the time between calibrations, the greater the risk of a high rate of calibration decay. Also, we observed an increase in variance after 6 minutes, with decay sharply increased for some participants. We have not attached great weight to the results beyond validation number 9, which were indeed likely to be worse. However, we cannot be sure how much this is affected by our participants tiring, rather than poor performance of the glasses. Furthermore, there were fewer than 10 participants who recorded validations beyond that point.

Overall, the results from this study show that, in its current form, this setup does not have the required spatial accuracy for mammography experiments on dual screen setups. The need to recalibrate during the experiment every 5 minutes may disrupt readers and negatively impact their performance. For experiments using mammograms, typically accuracies of around 0.5° are required, since targets can often be small. The initial significant decay of the glasses is one of the key issues. Reducing or eliminating this will vastly improve the feasibility of using them. The manual analysis is also cumbersome, where the calibration time window has to be selected and surface definitions have to be checked. Although, the use of surface tracking reduced much of the manual effort required for systems such as Tobii and SMI, and makes it easy to map coordinates onto the screens. The head pose tracking plugin with Pupil Player also makes it simple to calculate parameters that depend on distance-from-screen, such as the useful FOV. Another key benefit is the affordability compared to other commercial systems, at around £2,000 compared to £10,000+. Pupil Core is lightweight and was comfortable for participants to wear for the experiment, which could last as long as 1 hour. They also allow for free movement of the head and were therefore less restrictive than using a chin rest with a remote tracker as described in Chapters 6 and 7.

Our secondary aim was to measure the detection rate of a set of mass targets. By using various sizes and contrast levels, we had a range of detection rates that we are able to choose from for future experiments. The median detection rate was 62.5% (range 30.0%–95.0%), which is similar to the rates obtained in experiments with microcalcification clusters (Chapter 6 and 7). Smaller masses tended to have a numerically higher detection rate, which was significant between small and large high contrast masses. This appears at first to be counter-intuitive, however, larger masses tended to blend in more with the background regions, while smaller masses were more salient since they were in contrast with the random large structures in the background.

8.5 Summary

Dual-screen eye tracking poses a difficult problem and does not have a simple high-accuracy solution. Methods typically involve using multiple remote desktop trackers or a head-mounted tracker. Head-mounted trackers are light weight, making them suitable for long experiments, and do not restrict head movements, which is essential for use with large displays so that subjects can move naturally to adjust their gaze position. The key aim of this work was to test the suitability of Pupil Core for use with a clinical radiology setup.

In order to track gaze across the dual-screen setup, we used the Pupil Labs surface tracking feature, which involved displaying markers around the edge of the displays. These were automatically detected by the software and used to define surfaces over the screens, which gaze positions and fixations were mapped onto. Another software feature used was head-pose tracking, which monitored the distance of the participant from the screen and allowed for the calculation of the useful field-of-view.

To test the capability of Pupil Core, we conducted a visual search study with 26 non-expert participants. They searched a total of sixty simulated backgrounds for mass-like signals, split into three sets. The mass signals varied in size, contrast, and shape. No CAD prompts were present in the experiment. During the experiment, starting at one minute after the initial calibration, a validation sequence was completed each minute after the previous one had ended. This allowed us to measure the accuracy across the display over the course of the experiment.

The accuracy of the glasses significantly decayed after the first 1.5 minutes, but it was not significantly larger than 1.4° for the first 11 minutes after initial calibration. The variability in the accuracy of the glasses increased at around 6.5 minutes. The precision showed little change during the experiment and remained below 0.4° for the first 9 minutes. Based on the accuracy measurements, by recalibrating the glasses every 5 minutes, gaze can be tracked with less than $\approx 1.25^\circ$ error for long experimental times of up to an hour in our case. However, it is important to note the need for time-consuming manual analysis with this approach and overall lower accuracy compared to desktop eye trackers ($\approx 1^\circ$ versus $\approx 0.5^\circ$).

The set of mass targets used in the experiment gave an estimate for their detectability. These were used in Chapter 9, in addition to microcalcification clusters, to measure the interaction between prompts of different target types.

We have investigated the feasibility of using Pupil Core for eye tracking across a clinical radiology setup. Further work is needed to improve the spatial accuracy of this setup, particularly the significant initial decay. This setup was planned to be used for an experiment with expert readers in a clinical setting, discussed in Chapter 10.

Chapter 9

Interactions in CAD

9.1 Introduction

The safety-net experiment discussed in Chapter 6 and the interactive CAD study in Chapter 7 used single CAD operating points. However, commercial CAD systems typically use multiple algorithms with different operating points for the detection of masses, architectural distortions and microcalcification clusters, etc. As such, there may be multiple different prompts marking different abnormalities on the same image. It is not yet clear what the effects of these are on the reader and whether there is any interaction between these prompts.

To investigate this, we designed an online experiment for non-expert observers. This was setup so it could be run from participants own homes on their web browsers. The experiment consisted of 50 images in a single CAD condition, with two target types with their own prompts and operating points. We were primarily interested in how prompts of one target type affect how observers behave with prompts of the other target type, both when prompts mark a target and when they mark the background.

9.2 Experiment design

This was an online visual search task investigating how observers respond to single prompts compared to multiple prompts, both of the same and different target types. We used synthetic images and non-expert participants with CAD as a second reader in a single condition. The experiment was built in PsychoPy3 (Peirce et al., 2019) using the Builder interface and hosted on Pavlovia.org. Participants were sent a link to complete the experiment on their web browser.

A total of 87 participants (median age 20) took part in the study. Our participants were undergraduate psychology students and received course credit for taking part. The study was approved by the University of Manchester Research Ethics Committee (2020-10677-17195). Participants gave informed consent through the online experiment interface.

9.2.1 Stimuli

The images used in this experiment are a subset of those used in the experiments described in Chapters 6, 7, and 8. They were synthetic images created in MATLAB using open-source code (Methven and Qi, 2018). Two target types were used in this experiment. The first target type were malignant microcalcification clusters, which were extracted from magnified images of slices of mastectomies (Warren et al., 2012). The second target type were masses, which are simulated by Gaussian blobs. Both of these target types were inserted into the synthetic backgrounds by multiplying the target pixels with the background pixels. The images were initially set at 800×800 pixels, but were scaled onto the participants screen depending on their monitor size.

In this experiment, we refer to the microcalcification clusters as ‘calcs’, which was explained to the participants at the start of the training and used throughout the experiment. We will continue to use this nomenclature throughout this chapter. We required 10 mass images and 10 calc images. To choose the images for this experiment, we chose 10 images of each target type at random from our pool of target images that had been used previously in lab experiments. For each of these target images, we had an associated target sensitivity, which corresponds to their 'detectability' – the number of participants in previous experiments that had found them in the no-CAD condition. The target sensitivities of the randomly chosen targets were compared using a Kolmogorov–Smirnov test, and if the p value was >0.95 and the mean and median values were within 0.5% of each other then the condition was satisfied for these images to be used for the experiment. The distributions of target sensitivities of the chosen targets are shown in Figure 9.1. The 30 background images had also been used in previous experiments.

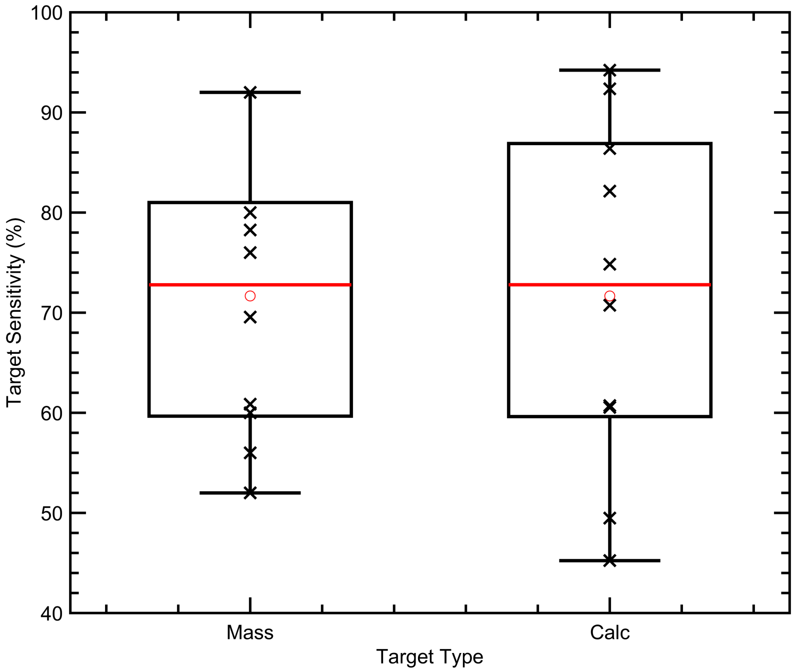
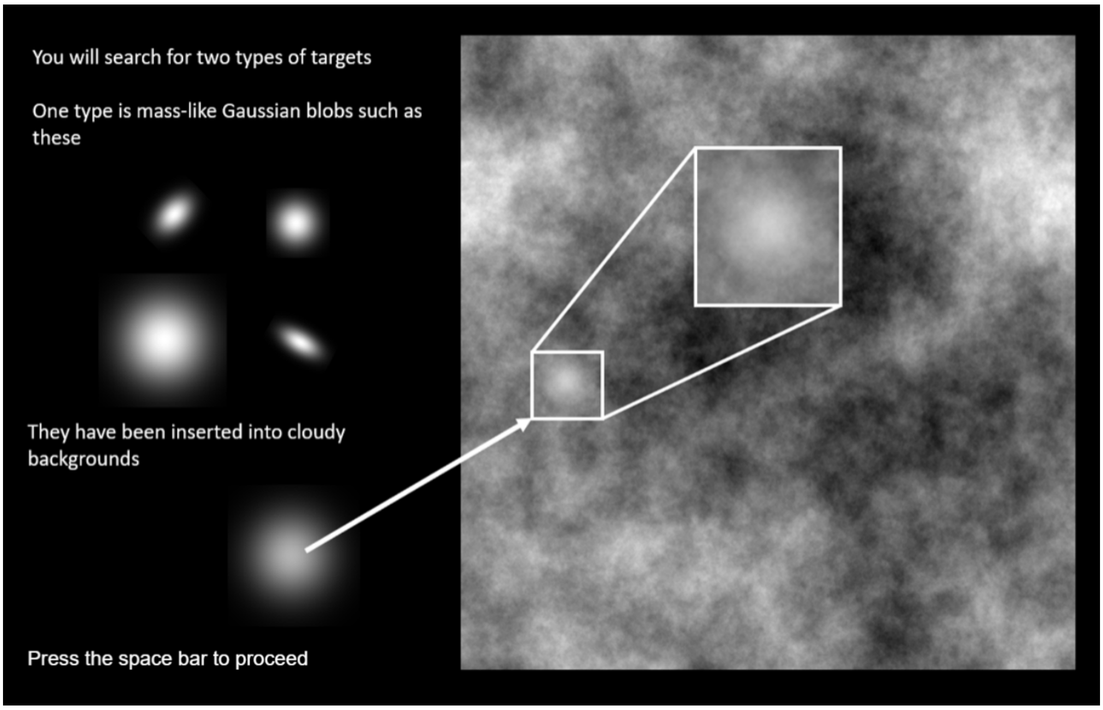


Figure 9.1: Target sensitivities from previous experiments based in the lab. The black crosses are the individual data points for each target type. The red circle is the mean and the red line is the median.

The radius of masses ranged from 82 to 393 pixels (median of 101.25 pixels), with one elliptical mass with an eccentricity of 2. The calcification clusters had an area of between 700 and 3657 pixels^2 (median 1474 pixels^2), with between 4 and 12 particles per cluster (median of 8 particles).

9.2.2 Training

Before the participants started with the experiment, they went through a training procedure. First, they were presented with each of the target types. The targets were shown as isolated examples and inserted into the synthetic backgrounds, as shown in Figure 9.2 for masses and Figure 9.3 for calcs.



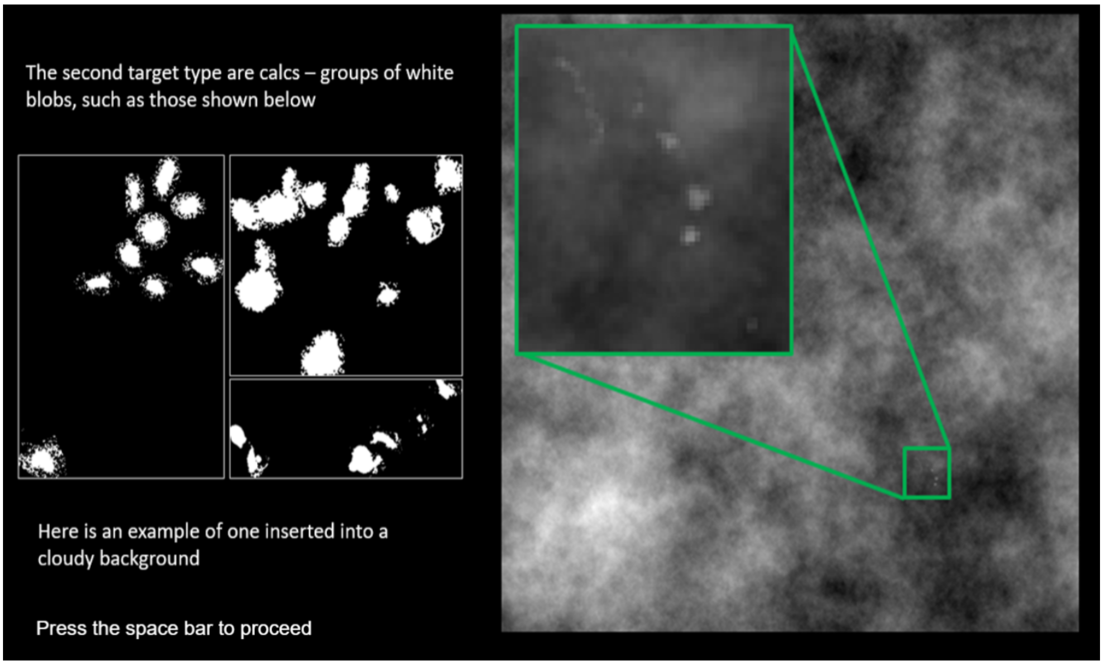
You will search for two types of targets

One type is mass-like Gaussian blobs such as these

They have been inserted into cloudy backgrounds

Press the space bar to proceed

Figure 9.2: Training screen shown to participants to introduce them to the mass targets. On the left are isolated examples of masses and on the right is an example mass inserted into a background with a magnified view of the target. The instruction to press space to move on appeared after 8 seconds, so participants were not able to continue to the next screen before this time period.



The second target type are calcs – groups of white blobs, such as those shown below

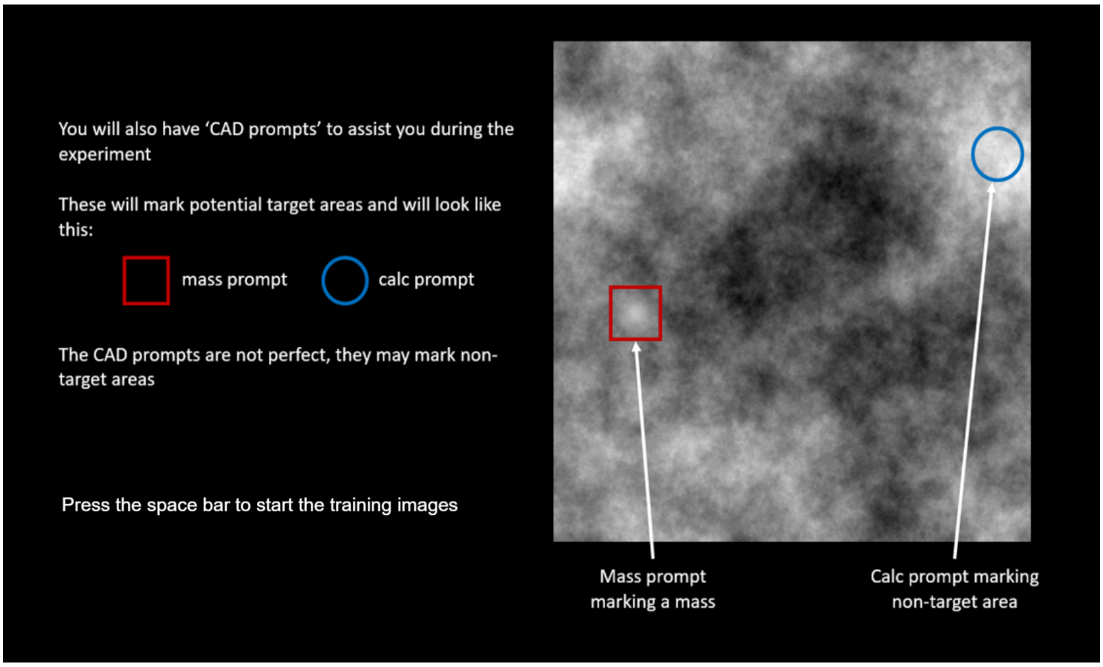
Here is an example of one inserted into a cloudy background

Press the space bar to proceed

Figure 9.3: Training screen shown to participants to introduce them to the calc targets. On the left are isolated examples of calcs and on the right is an example calc inserted into a background with a magnified view of the target. The instruction to press space to move on appeared after 8 seconds, so participant were not able to continue to the next screen before this time period.

After the participants had seen both screens introducing the target types, there was a screen that explained what CAD is and how it would work in this experiment (Figure 9.4). In this experiment, there was only a CAD condition, and it was operated in second reader mode. Thus, participants initially viewed the image without CAD and then pressed C to turn on the CAD prompts and search the image again with the aid of the prompts. This mode of operation was explained to participants throughout the interactive training section.

Next, they were shown 9 interactive training images. An example of one of these training images is shown in Figure 9.5. On the first training trial, they were told that they should first search the image initially unaided and then when they were finished,



You will also have 'CAD prompts' to assist you during the experiment

These will mark potential target areas and will look like this:

mass prompt
calc prompt

The CAD prompts are not perfect, they may mark non-target areas

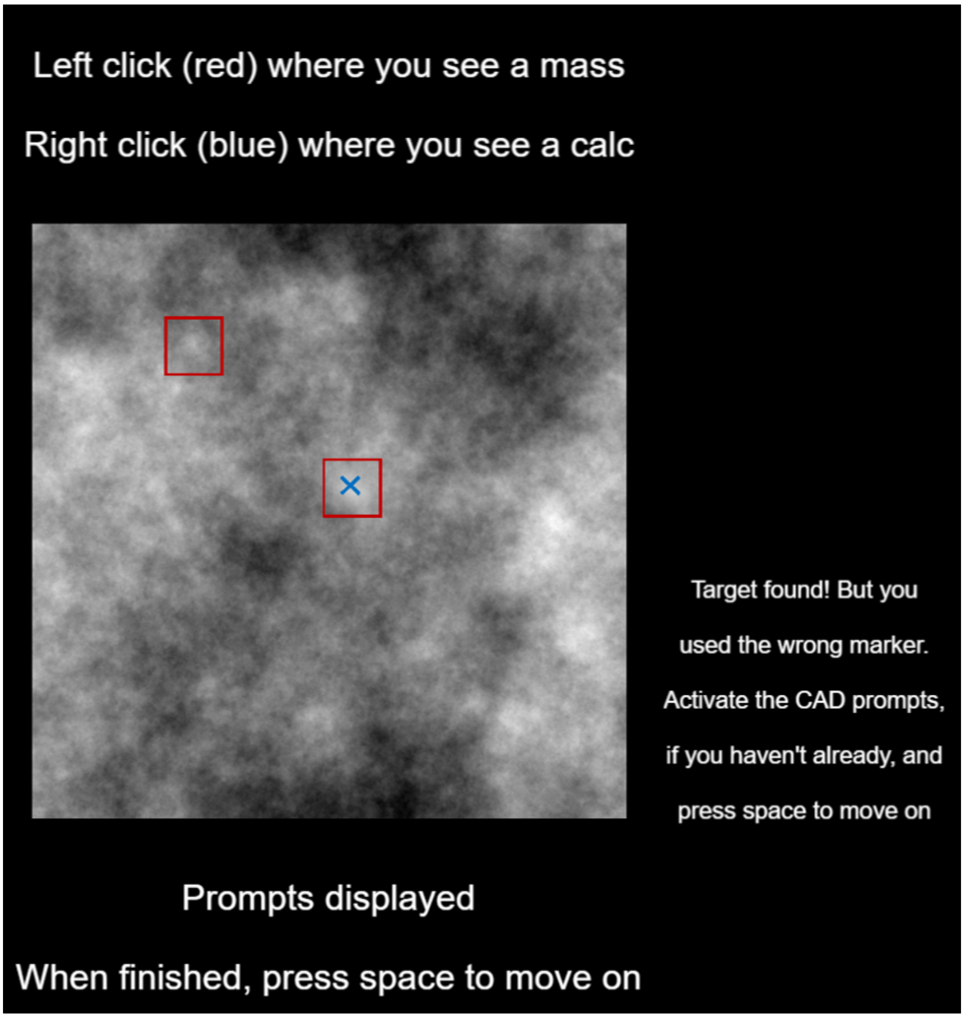
Press the space bar to start the training images

Mass prompt marking a mass

Calc prompt marking non-target area

Figure 9.4: Screen shown to participants to explain how the CAD prompts will work in the experiment. In this example there is a true mass prompt and a false calc prompt. The instruction to press space to move on appeared after 8 seconds, so participants were not able to continue to the next screen before this time period.

press C to display the prompts and search again with the aid of the prompts. When they successfully mark a target, they get a message on the right-hand side of the screen to notify them that they have correctly identified the target. After 30 seconds, a text notice appeared on the left-hand side to ask if they would like to press V to reveal the target location if they were stuck. There was a two minute time limit imposed on the training trials. Across the 9 trial images, there were a variety of prompt combinations, including one image without any prompts with a message to inform participants that there would be a number of images without prompts in the experiment trials. There was also a target absent image with an accompanying message to inform the participants that below half the experiment trials would not contain a single target.



Left click (red) where you see a mass

Right click (blue) where you see a calc

Target found! But you used the wrong marker. Activate the CAD prompts, if you haven't already, and press space to move on

Prompts displayed

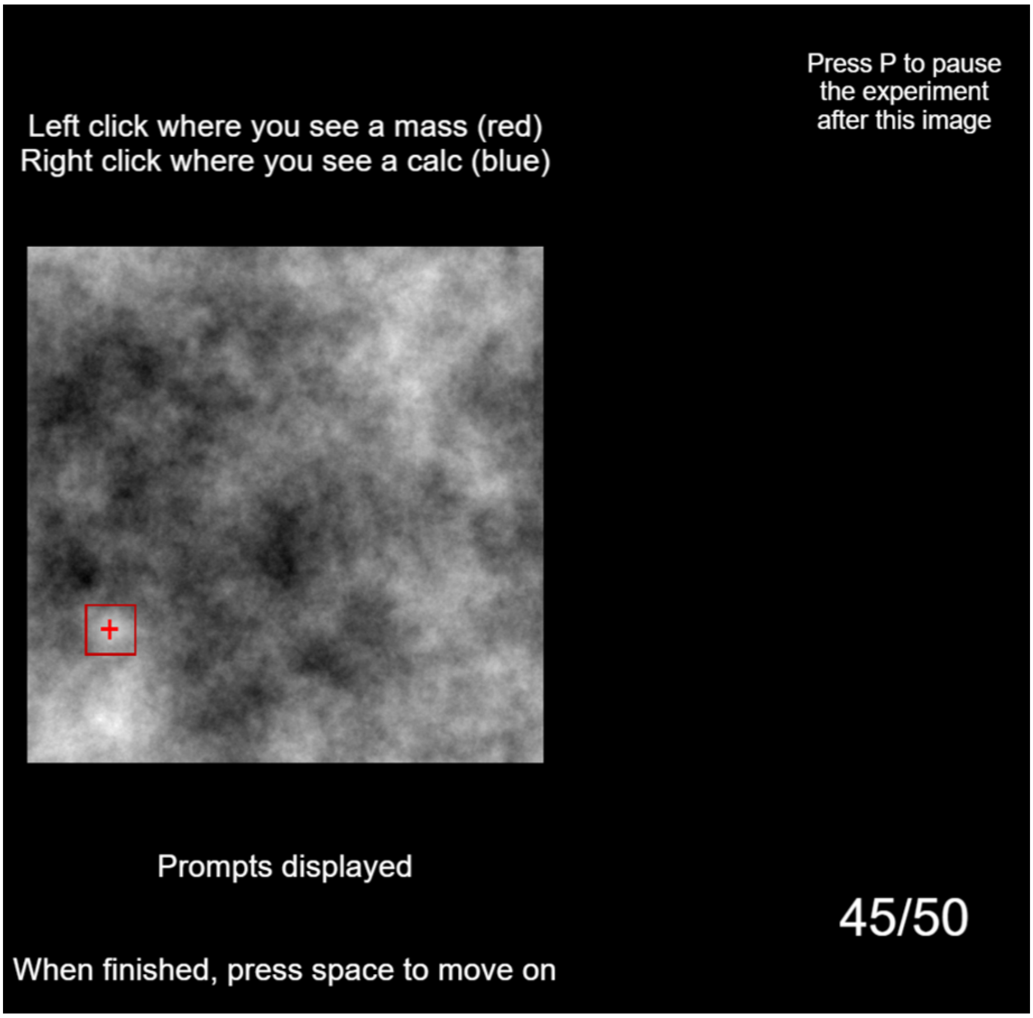
When finished, press space to move on

Figure 9.5: An example of one of the interactive training images. Participants were given instruction throughout the training. When they activated CAD by pressing C, the message below the image was displayed. On this example, there are two mass prompts, the lower one is a true prompt and the higher one is a false prompt. The participant has incorrectly marked the true prompt with the blue cross (denoting a calc) and so they received the feedback on the right-hand side.

9.2.3 Experimental procedure

Before participants went through the 50 experiment images, they were shown 3 final instruction screens. These reminded them firstly of the experimental procedure: total number of images, ability to pause the experiment, time limit, and ability to place multiple markers. Secondly, the experiment parameters: fewer than half the images contain a target, you do not have to place a marker, how to remove a marker after placing one, you will not get feedback on your clicks in the experiment. Finally, there was a screen to remind them on how to use CAD: search the images first without prompts, then press C to display prompts, then search the image again with prompts.

Figure 9.6 shows an example of an experiment trial. There were 50 images in total, 20 target present (10 mass targets and 10 calc targets) and 30 target absent. There was a 30-second time limit imposed on each image. Participants could press P to pause the experiment, which would bring up a pause screen after they had finished with the image they were currently on. There was also an opportunity for a break half way through the experiment. No feedback was given on participant clicks in the experiment trials, for example, if they were successful in finding targets. If participants changed their mind after placing a marker, they could click on that same marker to remove it. All participants saw the same 50 images, but the order of presentation was randomised between participants.



Press P to pause
the experiment
after this image

Left click where you see a mass (red)
Right click where you see a calc (blue)

Prompts displayed

When finished, press space to move on

Figure 9.6: An example of one the experiment trials. This trial contains a single true mass prompt, marked by the participant. In the bottom right there is a counter to tell them how far into the experiment they are. There is an instruction in the top right to tell them how to pause the experiment. The notice above the image remains there throughout the whole experiment. Before the CAD prompts are displayed, the text below the image reads “Press C to display CAD prompts”.

9.2.4 CAD Prompts

In this study, 80% of the masses and 90% of the calcs were marked by a prompt. The false positive prompt rates were 0.50 and 0.25 per image for masses and calcs, respectively. This corresponded to a total of 8 and 9 true positive mass and calc prompts, respectively, and a total of 25 and 13 false positive mass and calc prompts, respectively. This false positive rate is consistent with traditional CAD algorithms, which typically have around 2.4 false positive marks per case (0.6 per image) (The et al., 2009; Cole et al., 2012). The number of prompts in each image type (mass, calc or normal) is listed in Table 9.1.

Table 9.1: Distribution of prompts across images. False prompts were distributed evenly across the image types according to both the CAD operating points and the number of images available in that category. True prompts were assigned according to the CAD operating points.

Image type	**Images**	**TP calcs**	**TP mass**	**FP calc**	**FP mass**
Mass	10	0	8	3	5
Calc	10	9	0	2	5
Normal	30	0	0	8	15
Total	50	9	8	13	25

Across the images, the prompts were distributed such that they were evenly spread between the categories in Table 9.2. This shows the different image categories depending on the prompts they contain. There are 50 images, 33 of which contain prompts. They are defined by the prompts they contain: category A – a single prompt, either a mass or a calc prompt (true or false); category B – 2 prompts of the same target type, e.g. a true mass prompt and a false mass prompt; or category C – 2 prompts of different target types, e.g. a false mass prompt and a false calc prompt; finally, category D – no prompts at all.

Table 9.2: Image categories depending on what prompts they contain and the number of images in each category. There were 11 images in each of the categories that contained prompts and 17 images that did not contain prompts.

Category	**Definition**	**Total images in this category**
A	Image contains a single prompt, either true or false	11
B	Image contains multiple prompts of the same target type (mass or calc)	11
C	Image contains multiple prompts of different target types	11
D	Image contains no prompts	17

9.2.5 Analysis

Since the experiment was not based in the lab, we do not have access to eye tracking. Therefore, we were only able to collect behavioural data. The overall sensitivity of the mass and calc targets was of interest, also relative to previous lab-based experiments. The number of false positives per image was also measured. A key parameter was the number of prompts acted on, i.e., the proportion of available prompts in any given image that a participant clicks on. This was used to compare how participants change their behaviour when presented with images from categories A, B, and C. A paired t-test was used to compare within-participant differences between the sensitivity and false response rate for different image categories.

9.3 Results

In this section, results are presented for various parameters for the single CAD condition. However, in some instances it is interesting to look at how participants behave before and after they have activated the CAD prompts. This will be referred to as pre-CAD and post-CAD. There was not a no-CAD condition in this experiment.

9.3.1 Trial times

The median total trial time was 11.06s, as shown in Table 9.3. The majority of time spent viewing the images was spent before the prompts were activated, with a median pre-CAD trial time of 7.29s. The difference of 3.37s is how long participants spent on reviewing the images after they had activated the CAD prompts.

When looking at how the participants behaved post-CAD, in terms of time spent viewing the images, it is important to look at the data in the context of images where prompts were present and were absent. When no prompts were present, the participants had pressed C to activate CAD and received a message informing them there were no prompts available for that image. The bottom two rows of Table 9.3 highlight the difference in the behaviour for the 17 images where prompts were absent and the 33 images where prompts were present. When there were no prompts present, the vast majority of participants were quick to move on to the next image after activating CAD, with a median trial time of 1.96s. When prompts were present, the median trial time was 4.02s.

Table 9.3: Summary of the time spent on each image in total, pre-CAD, and post-CAD. For the post-CAD trial times, we distinguished between trials where prompts were present and absent. 25th and 75th percentiles are given with the medians.

Trial time (s)	**Number of images**	**Median (25% – 75%)**
Total	50	11.06 (8.47 – 13.33)
Pre-CAD	50	7.29 (5.43 – 9.15)
Post-CAD, Prompts Present	33	4.02 (2.89 – 5.29)
Post-CAD, Prompts Absent	17	1.96 (1.04 – 2.95)

As shown in Figure 9.7, mean trial time decreased steadily over the course of the experiment. There was a sharp initial decrease in the mean trial time; by the 13th image, just over one quarter of the way into the experiment, there was a 24% (4.19s) decrease in trial time. From the 14th image to the 50th image, there was a steady decrease of 22% (2.69s).

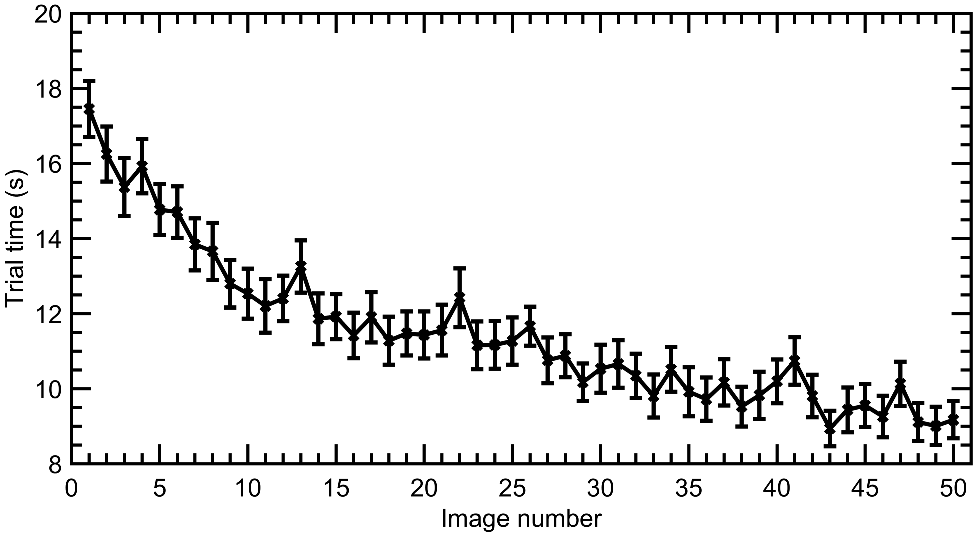


Figure 9.7: Time spent on each image over the course of the experiment, from image number 1 to 50. Error bars are the standard errors across participants.

9.3.2 Observer performance

Observer performance was measured in terms of sensitivity and the number of false positive clicks per image. Participants could either left-click to signify a mass or right-click for a calc. Individual target sensitivities were measured as the percentage of targets that a participant detected (out of a possible 10). Observer sensitivities are shown in Figure 9.8. The total target sensitivity was calculated as the mean of those two sensitivities (labelled as 'Masses & Calcs' in Figure 9.8). The mean mass sensitivity was (81.7 ± 13.1)% and was numerically higher than the mean calc sensitivity of (70.5 ± 23.6)%.

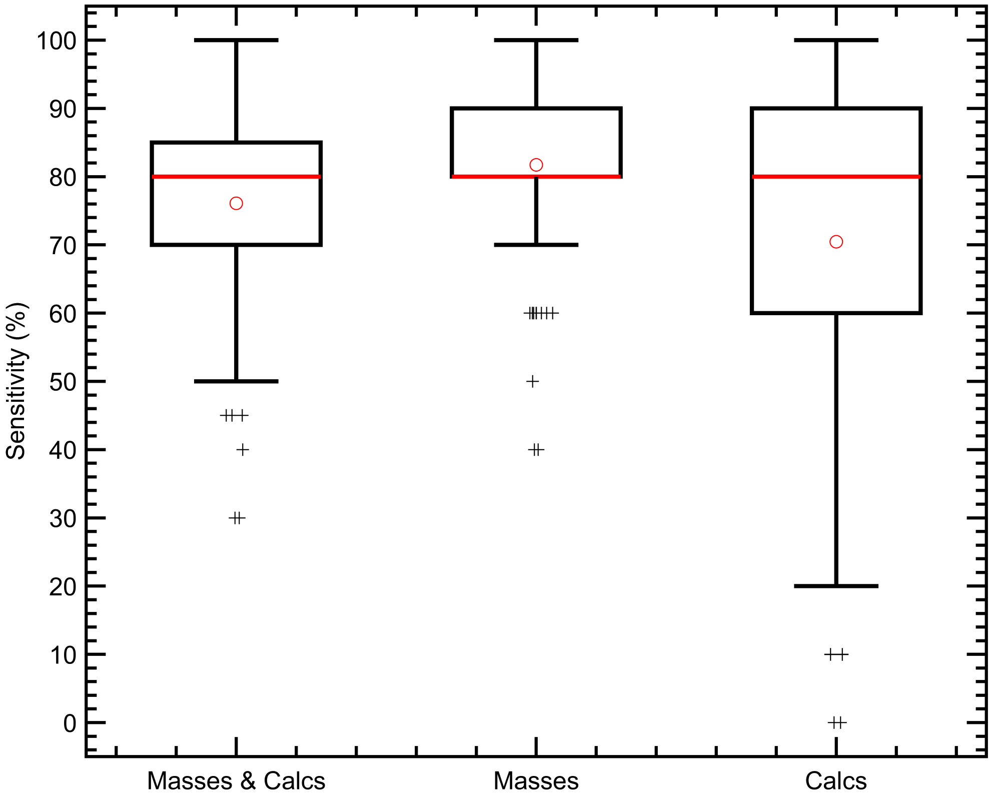


Figure 9.8: Observer sensitivities given for combined masses and calcs and individually. 'Masses & Calcs' is the mean of the Masses and Calcs data.

False positives were defined as any click not on a target region. The observer false positives per image rates are shown in Figure 9.9. The mean number of false positive mass clicks per image was 0.64 ± 0.34 versus 0.31 ± 0.26 for calcs. This difference was significant (t(86)=7.30, p<0.001). Total mean false positives per image was 0.95 ± 0.51.

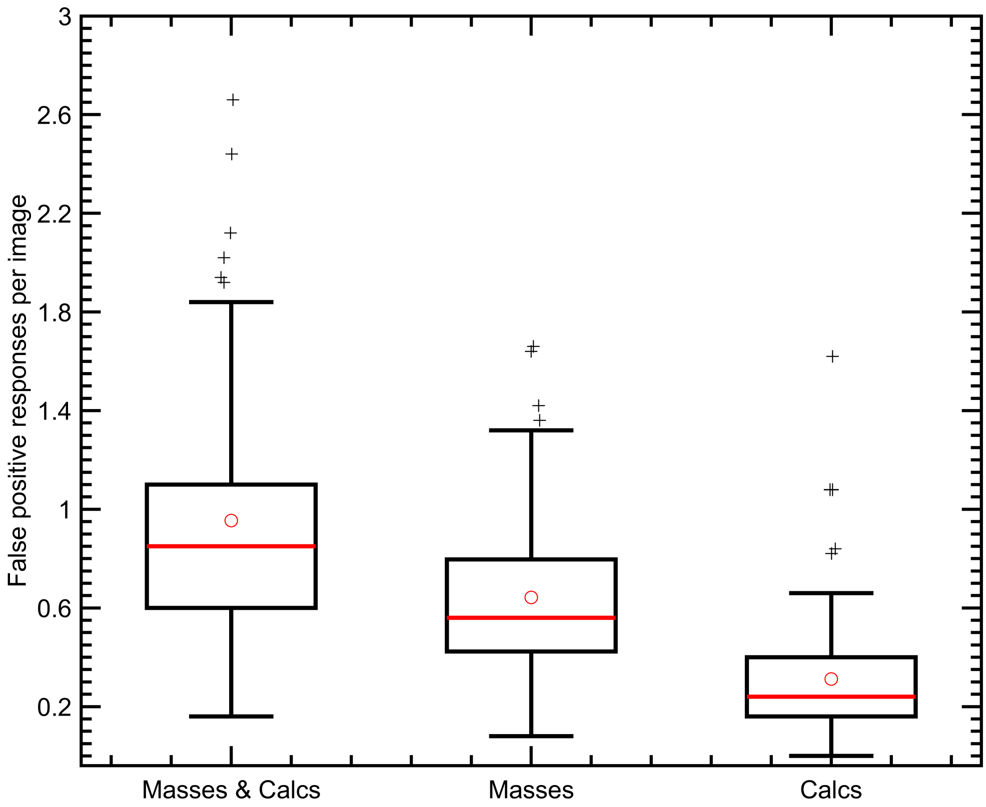


Figure 9.9: False positives per image from participant clicks for masses and calcs and individually. ‘Masses & Calcs’ is the sum of the Masses and Calcs data.

9.3.3 Effect of prompting

To investigate the effect of CAD prompts on observer behaviour, we first looked at the number of true and false positive markers that were placed and removed once CAD had been activated. This is shown in Figure 9.10. As expected, all true positive responses placed were on prompts. There was a median net sensitivity increase of 35% when CAD was activated, and net false positive responses increased by a median of 0.36 per image. The median net false positive increase of markers placed on prompts was 15. Therefore, with reference to Table 9.1, given that there was a total of 17 true prompts and 38 false prompts, participants marked, on average, 41% of true prompts and 39% of false prompts. This is just what participants marked once CAD was activated, and so we also measured the number of true prompt regions that participants marked before CAD was turned on. On average, 39% of true prompt regions were marked in the pre-CAD condition. In these cases, there was no need for the participant to act on the region post-CAD when the prompt appeared on their marker.

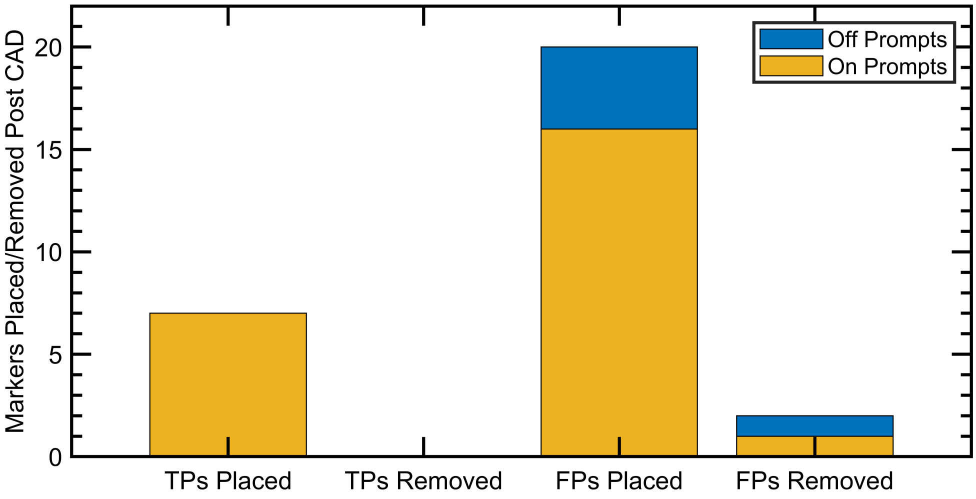


Figure 9.10: Median number of markers placed and removed by participants post-CAD both on prompts (yellow) and off prompts (blue). TP = true positive, FP = false positive.

The interaction of the mass and calc prompts is illustrated in Figure 9.11. We measured the observer sensitivity of prompted targets in the context of the three image categories containing prompts. There were 5 images containing a single true mass or calc prompt (category A), 6 images containing a true prompt and a false prompt of the same target type (category B), and 6 images containing a true prompt and a false prompt of different target types (category C). We indicate whether participants marked the targets before or after CAD had been activated and therefore had a prompt on the target (all categories), and whether there was a false prompt in the image (categories

B and C). The comparisons between categories A (single true prompt) and B/C (true prompt plus false prompt) reveal whether false prompts impact the detection rate of targets, and any difference between B and C indicates whether it matters if the false prompt and true prompt are the same target type. Sensitivity for prompted targets in image category A was 70.8%, which was not significantly different compared to images in category B at 88.1% (mean difference 17.3%, 95% CI, −7.1%−58.4%; t(86)=1.05, p=0.30) and compared to images in category C at 84.9% (mean difference 14.1%, 95% CI, −8.0%−33.8%; t(86)=0.85, p=0.40). There was also no difference in sensitivity between prompted targets in image categories B and C (mean difference 3.3%, 95% CI; −18.0%−7.6%, t(86)=0.70, p=0.48).

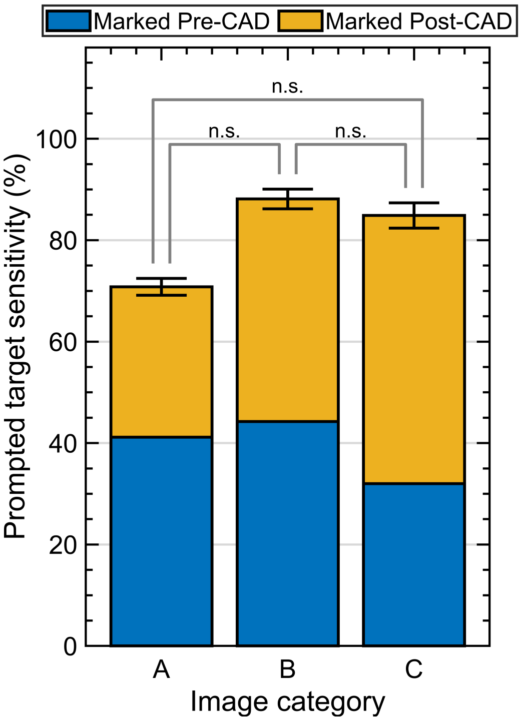
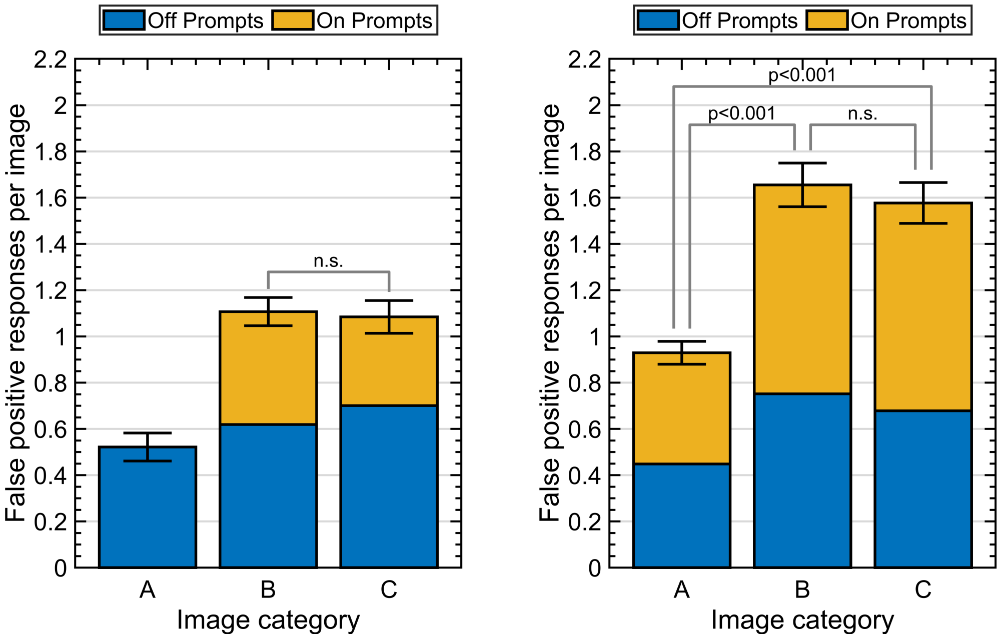


Figure 9.11: Observer sensitivity for images with prompted targets for each image category. Bars are split between the proportions of the targets that were marked before and after CAD prompts were activated (pre-CAD and post-CAD). The error bars are the standard errors on the means.

The variation in false positive response rate per image for each image category is shown in Figure 9.12, where analysis has been separated for images with and without true prompts. Figure 9.12a shows the the false positive response rate in images where true prompts were present, which are the same images as in Figure 9.11. Since there were no false prompts present in image category A, we did not make any comparisons between category A and categories B and C. The comparison between B and C investigated whether the false response rate was different in images where the true and false prompt were the same target type (category B) and different target types (category C). The total false positive response rate in category B was 1.11 (0.49 on prompts) and 1.08 (0.38 on prompts) in category C, a non-significant difference of 0.03 (95% CI; –0.26%−0.31%, t(86)=0.16, p=0.88).

The false response rates for images where there were no true prompts are shown in Figure 9.12b. There were 6 images containing a single mass or calc false prompt (category A), 5 images containing two false prompts of the same target type (category B), and 5 images containing one false prompt of one target type and one of the alternate target type (category C). This analysis was designed to measure the difference in false positive response rate for single versus multiple false prompts, and where there were multiple false prompts whether it mattered whether they were the same or a different target type. The total false positive response rate on category A was 0.93 (0.48 on prompts), significantly lower than the false positive response rate of 1.66 (0.90 on prompts) for category B (mean difference 0.73, 95% CI, 0.62−1.13; t(86)=6.58, p<0.001) and 1.58 (0.90 on prompts) for category C (mean difference 0.65, 95% CI, 0.53−1.04; t(86)=5.91, p<0.001). The mean difference of 0.08 between categories B and C was not significant (95% CI; –0.29−0.13, t(86)=0.73, p=0.47).



(a) (b)

Figure 9.12: The observer false positive response rate per image on and off prompts for each image category in (a) images with true prompts and (b) images without true prompts. The error bars are the standard errors on the means.

9.3.4 Lab versus online

To compare how participants performed in this study versus previous studies in the lab, we used participant data from the no-CAD conditions of the studies from Chapters 6 & 7 for calcs and Chapter 8 for masses as the lab target sensitivities. This was plotted against the online target sensitivities, as shown in Figure 9.13. There was no difference in the median lab and online target sensitivities for masses (72.8% versus 90.8%, z=1.85, p=0.064) or calcs (72.8% versus 78.2%, z=0.04, p=0.97), as determined by Wilcoxon rank-sum tests. Two of the unprompted targets in the online study had a greatly decreased sensitivity compared to the lab. There was also a particular target that participants in the online study only detected <20% of the time, despite participants in the lab finding it 60% of the time.

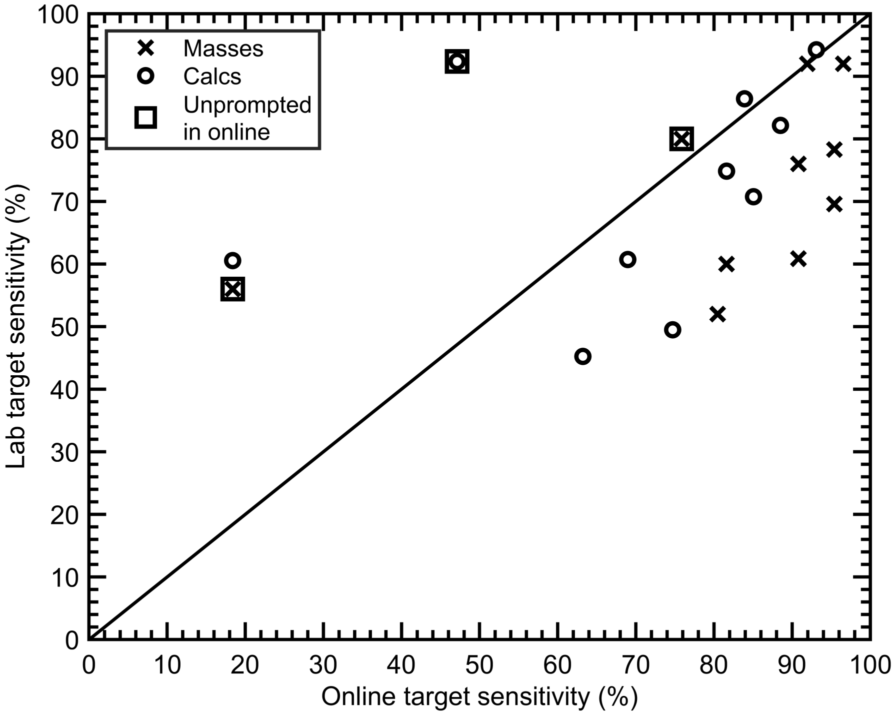


Figure 9.13: Target detectability in lab versus online experiments. The three targets that were not prompted by CAD in the online experiment are highlighted by squares. A line of equality is also plotted.

Comparing the participants in the online study to participants in the CAD conditions from the studies described in Chapters 6 and 7, the online study participants had the lowest average trial time of 11.48s vs 12.95s and 13.43s for the safety-net and interactive CAD studies, respectively. The online study had a higher false positive CAD rate of 0.75 false prompts per image (mass prompts + calc prompts) compared to the safety-net study (0.5 false prompts per image). The false positive response rate of participants in the online study was more than double that of the safety-net experiment (0.95 vs 0.35). These were both higher than the interactive CAD experiment of 0.22 false clicks per image.

9.4 Discussion

This study investigated the interaction of multiple CAD prompts on single images, and how participant performance was affected by their presence. Prompted target sensitivity was unaffected by the presence of a false prompt. There was no difference in how participants acted on prompted targets in images with a false prompt of the same or different target type present. Participants in the online study performed better for masses in terms of their sensitivity and the same for calcs compared to the participants evaluating the same target images in lab experiments, but the majority of the targets in the online study were prompted.

False prompts did not distract participants when targets were prompted by CAD. This was the case when the false prompt was of the same target type or of the alternate target type. These results are consistent with a similar study that looked at distraction caused by false positive prompts (Ionescu et al., 2018). They found that for prompted targets, the presence of false prompts did not affect observer sensitivity. Although, they only reported on one target and prompt type, so it is a new result that sensitivity remains unchanged when an additional target and prompt type is introduced in our study.

In the study by Ionescu et al. (2018), it was only when CAD failed to mark a target that sensitivity was significantly reduced, which has also been reported in other studies, for non-experts (Drew et al., 2012; Kunar et al., 2017) and experts (Alberdi et al., 2004; Zheng et al., 2004). Since in our study only 3 of our 20 targets were unprompted, we cannot comment on the sensitivity of those cases.

The CAD signal in this study, the ratio of true to false prompts presented to the participants, was 1:3.13 and 1:1.44 for mass and calc prompts, respectively. The mean ratio of 1:2.24 was higher than the ratio of 1:1.56 in the previous study with traditional CAD in the lab (Chapter 6). Participants saw more false positive prompts for every true prompt in the online study compared with the lab study, but they were more likely to act on true prompts when they did appear, at 43% versus 34%. In the lab study, participants marked 55% of true prompt regions in the pre-CAD condition compared to 39% in the online study. Despite the CAD signal in the online study being weaker than the lab study due to the higher ratio of false to true positive prompts, participants were still marking more true prompts post-CAD than in the lab study. Online participants marked true and false prompts at a similar rate, suggesting they had become overreliant on CAD to make decisions, and were not capable of distinguishing between true and false prompts.

For trials where true prompts were absent, false positive responses were significantly higher by a factor of two in images with twice the number of false prompts. However, in the study by Ionescu et al. (2018), false positive responses were not correlated with the number of false prompts, suggesting participants in our study were much more reliant on CAD. Kunar et al. (2017) argue that search tasks with multiple target types can cause an overreliance on CAD due to target representations being weakened in participants' visual working memory (see Section 4.3). Our results were consistent with Kunar et al. (2017), with correct responses when CAD marked targets and false positive errors when CAD marked non-target regions. The false prompt rate for masses was significantly higher than for calcs. This can be explained by the fact that in Chapter 8, the FP/image rate for masses (no CAD prompts were used) was 0.70, compared to the rate of 0.28 per image for calcs in Chapter 6. This suggests that there were more mass-like features in the image than calc-like and therefore were more likely to be acted upon, especially if they were prompted.

The performance of the participants in terms of their sensitivity demonstrated the viability of running this sort of experiment online. Participants achieved similar or higher sensitivity compared to participants in the lab. The observer sensitivity in the online experiment was boosted by the aid of CAD prompts. Since the task was unsupervised and with only on-screen instructions, we ensured the task was short enough to maintain their concentration levels, and as simple as possible since they were not able to ask a researcher questions during the experiment like they usually would. However, without a no-CAD condition, we do not have a true measure of their performance level to compare to the lab study.

In addition to these experiment design considerations, there are further limitations with running an online study. Firstly, we were limited by how long the experiment could last so that participants were more likely to engage with the task for the whole duration. The short experiment time meant that for each participant we did not collect a lot of data points per condition and limited what we were able to conclude from the study. We were also restricted to a single CAD condition, and while this did not affect the analysis of behaviour with different combinations of prompts, it would have allowed us to benchmark reader performance against the CAD condition.

Furthermore, we used height units for the images, which meant that they were scaled to participant displays, and therefore were not necessarily the same size for each participant. Participants may also be interrupted, or may themselves choose to stop the experiment at any point. We added a pause feature to the study to control for this, allowing participants to stop and resume the study at any time. We were also limited to behavioural data only and unable to track eye movements. Parameters such as proportions of targets and prompts fixated and their associated dwell time, total image coverage, and time taken to fixate the target can add crucial insight into participant behaviour. We were also not in control of the light conditions, where in all previous experiments we have operated them in with the lights off for the benefit of the eye tracking, it may also enhance the clarity of the displays. There will have been a large variation in the setting that each participant completed the experiment in, including the size of their computer monitor, which we were not able to account for.

9.5 Summary

We have described an online experiment for investigating the effect of multiple CAD prompts on observer behaviour in terms of their sensitivity and specificity. Participants completed 50 images with the aid of CAD, and were tasked with finding both masses and calcs. CAD marked 80% of the masses with 0.5 false prompts per image and 90% of calcs with 0.25 false prompts per image.

Analyses of the images containing prompts were categorised by the prompts they contained. The three categories were: a single prompt, two prompts of the same target type, and two prompts of different target types. The prompted target sensitivity did not vary across image categories, suggesting that participants were not distracted by false positive prompts, regardless of the target type of the false prompt.

Participants became overreliant on CAD, which in images without true prompts present was detrimental; false positive responses were proportional to the number of false prompts present on the image. Again, this was unaffected by whether there were two false prompts of the same or alternate target types.

The performance of participants was encouraging for running observer studies online to test the effect of CAD with non-experts, achieving similar sensitivities as previous lab studies. However, methodologies must be adapted to suit the nature of online tasks. We shortened the length of the experiment compared to usual lab tasks and removed the no-CAD condition to focus entirely on the effect of prompts, and were specific in our instructions to make sure participants understood the task at hand without the need to ask a researcher any questions before participating.

Chapter 10

From lab to clinic

10.1 Introduction

The experiments outlined in Chapters 6, 7 & 9 have highlighted various advantages and disadvantages of using CAD. However, they were all performed with non-expert readers. Therefore, it was planned that an eye tracking experiment would be performed with expert readers, using the setup outlined in Chapter 8, to investigate how prompts affect their visual search behaviour and performance. Unfortunately, data collection could not be completed due to government restrictions in the COVID-19 pandemic. However, we include in this chapter a detailed description of the planned experiment and the work that went into its setup.

The effect of false prompts on observer sensitivity was investigated by Ionescu et al. (2018) with non-expert readers. A single prompt type was used and sensitivity of prompted targets was unaffected in the presence of a false prompt. For unprompted targets, sensitivity was significantly reduced, similar to other studies with non-experts (Drew et al., 2012; Kunar et al., 2017) and experts (Alberdi et al., 2004; Zheng et al., 2004). In an eye tracking study with non-expert participants, CAD prompts led to a significant reduction in image coverage, with attention focused on the prompted regions (Drew et al., 2012). The operating point is an important factor and a higher rate of false prompts has been shown to reduce performance (Zheng et al., 2004). Without eye tracking, it is difficult to conclude whether this reduction in performance is a result of prompts distracting readers attention away from abnormalities, or whether trust in the prompts was reduced with a higher false prompt rate resulting in more true prompts being ignored. Most likely it is a combination of the two. A study tracking eye movements of expert readers with a clinical setup has yet to be published to investigate the effect of CAD prompts. It is also important to operate CAD as a second reader since this is how it was designed to be used, but in many observer studies, such as those mentioned above, prompts were displayed from image onset.

In Chapter 6, CAD improved observer sensitivity at the cost of an increase in the false positive response rate, a finding that is common in literature (see Section 3.5). This was using a single target and prompt type. The use of multiple targets, each with a specific prompt type, was investigated in Chapter 9. Observer sensitivity for targets that were prompted did not change in the presence of a false prompt. Participants demonstrated an overreliance on CAD, resulting in a high number of false prompts being accepted as true. Participants across all previous experiments showed a high level of trust in prompts, and in Chapter 7 this trust scaled with prompt confidence values. Expert readers are not likely to be as trusting. In fact, previous studies have shown that they are often dismissive of prompts even when they are correct (Nishikawa et al., 2012). Therefore, it would be interesting to observe how these results translate to a study with expert readers.

This experiment was designed to test how CAD prompts affects reader behaviour with a realistic clinical setup. In particular, whether false prompts distract readers from targets and whether there is an interaction between multiple prompts on the same case. We are interested in how false prompts impact the way readers react to prompted abnormalities. Since prompts are displayed across different image views and breasts in mammography screening, it would be interesting to see how this results in differences from the previous experiments. Readers are also more experienced with the images than non-experts, although mammograms are also more complex than the images used with non-experts. Since the safety-net effect was previously observed (Chapter 6), this study will continue to use a no-CAD and a CAD condition, rather than using the pre-CAD search as the no-CAD condition as in many cross-sectional CAD studies. This will create a more valid comparison between no-CAD and CAD observer behaviour.

10.2 Methods

10.2.1 Stimuli

The images in this study are digital mammograms from the OPTIMAM Mammography Image Database (OMI-DB). The database is sourced from multiple screening centres across the UK and was designed with the purpose of sharing with external research groups (Halling-Brown et al., 2020). OMI-DB contains over 2.5 million images, collected from 173,319 women attending the NHSBSP. The database comprises of unprocessed and processed images, and unannotated and annotated images, with medical expert readers providing annotations and ground truth labels.

We were provided with a subset of 6500 images, the details of which are outlined in Table 10.1. Malignant and benign images were classified by screening, surgery, biopsy, or have been previously assessed as malignant or benign. Normal images are those that were not classified as either malignant or benign with no history of malignancy. From the available cases, we aimed to select 92 cases, with 56 normal, 18 with a malignant mass, and 18 with a malignant microcalcification cluster. This gave a target prevalence of 20% for each abnormality type.

Table 10.1: Characteristics of OPTIMAM images we had access to.

Image type	**Number of images**
Annotated malignant	2594
Annotated benign	1387
Unannotated malignant	719
Unannotated benign	800
Normal	1000
Total	**6500**

To select the abnormal cases for the experiment, we first selected only those cases which had a lesion conspicuity rated as ‘very subtle’ to provide a difficult case set. Abnormal cases were further stratified by those that did not have an associated age, since this was to be used for density matching between normal and abnormal cases (discussed below). We excluded any case that contained additional abnormalities to the targeted one. We ensured that all cases were obtained from systems of the same manufacturer, this was chosen as the manufacturer with the most potential cases (Hologic, Inc.). Once a final subselection of potential images was obtained, these were visually inspected to check for any unwanted features (such as surgical staples) and removed if necessary. From the remaining selection of cases, 18 images were randomly chosen for each abnormality type.

Once the 36 abnormal images were chosen, we started to select potential normal cases. To keep the overall image features as similar as possible between normal and abnormal cases, we matched the distribution of breast densities between the image sets. Density was not provided in the associated information for each case, and therefore we used an algorithm that predicts percentage density (Ionescu et al., 2019), trained on the predictions of expert readers using a visual analogue scale (VAS).

For each abnormal case, five unique normal cases with an age within five years of the abnormal cases were chosen as a potential case to use (total of 180 potential normal cases). Matching within an age range would be more likely to be closer in density, giving a better chance that similar cases could then be selected to match the abnormal densities. The breast density of all abnormal cases and potential normal cases were predicted with the algorithm described in Ionescu et al. (2019). The mean value was taken across the four image views to give a single average density for each case. To match densities between density distributions, the 36 abnormal cases were compared against 80 randomly selected normal cases from the potential normal cases. Distributions were compared using a Kolmogorov-Smirnov test, where the normal cases were selected for use if $p>0.99$. We selected 80 cases, more than the required 56, so that cases could be removed if they contained unwanted features on any image view. Once images were removed, the final 56 images were randomly chosen. Comparing the distributions of the 56 normal images with the 36 abnormal cases gave a p-value of 0.96 using a Kolmogorov-Smirnov test. As in Chapter 9, we will refer to malignant microcalcification clusters as calcs throughout this chapter.

10.2.2 CAD prompts

The chosen CAD operating points for each prompt type were selected to reflect commercial CAD algorithms. CAD prompts marked 80% of masses and 90% of calcs. The false positive prompt rates were 1.6 FP/case (0.4 FP/image) for masses and 0.8 FP/case (0.2 FP/image) for calcs. The mean number of false prompts per case was therefore

2.4, in line with values reported in literature of 2.3 per case (The et al., 2009) and 2.56 per case (Cole et al., 2012).

The abnormal cases that will not be marked by a true prompt were selected randomly, with the true prompts placed on the remaining cases the abnormal regions outlined by OMI-DB. For false prompts, potentially suspicious regions had to be selected across normal and abnormal cases. To do so, we presented the 92 cases (plus 4 training cases) to two consultant breast radiographers. The radiographers were instructed to place prompts on any potentially abnormal or suspicious regions and those which could be recalled. They were able to drag and drop prompts onto the selected regions, with a different prompt shape denoting the abnormality type (mass or calc). They were not told the desired number of prompts and were free to not place any on cases where they did not see any potential regions. They reviewed the cases independently of each other.

In total, 222 unique mass prompts and 17 unique calc prompts were placed by the radiographers. For mass prompts, 148 had to be selected out of the 222 to meet the target operating points. Any prompt that the radiographers agreed upon was accepted, and the remaining prompts were randomly chosen (see Table 10.2). The 17 calc prompts selected by the radiographers was much lower than the required 74 to meet the target operating point.

Table 10.2: False prompt distribution across different images types.

Image type	**Mass prompts**	**Calc prompts**
Mass	24	24
Calc	19	18
Normal	105	32
Total	148	74

To make up for the deficit in calc prompts, we added further prompts by identifying potential image regions based on the areas that prompts were placed by the radiographers. First, filtering was repeatedly applied to the images that contained the 17 radiographer-placed prompts, with structuring elements of the filter process shown in Figure 10.1 and outlined in Zhang et al. (2013). The filters were applied serially to these images to validate whether this method was capable of highlighting potential calc regions. This top-hat filtering process was completed in Matlab using the *imtophat* function, repeated for each structuring element shown in Figure 10.1.

Once the images were filtered, we highlighted the potential calc regions by thresholding and converting the image from greyscale to binary by setting all values below the threshold to 0 and all above the threshold to 1. It was possible to highlight 14 out of 17 of the regions where the radiographers placed prompts, such as in Figure 10.2. Since this method produced acceptable results, it was applied to all images to highlight potential regions and calc prompts were added to a portion of those regions such that prompts were distributed evenly across image types (see Table 10.2).

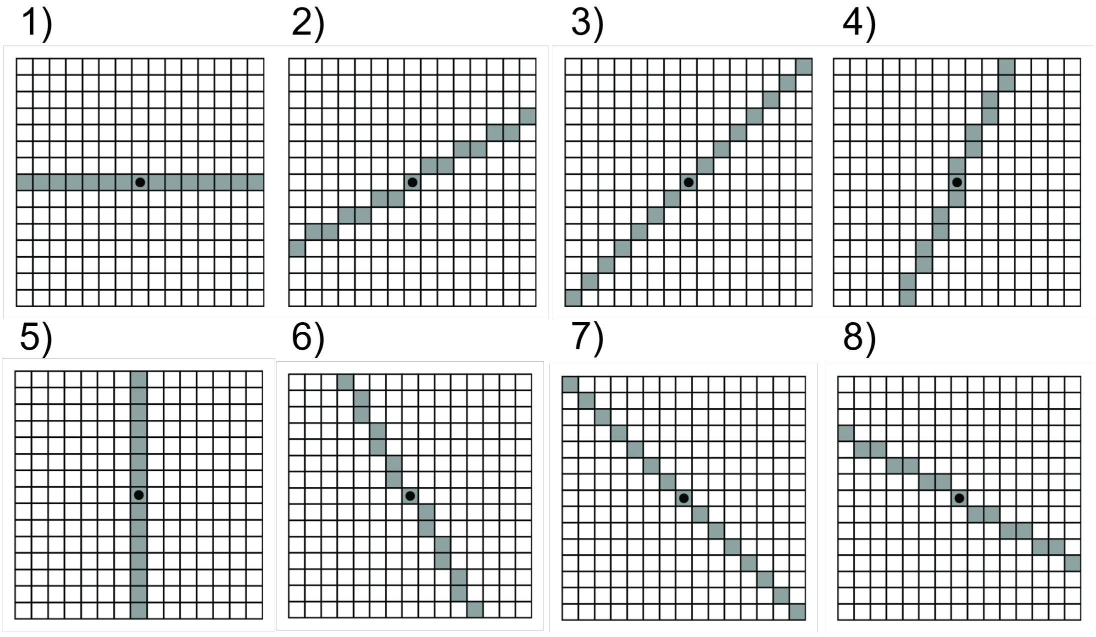


Figure 10.1: Visualisation of the structuring elements used with top-hat filtering applied to images for highlighting calc regions. Filled squares represent 1s and unfilled squares represent 0s. Elements were applied serially from 1 to 8. Images from Zhang et al. (2013).

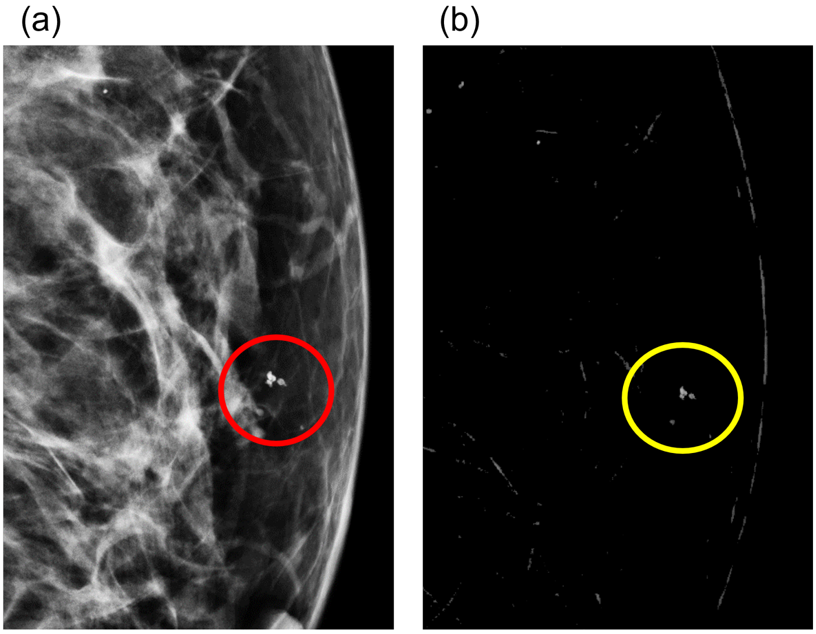


Figure 10.2: (a) Unfiltered mammogram with prompt placed by radiographer (red circle) and (b) Mammogram with filters applied to highlight calc regions (yellow circle).

10.2.3 Experimental setup

The experimental setup is the same as that used in Chapter 8, except that it was to be set in a reporting room at a hospital to better replicate clinical conditions and convenience for readers. The experiment was operated with a mouse and keyboard and did not make use of a side monitor as cases were automatically loaded one after another.

The experiment framework, created in PyGaze v0.6.0 (Dalmaijer et al., 2014), was designed to be simple for the reader to operate while allowing for features they are used to. As shown in Figure 10.3, readers were able to change the image view using the option box in the top left of the screen, between MLO (default view), CC, and both CC&MLO at the same time.

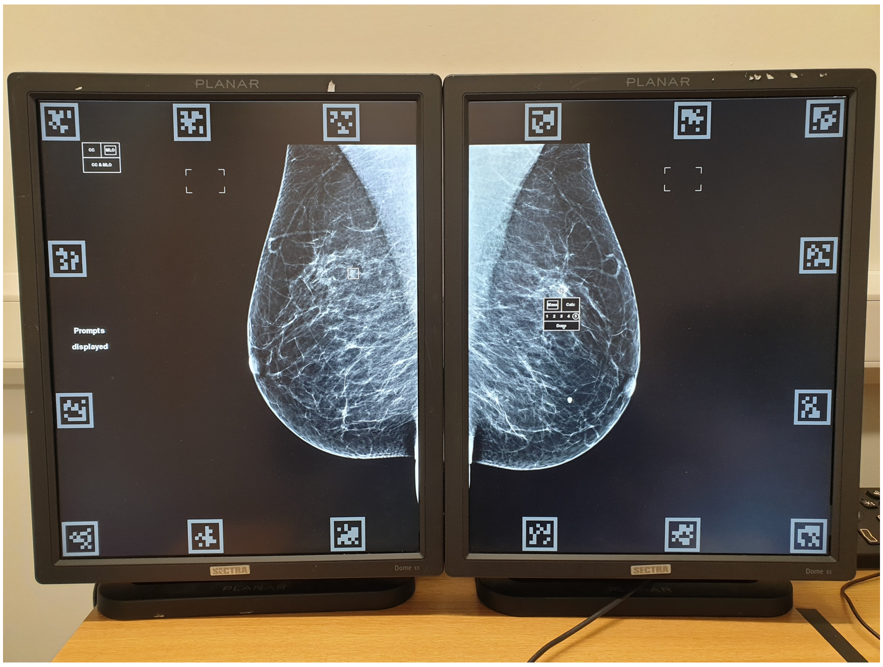
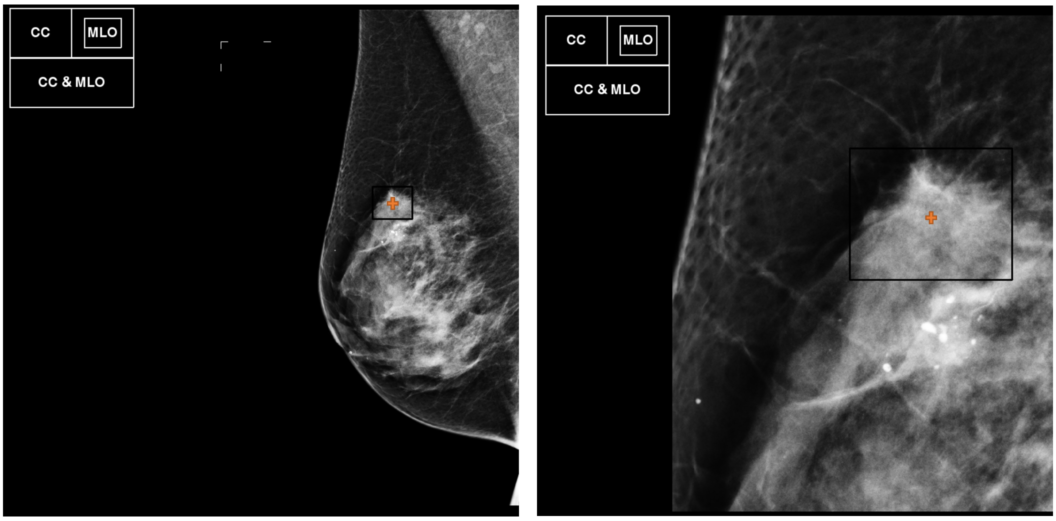


Figure 10.3: Photograph of experiment running on dual screen display. The option box for selecting the image view is in the top left of the displays. On the right screen is the pop-up option box for choosing the abnormality type and confidence value following a left-click to place a marker.

Furthermore, readers could zoom the image by a fixed factor of 4 by pressing the Z key. The image zooms around the cursor position, shown in Figure 10.4. Readers are not be able to pan across images. Readers left click to place a marker where they believed an abnormality to be. This brings up an option box where the reader first chooses the abnormality type (mass or calc) and then their confidence in that decision, rated on a scale from 1 (least) to 5 (most) confident that the marked region is abnormal.



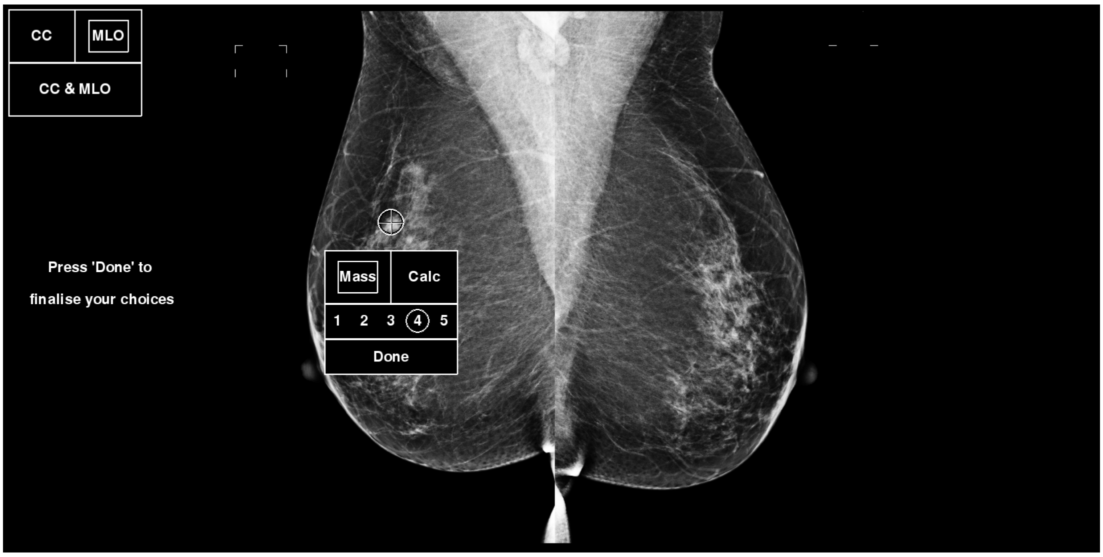
(a) Unzoomed view (b) Zoomed view

Figure 10.4: Screenshots of experiment screen demonstrating the zooming feature. (a) Unzoomed image view and (b) zoomed image view with a magnification factor of 4. The image zooms around the position of the cursor, which is denoted by the orange plus here. The black box outlines the position of a malignant mass, for illustrative purposes only.

10.2.4 Training

At the start of the experiment, readers will have a training set of cases to become familiar with the controls and experimental setup. They will receive on-screen instructions and feedback on their clicks throughout the training. The training set was designed to wait for the specific instructions to be completed before the participant is able to move onto the next case, to make sure they have correctly interacted with it. A total of 11 cases were used for training, with 7 used to explain how to operate the experiment software and the other 4 to practice with (2 CAD and 2 no-CAD).

The explanation of the software first involved how to change views and how to zoom in and out of images. A single view could be zoomed or multiple zoomed views could be displayed at once, and images were reset when the view was changed. Next there was a walkthrough on placing markers on regions where readers were suspicious of, shown in Figure 10.5, where the scale for rating confidence was explained. Choices made for a marker were displayed next to a marker once it has been placed. It was also explained how to remove a marker should a reader change their mind. Markers were mapped to the correct positions when image views were changed between single and multiple view states and when zooming, which was demonstrated in the training.



Press 'Done' to finalise your choices

Figure 10.5: Screenshot of experiment training. On this case, the participant was instructed to place a marker by left clicking, with the text on the left hand side updating as they made their choices. Once Done is pressed, the choices appear next to the marker, in this case ‘M4’ for a mass with confidence 4.

It was explained how to operate the CAD prompts - pressing C activated the prompts and C again to toggle them off. Square prompts represented a potential mass and circular prompts represented a potential calc. Readers will be told that they should review the image first without prompts and when they have finished the initial view, review the image again with the aid of prompts. Finally, readers will be provided with four practice cases, with and without CAD, without any feedback to practice before the experiment. Readers will be free to ask any questions during the training.

10.2.5 Experimental procedure

The experiment was planned to occur over 2 sessions for each reader, each 1 hour in length. The sessions were due to take place at the Nightingale Centre at Wythenshawe Hospital. At the start of the first session, the readers will be briefly interviewed about their experience with breast screening and CAD. This would allow us to take any effect of experience with mammography reading or with CAD into consideration in the discussion. A total of 10 readers had agreed to participate in the experiment. These readers were either breast radiologists or radiographers, or advanced imaging practitioners. This sample size was determined by convenience sampling, based on the number of available expert readers at the Nightingale Centre.

A total of 92 images will be used in this experiment. All cases are interpreted with and without CAD over the two sessions. In each session, 46 images will be completed with CAD and 46 without, with the other half of the images for each condition read in the next session. Of the 92 mammograms to be used in the experiment, 20% (18 images) of the images contain a malignant mass and 20% (18 images) contain a malignant microcalcification cluster. Readers will not be told the exact number of abnormal cases but instead that it is an enriched dataset.

Readers complete the training set at the start of the first session. They will then be given the opportunity to ask any questions they may have about the experiment. When the participant is satisfied that they fully understand the task, the eye tracking glasses will be setup and calibrated. This process is the same as that explained in Chapter 8, with 15 calibration targets displayed for 3 seconds each. A practice calibration will be completed before the real one to familiarise readers with the process in case they have no prior eye tracking experience. Then we will validate this process to check the calibration accuracy, a similar process, but the targets appear for a shorter amount of time, lasting only 30 seconds in total.

Each session consists of 2 conditions: no-CAD and CAD. The order of the conditions will be evenly varied across readers, and counterbalanced for the same participant between sessions. During the experiment, there will be a calibration sequence every 5 minutes to maintain the accuracy of the eye tracking glasses, based on a previous experiment with this setup that determined this to be the optimum recalibration time (see Section 8.3.2). Readers click where they believe a target to be, specifying what target type they think it is and assign a confidence rating to that decision on a scale from 1 to 5.

In the CAD condition, the readers first search the image unaided and then press a button to review the image again with the aid of the CAD prompts. In the CAD condition, 80% of the masses and 90% of the microcalcification clusters are marked by a prompt. The false positive prompt rates are 0.4 and 0.2 per image for masses and microcalcification clusters, respectively. This produced a mean prompting sensitivity of 85% with 2.41 false prompts per case. Readers will be informed the approximate operating point of the CAD system of between 80% to 90% sensitivity and to expect 2 to 3 false positives per case. This allows readers to set suitable expectations of the prompting system, rather than making their own assumptions on the prompt accuracy from the initial cases and changing their behaviour according to that.

10.2.6 Analysis

The sensitivity, defined as the proportion of abnormalities correctly located, in addition to the number of false positives per image, will be reported. To compare the overall performance of readers between conditions, we will use JAFROC analysis. To differentiate lesion localisations from non-lesion localisations on abnormal images, we will compare the locations of responses to the image annotations provided by expert readers as part of the OMI-DB, with a tolerance of 50 pixels around the annotation boundary. We will implement the variation of JAFROC where false positive responses on abnormal cases are taken into account (Chakraborty and Yoon, 2009). The difference in performance between the no-CAD and CAD conditions will be compared using the calculated JAFROC scores. To determine whether statistical differences were observed between conditions, a Mann–Whitney U-test will be used.

We will measure the percentage of prompts correctly acted on (true prompts marked and false prompts ignored), and compare between cases where there is a single prompt with those where there are multiple prompts present. The comparison will be bootstrapped following the approach outlined in Section 6.2. This gives an insight into how multiple prompts affect the behaviour of readers compared to a single prompt. In addition, we will measure the percentage of true prompts fixated and the associated dwelltimes with and without false prompts present. The overall trial time and image coverage will be compared between the CAD and no-CAD conditions using a bootstrap approach to quantify how CAD impacts the thoroughness of reader search.

The confidence ratings of responses will be compared between the CAD and no-CAD conditions for normal cases, and between all cases with and without false prompts present. The significance of these differences will be measured with a Wilcoxon signed-rank test. The purpose of these comparisons is to test whether the presence of true prompts improves reader confidence in their responses and if false prompts on cases reduces confidence.

10.3 Discussion

Overall, we would expect that CAD will improve sensitivity at the cost of an increase in the false alarm rate. However, it is expected that expert readers will be less trusting of CAD than non-experts. Even across readers trained to interpret mammograms, less experienced readers tend to rely more on CAD (Nishikawa et al., 2012; Hupse et al., 2013). The CAD operating point is important to gaining a benefit, as it has been shown that a CAD sensitivity of 80% with 0.5 false prompts per image can improve reader performance, but 1.2 false prompts per image have no effect or may even be detrimental (Zheng et al., 2004).

Across cross-sectional studies that examined the effectiveness of CAD, for those that reported CAD sensitivity and false prompt rate, on average CAD marked 70% abnormalities with 2.9 false prompts per case (Freer and Ulissey, 2001; Birdwell et al., 2005; Dean and Ilvento, 2006; Ko et al., 2006; Morton et al., 2006). In these studies, CAD improved reader sensitivity by 10.4% with a 16.3% increase in recalls. This can be contrasted with the results from Chapter 6, where operating CAD at 80% sensitivity and 0.5 false prompts per image resulted in a 15.8% increase in observer sensitivity with the number of false alarms per image increasing by 37.4% for target absent images (most comparable to recall rate in this case). Non-expert readers in our study had over double the false response rate compared with experts, so while an abundance of false CAD prompts are expected to remain an issue for expert readers, that impact is not likely to be as extreme.

A difference between experts and non-experts would be in part due to expert readers having a much deeper understanding of the search task and of CAD. Generally, familiarity with the images will likely make experts better at identifying true prompts and dismissing false prompts. However, prior knowledge of the high false positive rate of CAD could negatively prejudice readers to be inherently distrusting of prompts, compared to non-experts who tend to rely on CAD to make decisions.

We studied the impact of false prompts on the proportion of true prompts acted on in Chapter 9. Non-expert readers were trusting of both true and false prompts and no difference in terms of prompted target sensitivity was observed between images with and without false prompts. The presence of false prompts significantly increased false positive responses, increasing linearly with the number of false prompts. Expert readers are likely to more carefully consider their actions on prompts and therefore not have as dramatic of an increase in false responses. However, since expert readers are known to increase their recall rates with CAD, it is clear that false prompts have an impact.

As has been previously reported, CAD can significantly alter visual search, resulting in more attention being focused around prompts than the rest of the image (Hatton et al., 2004; Drew et al., 2012; Helbren et al., 2015; Drew et al., 2020). These studies operated CAD with prompts displayed from image onset (or they appeared throughout the CT colonography video in Helbren et al. (2015)). In Chapter 6, we demonstrated that when CAD was operated as a second reader, the overall image coverage was the same in the no-CAD and CAD conditions. The initial search before CAD prompts are activated allows readers to interpret the image without being biased by prompts. We expect that this pre-CAD search will be truncated compared to no-CAD search due to the safety-net effect we observed in Chapter 6, but it is not clear whether this will lead to a reduction in the number of targets fixated or detected prior to prompt activation.

Confidence in decisions on mammograms has been reported to be affected by prevalence (Gur et al., 2007) and expectation of prevalence (Reed et al., 2014). One study by Tchou et al. (2010) demonstrated that CAD can impact reader image confidence ratings on mammograms, with a change in confidence in 22% of cases (14% increased confidence and 8% decreased). The confidence ratings in Tchou et al. (2010) were for the overall image, measuring the effect of CAD on diagnostic confidence rather than detection. Confidence ratings will likely be increased on regions where CAD is in agreement with reader judgement, even if this is a false reassurance for regions that are actually normal. How much of an impact the presence of a false prompt will have on confidence in a decision on another region is not clear, but will likely depend on whether the region marked by the reader is also prompted.

Ten readers had agreed to participate in this study. This was a relatively small number of readers compared to the studies completed with non-expert readers, but is similar to other expert reader studies; the number of readers ranged from six to seventeen (median 8.5) across CAD observer studies (Nishikawa et al., 2012; Evans et al., 2013; Hupse et al., 2013; Rodríguez-Ruiz et al., 2019a; Watanabe et al., 2019; Wolfe et al., 2021b). The sample size is based upon those who were based at the breast centre and were willing to give up their time. The time constraint also limited the number of cases we were able to use, with each session planned for an hour with time for training and any necessary discussions about the study.

We did not have access to a commercial CAD algorithm for this study. The CAD prompts used in this study were either placed by consultant breast radiographers or for the majority of calc prompts were modelled on their placements. This gave us the freedom to decide on the exact number of prompts for each image type. Since we had the ground truth for abnormal prompts, we were able to select the exact prompt sensitivity for each abnormality. However, it is not clear whether the false prompts placed by the radiographers would also have been placed by commercial CAD algorithm.

10.4 Summary

The purpose of this study was to investigate the impact of prompts on both visual search and behaviour on a clinical radiology setup, with eye movements tracked using the methodology outlined in Chapter 8. The main focus was on false prompt distraction, in terms of the proportion of prompts fixated and acted on, and how multiple prompts for the same case interact to influence reader behaviour. However, the data collection was unable to be completed due to government restrictions.

We used 92 cases selected from the OPTIMAM Mammography Image Database from a total of 6,500 cases. Of the 92 cases selected, 36 were abnormal and 56 were normal. The abnormal cases were deemed to be 'very subtle' and contained either a single mass or microcalcification cluster with no other abnormalities or benign features. The density distributions were matched between the abnormal and normal cases using an AI model that predicted reader assessed density.

The CAD operating points were chosen as 80% sensitivity with 1.6 false prompts per case and 90% sensitivity with 0.8 false prompts per case for masses and microcalcification clusters, respectively. False prompts were added to the cases by two consultant breast radiographers. Due to an insufficient number of false calcification prompts to meet the operating point requirement, further prompts were added to cases using a serial top-hat filtering technique to identify potential regions for calcification prompts.

The experiment was due to take place in a clinical reporting room on a dual screen radiology workstation. The experiment software allows readers to zoom on images and change between CC and MLO views. Readers mark regions that they believe to be abnormal and provide a confidence rating between 1 and 5 for each mark placed, along with the abnormality type. The cases would be interpreted twice over two sessions, with and without CAD.

The overall performance would be assessed in terms of their JAFROC score, where it is expected that CAD will improve sensitivity and increase the number of false positives. Expert readers are not expected to rely as heavily on prompts as the non-experts in our previous experiments, but it is clear from literature that search and confidence can be affected by the presence of prompts. The extent to which false prompts distract readers from targets in terms of both their visual search and their sensitivity is not yet clear. In view of the COVID-19 pandemic, we have not been permitted to run the study.

Chapter 11

Conclusions and Future Work

11.1 Summary

This thesis explored the use of CAD in mammography and human-CAD interaction. We outlined the necessity for a solution such as CAD to improve cancer detection at screening and to reduce clinical workload. The relationship between the reader and CAD is complex and has not yet been optimised. The work described in this thesis aimed to gain a deeper understanding of this interplay by replicating mammographic CAD in the lab in visual search studies. The results of these experiments highlighted both how CAD can be useful and where improvements need to be made, as well as providing insight into how reader behaviour is affected by prompts.

Breast screening in the UK has been reported to reduce mortality rates. As part of the screening programme, each mammogram is read by two expert readers. Despite this, some cancers are missed at screening due to the difficulty of detecting subtle and infrequent abnormalities. There also exists a critical shortage of expert readers in the UK, which has been under more pressure with the age expansion trial in recent years increasing the screening age range from 50–70 to 47–73. CAD systems have been investigated as a potential solution, but are yet to be adopted in the UK. If it can be demonstrated that single reading with CAD can achieve similar or better detection rates and does not increase recall rates compared to double reading, it could be used instead of double reading, reducing clinical workload. There is also an argument that single reading with CAD would be accepted with a slight drop in the detection rate if recall rates were also to reduce.

Previous studies evaluating the efficacy of CAD have yet to demonstrate that single reading with CAD is superior to double reading and the use of CAD often leads to a significant increase in the recall rate. The lack of improvement with CAD is in part due to the imperfections in current CAD software, where normal regions are often prompted and cancers are missed. However, even if the standalone performance of CAD was improved, the combination with a human reader needs to be optimised to gain the full benefit. Readers will often dismiss correct CAD prompts (under-trust), or experience a reduction in sensitivity when CAD fails to mark a cancer and an increase in the recall rate by marking false prompts (over-trust). CAD methodologies such as interactive CAD offer an alternative approach to traditional CAD and typically also provide prompt-level and case-level confidence values. This extra information may improve trust in the CAD system, which will also be improved by not being subjected to an excess of false prompts.

Visual search experiments provide a deeper understanding of how readers interpret images and their relationship with CAD, and often use eye tracking to further analyse search behaviour. While CAD has been shown to improve sensitivity for targets it marks, it often reduces sensitivity for targets that it fails to mark. This is a result of overreliance of readers on CAD prompts, where the lack of a prompt leads observers to believe the suspicious area is less likely to be a target or in some cases more likely to be benign. CAD prompts displayed from image onset were found to significantly reduce the proportion of the image that is viewed, with attention focused around prompts. At low prevalence, the miss errors for cancers in mammography and for targets missed by CAD have been shown to increase. There is some evidence that for non-expert readers that this prevalence effect for unprompted targets is eliminated with the use of interactive CAD. The research presented in this thesis aimed to build on the results of these studies and answer the remaining questions surrounding the use of CAD in mammography.

For all experiments conducted in this thesis, we used non-expert readers and synthetic images. The targets used were either malignant microcalcification clusters or Gaussian blobs simulating masses. Previous studies investigated how displaying prompts from image onset affects search behaviour. When CAD is operated as a second reader, there is an initial unaided view of the image where readers are not directly influenced by prompts. However, the fact that this view is preliminary to a further search with CAD leads to the possibility of a safety-net effect where the initial search is truncated compared to search without CAD. We conducted an eye tracking study to investigate the existence of this effect and how readers are influenced by prompts.

The results of our study (Chapter 6) showed that the initial unaided search was reduced in thoroughness in terms of coverage, trial time, and proportion of targets fixated – evidence of a safety-net effect. Cross-sectional CAD efficacy studies use this initial unaided search as the no-CAD condition and compare it to the search with CAD. The fact that this initial search appears to be indirectly influenced by the anticipation of CAD prompts suggests that studies should use separate no-CAD and CAD conditions. However, readers would be required to complete both conditions, resulting in additional work. If each reader only completed one condition, there would be issues with power and reader variability. The overall search with CAD was found to be equal to unaided reading in terms of image coverage, despite lasting longer on average. This shows that when CAD is operated as a second reader, the overall search thoroughness is unaffected, compared to displaying prompts from image onset. We did not observe a sensitivity cost for unprompted targets in our second reader CAD study, but this should be verified with a greater number of images containing unprompted targets.

Interactive CAD systems have been shown to be capable of improving sensitivity of expert readers without increasing the false positive rate. The use of prompt and image scores provides more detailed information to the reader, and influences how they interpret images. Withholding prompts until they are queried results is a different mode of image interpretation than traditional CAD that is more likely to resemble unaided viewing. Through two eye tracking studies, we aimed to explore how search with interactive CAD compared to unaided viewing, and how prompt and image scores influenced behaviour.

In Chapter 7, we demonstrated that interactive CAD did not impact the overall image coverage, suggesting that readers search in a similar way as in unaided viewing. The likelihood of readers accepting prompts as a target scaled with prompt confidence, and both the time spent on images and the false response rate scaled with the overall image score. When both a prompt confidence value and an image score were present, it was the confidence value that determined whether a reader would act on the prompt.

With an interactive CAD system, to make the most of the prompts, readers must have a knowledge of where abnormalities might appear and query appropriate regions. With non-experts searching simulated images lacking anatomical structure, we did not see an improvement in observer sensitivity with CAD, even when an image score was provided. Some level of recognition as a potential target is required to query a region. For targets that were not marked by participants, we showed that even when they were fixated, only a small fraction were then queried, meaning they would not gain the benefit of a true prompt on those targets.

In our final CAD observer study (Chapter 9), we focused on how readers behaved when there were multiple CAD prompts present. CAD algorithms in mammography use different prompt types to indicate which potential abnormality they mark. Most visual search studies with CAD use a single prompt type and operating point. Therefore, it is not well understood how the presence of different types of prompts on the same image can influence how observers interpret prompts. We conducted an online experiment to investigate this, using a single CAD condition and two target types (masses and microcalcification clusters).

In our study, we found that participants became overreliant on the prompts, where the false response rate was significantly increased with the number of prompts. However, this result was consistent across images with false prompts that indicated the same and different target type. Therefore, the increase in false positive rates were due to the fact there were multiple prompts rather than different prompt types. Prompted target sensitivity was unaffected by the presence of false prompts, whether or not the false prompt was the same type as the true prompt. This experiment also demonstrated the viability of running studies online to investigate CAD with non-experts. The sensitivity rate was similar in this study compared with the other studies, suggesting that they were appropriately engaged with the task as they would be in a lab setting.

To accommodate the limitations of eye tracking systems, visual search studies with expert medical readers often simplify the experimental setup, such as using only a single monitor. It is desirable to replicate normal clinical conditions as far as possible, and thus we explored the feasibility of using eye tracking glasses with a dual-screen mammography setup. In Chapter 8, we carried out an experiment in which participants searched for mass-like targets in synthetic images. A validation procedure was repeated after each minute of search to assess the accuracy and precision of the eye tracking glasses.

The eye tracking glasses used were lightweight, and relatively inexpensive compared with similar devices. To track the position of the displays and the participant-screen distance, we used the built-in surface and head-pose tracking software. Gaze coordinates were mapped onto the surfaces defined on the displays. The accuracy significantly decayed at the start of the experiment, but remained relatively stable following this for the next 10 minutes. We determined that recalibrations should be performed every 5 minutes during an experiment to maintain a reasonable accuracy ($<1.25°$) and precision ($<0.4°$).

An experiment with expert readers was planned to be conducted to investigate how they interacted with multiple CAD prompts, using the dual-screen eye tracking setup (Chapter 10). However, the data collection could not be completed due to COVID-19 restrictions. We designed the experiment software to allow for observers to zoom on mammograms and change between image views as they would in the clinic. Readers would also provide confidence ratings on every response, which would allow us to monitor how prompts affected the confidence in their decisions. The main aim of this experiment was to determine the effect of false prompt distraction and how multiple prompts on the same case affect behaviour, which could be compared to experiments with non-expert readers.

11.2 Contributions to Knowledge

The identification of a safety-net effect when operating CAD as a second reader has implications for CAD observer studies and has the potential for reducing the performance of radiologists. This is the first eye tracking study to investigate the search behaviour of observers with and without prompts, with CAD operated in second reader mode as with most CAD systems in mammography (Du-Crow et al., 2019). Previous work demonstrated that the overall image coverage is reduced with prompts (displayed from image onset). This was not found to be the case in our study. This argues for the importance of an initial unaided view. For evaluating the efficacy of CAD, we argue that it would be more appropriate to use a study design with separate CAD and no-CAD conditions to avoid the consequences of the safety-net effect.

Interactive CAD systems reduce the number of false prompts encountered by the reader compared to traditional CAD, and have been shown to improve expert reader sensitivity without increasing the recall rate. Non-expert readers were unable to gain improvements in sensitivity in part due to the nature of the images, which lacked clear anatomical structures (Du-Crow et al., 2020). The absence of improvement in performance highlights a potential limitation of interactive prompts, since targets that are unnoticed or are not recognised as a potential target will not be queried and therefore the reader will not benefit from a prompt. The study also demonstrated that the impact of image scores is the same for non-expert readers as it is for expert readers, with viewing time and false positive responses increasing for high scores. We expanded upon this and showed that the prompt confidence, not the image score, would determine the likelihood of a participant marking a prompt as a target.

CAD systems in mammography often display different types (shapes, symbols, etc.) to denote the different abnormalities present in mammograms. The interaction of different prompt types on reader behaviour has not been well studied. We showed that when targets were prompted, the presence of a false prompt did not affect observer sensitivity, whether the false prompt was the same type as the target or the alternate type. This is further evidence that second reader CAD does not lead to false prompt distraction for prompted targets. However, false prompts significantly increased the false positive response rate. This effect was also unaffected by the type of prompts present.

For observer experiments using eye tracking in realistic clinical settings, eye movements must be tracked across large or multiple displays. We outlined a methodology for tracking reader eye movements across a dual-screen mammography setup. This was achieved using eye tracking glasses with a built in surface tracking module. However, the spatial accuracy of the glasses was too low (approximately 1.25 degrees visual angle) for use in experiments in mammography, and therefore requires further work to improve the accuracy to around 0.5 degrees.

11.3 Future Work

11.3.1 Unprompted targets

The results from Chapter 6 suggest that by operating CAD as a second reader, the miss cost of unprompted targets may be reduced or even eliminated. The increased error rate for unprompted targets was reported in studies operating CAD from image onset (Alberdi et al., 2004; Drew et al., 2012; Russell and Kunar, 2012; Kunar et al., 2017; Ionescu et al., 2018). In second reader mode, the extent to which the initial unaided viewing of the images can nullify this effect remains to be established by testing this with a greater number of images containing unprompted targets than were included in our study. It is important for any study investigating this to maintain a reasonable operating point for CAD. That is to say, not reduce the CAD sensitivity as a means to increase the number of unprompted targets, since the trust level may be reduced as a consequence. Instead, it is necessary to run a study with a greater number of images.

It would also be of interest to determine whether the presence of false prompts in images with unmarked targets further reduces observer sensitivity for second reader CAD. A study by Russell and Kunar (2012) found no difference in error rates between unprompted target trials with and without false prompts. However, Ionescu et al. (2018) reported that there was a significant reduction in sensitivity for unprompted targets when false prompts were present compared to when they were absent. Since both of these studies displayed prompts from image onset, a study could investigate how second reader CAD would compare. Furthermore, it could also be investigated whether the unprompted target sensitivity is affected when there are multiple prompt types present, which follows on from the study in Chapter 9. We had insufficient unprompted targets to measure this in our study, but it would have been of interest to know whether the type of the false prompt (same type as target or different) would have had an impact. It was clear that sensitivity was unaffected for prompted targets but unprompted targets are where false positive distraction has been reported previously (Russell and Kunar, 2012; Ionescu et al., 2018).

With interactive CAD, the significant increase in miss cost at low prevalence for unprompted targets can be eliminated (Drew et al., 2020). In our study, when prompts were available, participants acted in line with the confidence score as opposed to the image score. However, what was not clear was whether the image score would increase in importance for images where CAD failed to mark a target. In our study (Chapter 7), the image score impacted the number of false positive responses, many of which were not prompted. Since about 90% of cancer cases have an image score of 9 or 10 with one clinical interactive CAD system (FDA, 2018), it may be that the image score can further improve the detection of unprompted targets, but this should be tested directly.

11.3.2 Multiple targets

In Chapter 9, our study used images with a single target present. We were constrained by the fact that we had to reuse images from previous studies in order to estimate target detection rates. However, it is also of interest to test how multiple prompts affect reader behaviour in images with multiple abnormalities present. In mammography, microcalcification clusters may be present in cases with a mass. A previous study has suggested that in cases with multiple abnormalities, there is a satisfaction of search effect caused by a suppression of recognition for additional abnormalities that differ from the first one that was detected (Mello-Thoms et al., 2014). The use of a non-interactive CAD system may overcome satisfaction of search if prompts are able to make readers re-examine abnormalities that they have fixated but dismissed. A study could compare the detection rate of multiple abnormalities in aided and unaided viewing. Readers might be more reluctant to accept further prompts if they have already reported on one target.

11.3.3 Reducing the impact of false prompts

The increase in false positive responses due to false positive prompts has been discussed throughout this thesis. CAD systems are constantly updated to improve their sensitivity and crucially their specificity. There are also algorithms that reduce the number of false prompts per image by analysing and removing the prompts of commercial CAD systems (Mayo et al., 2019). However, because false prompts are unavoidable, methods should be investigated to improve how readers deal with them. This will also be influenced by how much readers trust CAD. By indicating the confidence of CAD prompts, readers may keep their trust when they are confronted with a false prompt they would have dismissed anyway if the prompt is accompanied by a low confidence rating. However, if false prompts have high confidence ratings, they are more likely to be recalled. Varying the prompt size to indicate confidence has been shown to increase the recall rate for normal and cancer cases (Gilbert et al., 2008a). Therefore, the improvement in detection rate can come at the cost of an increase in false positives if readers are equally influenced by the confidence of true and false prompts.

As suggested by Jorritsma et al. (2015), other methods that could improve performance with CAD are providing global and local rationales. A global rationale would explain to readers how the CAD system works and what features are likely to cause false positive errors, which might make readers more confident in dismissing false prompts if they can identify it as a common error. The local rationale provides an explanation of why a specific region was marked. This may be difficult for many CAD systems in mammography due to the complex processes that lead to the placement of a prompt. However, if specific details can be presented, or perhaps similar regions from other cases can be displayed, readers may have a greater trust in the system and be able to identify false prompts more easily. These methods have not been studied well in the context of mammography but have the potential to improve performance with CAD. Studies could compare various methods of enhancing prompts with this information to measure their impact on performance. It is always important that these methods are implemented in such a way that does not make their use cumbersome and are developed with the end user in mind (Filice and Ratwani, 2020).

11.3.4 Reader experience

The participants in all experiments in this thesis were non-experts, with expert readers planned to take part in the study described in Chapter 10. The results of this thesis should be verified with expert readers. There have been many studies that have investigated differences between readers of varying experience levels. Even among medical readers, less experienced readers have been shown to differ from more experienced readers in terms of their search technique (Kundel et al., 2007). Furthermore, CAD may be more beneficial in terms of sensitivity for less experienced readers since more experienced readers may already be close to their performance ceiling and will not see as large of a sensitivity increase (Balleyguier et al., 2005). However, there are similarities between non-experts and expert readers in terms of errors and overall performance with CAD. The increase in false positive responses with CAD has been observed in studies with experts and non-experts (including those in this thesis). Furthermore, the increased target miss rate when CAD fails to mark a target has also been observed across experience levels, as has the effect of inattentional blindness. While there are indeed commonalities across reader experience, to be able to test CAD as it is used in mammography, the most effective way to do so is through experiments with expert readers and clinical setups.

11.3.5 Alternative modalities

CAD has been extended to other breast imaging modalities. Digital breast tomosynthesis (DBT) has demonstrated that it can achieve similar sensitivity and improved specificity compared to mammography (Gilbert et al., 2015). CAD has been reported to improve performance with DBT in a number of small-scale studies and, as with CAD in mammography, may increase recall rate due to false prompts (Harkness et al., 2015; Morra et al., 2015; Benedikt et al., 2018). Another modality is automated breast ultrasound (ABUS), which has demonstrated improved sensitivity when used as an adjunct to mammography (Kelly et al., 2010; Brem et al., 2015; Wilczek et al., 2016). One of the main aims with CAD in ABUS is to increase search efficiency (time taken to read cases). CAD was shown to improve efficiency with ABUS without affecting the performance of readers (van Zelst et al., 2018). Similar results have been reported for DBT (Benedikt et al., 2018). This is of course desirable for 3D modalities where reading time is typically longer than for mammography.

Future studies could be conducted investigating how CAD prompts used in these modalities affect visual search, using the eye tracking methodologies outlined by Aizenman et al. (2017) and Dong et al. (2018). One study with CAD in CT colonography videos demonstrated that gaze was attracted to prompts and readers spent less time viewing unprompted areas (Helbren et al., 2015). With these methodologies, since reading time is already longer than mammography, prompts are often shown from image onset. Therefore, it may be that similar results are seen with CAD for DBT and ABUS, where false prompts may distract from unprompted abnormalities.

The overall goal of these proposed studies, and those described in this thesis, is to develop a deeper understanding of human-CAD interaction which can in turn be used to improve CAD algorithms and their effectiveness when combined with expert readers.

References

Aizenman, A., Drew, T., Ehinger, K., Georgian-Smith, D. and Wolfe, J. (2017), 'Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: an eye tracking study', *Journal of Medical Imaging* **4**(4), 045501–1–045501–10.

Al-Ghaib, H. (2015), 'Structural similarity index algorithm for accurate mammogram registration'. [PowerPoint Presentation], Available at: `http://slideplayer.com/slide/9153733/` [Accessed 13 Dec. 2017].

Alamudun, F., Paulus, P., Yoon, H.-J. and Tourassi, G. (2018), 'Modeling sequential context effects in diagnostic interpretation of screening mammograms', *Journal of Medical Imaging* **5**(03), 1.

Alberdi, E., Povyakalo, A., Strigini, L. and Ayton, P. (2004), 'Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography', *Academic Radiology* **11**(18), 909–918.

Alvarez, G. and Cavanagh, P. (2004), 'The capacity of visual short-term memory is set both by visual information load and by number of objects', **15**(2), 106–111.

Argus Science (n.d.), 'Argus Science ETMobile Eye Tracking System'. Available at: `http://www.argusscience.com/file/ETMobileBrochure.pdf` [Accessed 20 Jun. 2022].

Arrington Research (n.d.), 'Head Fixed Eye Tracking System Specifications'. Available at: `http://www.arringtonresearch.com/scenetechinfo.html` [Accessed 20 Jun. 2022].

ASL (n.d.), 'ASL Mobile Eye-5 Glasses'. Available at: `https://est-kl.com/ima` `ges/PDF/ASL/Mobile%20Eye%205%20Specifications.pdf` [Accessed 20 Jun. 2022].

Astley, S. (2004), ‘Computer-based detection and prompting of mammographic abnormalities’, *The British Journal of Radiology* **77**(2), 194–200.

Astley, S. and Gilbert, F. (2004), ‘Computer-aided detection in mammography’, *Clinical Radiology* **59**(5), 390–399.

Astley, S., Mistry, T., Boggis, C. and Hillier, V. (2003), ‘Should we use humans or a machine to pre-screen mammograms?’, *Digital Mammography* pp. 476–480.

Atchison, D. and Smith, G. (2000), *Optics of the human eye*, Butterworth-Heinemann Medical, Oxford.

Ayer, T., Chen, Q. and Burnside, E. S. (2013), ‘Artificial neural networks in mammography interpretation and diagnostic decision making’, *Computational and Mathematical Methods in Medicine* **2013**, 1–10.

Azavedo, E., Zackrisson, S., Mejàre, I. and Heibert Arnlind, M. (2012), ‘Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic review’, *BMC Medical Imaging* **12**(1).

Bach, M. (1996), ‘The Freiburg visual acuity test—automatic measurement of visual acuity’, *Optometry and Vision Science* **73**(1), 49–53.

Bahl, M., Baker, J., Kinsey, E. and Ghate, S. (2015), ‘Architectural Distortion on Mammography: Correlation With Pathologic Outcomes and Predictors of Malignancy’, *American Journal of Roentgenology* **205**(6), 1339–1345.

Baker, J., Rosen, E., Lo, J., Gimenez, E., Walsh, R. and Soo, M. (2003), 'Computer-Aided Detection (CAD) in Screening Mammography:Sensitivity of Commercial CAD Systems for Detecting Architectural Distortion', *American Journal of Roentgenology* **181**(4), 1083–1088.

Balleyguier, C., Kinkel, K., Fermanian, J., Malan, S., Djen, G., Taourel, P. and Helenon, O. (2005), 'Computer-aided detection (CAD) in mammography: Does it help the junior or the senior radiologist?', **54**(1), 90–96.

Balta, C., Rodriguez-Ruiz, A., Mieskes, C., Karssemeijer, N. and Heywang-Köbrunner, S. H. (2020), Going from double to single reading for screening exams labeled as likely normal by AI: what is the impact?, *in* C. V. Ongeval, N. Marshall and H. Bosmans, eds, '15th International Workshop on Breast Imaging (IWBI2020)', SPIE.

Bao, B. and Prasad, A. S. (2019), Targeting CSC in a most aggressive subtype of breast cancer TNBC, *in* 'Advances in Experimental Medicine and Biology', Springer International Publishing, pp. 311–334.

Bassett, L. and Conner, K. (2003), The abnormal mammogram, *in* D. Kufe, R. Pollock, R. Weichselbaum, R. Bast, T. Gansler, J. Holland and E. Frei, eds, 'Cancer Medicine 6', 6 edn, BC Decker, Hamilton, Ont. Lewiston, NY.
URL: *https://www.ncbi.nlm.nih.gov/books/NBK12642/*

Bazzani, A., Bevilacqua, A., Bollini, D., Brancaccio, R., Campanini, R., Lanconelli, N., Riccardi, A. and Romani, D. (2001), 'An SVM classifier to separate false signals from microcalcifications in digital mammograms', *Physics in Medicine and Biology* **46**(6), 1651–1663.

Benedikt, R. A., Boatsman, J. E., Swann, C. A., Kirkpatrick, A. D. and Toledano,

A. Y. (2018), ‘Concurrent computer-aided detection improves reading time of digital breast tomosynthesis and maintains interpretation performance in a multireader multicase study’, **210**(3), 685–694.

Berbaum, K., Franken, E., Dorfman, D., Rooholamini, S., Kathol, M., Barloon, T., Behlke, F., Sato, Y., Lu, C., el Khoury, G., Flickinger, F. and Montgomery, W. (1990), ‘Satisfaction of search in diagnostic radiology’, *Investigative Radiology* **25**(2), 133–139.

Berbaum, K., Krupinski, E., Schartz, K., Caldwell, R., Madsen, M., Hur, S., Laroia, A., Thompson, B., Mullan, B. and Franken, E. (2015), ‘Satisfaction of Search in Chest Radiography 2015’, *Academic Radiology* **22**(11), 1457–1465.

Berlin, L. (2007), ‘Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five Decades?’, *American Journal of Roentgenology* **188**(5), 1173–1178.

Bilsky, A. B. and Wolfe, J. M. (1995), ‘Part—whole information is useful in visual search for size × size but not orientation × orientation conjunctions’, *Perception & Psychophysics* **57**(6), 749–760.

Bird, R., Wallace, T. and Yankaskas, B. (1992), ‘Analysis of cancers missed at screening mammography’, *Radiology* **184**(3), 613–617.

Birdwell, R., Bandodkar, P. and Ikeda, D. (2005), ‘Computer-aided detection with screening mammography in a university hospital setting’, *Radiology* **236**(2), 451–457.

Bolivar, A. V., Gomez, S. S., Merino, P., Alonso-Bartolomé, P., Garcia, E. O., Cacho, P. M. and Hoffmeister, J. W. (2010), ‘Computer-aided detection system applied to full-field digital mammograms’, *Acta Radiologica* **51**(10), 1086–1092.

Breastcancer.org (2016), 'LCIS and Breast Cancer Risk'. Available at: `https://www.breastcancer.org/symptoms/types/lcis/cancer_risk` [Accessed 30 Mar. 2021].

Brem, R., Tabár, L., Duffy, S., Inciardi, M., Guingrich, J., Hashimoto, B., Lander, M., Lapidus, R., Peterson, M., Rapelyea, J., Roux, S., Schilling, K., Shah, B., Torrente, J., Wynn, R. and Miller, D. (2015), 'Assessing improvement in detection of breast cancer with three-dimensional automated breast us in women with dense breast tissue: the somoinsight study', *Radiology* **274**, 663–673.

Bruno, M. (2017), '256 Shades of gray: uncertainty and diagnostic error in radiology', *Diagnosis* **4**(3).

Campanini, R., Dongiovanni, D., Iampieri, E., Lanconelli, N., Masotti, M., Palermo, G., Riccardi, A. and M., R. (2004), 'A novel featureless approach to mass detection in digital mammograms based on support vector machines.', *Physics in Medicine and Biology* **49**(6), 961–975.

Cancer Research UK (2020), 'Breast screening'. Available at: `https://www.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/screening/breast-screening#:~:text=These%20tiny%20breast%20cancers%20are,every%201%2C000%20women%20having%20screening.` [Accessed 7 Mar. 2021].

Cancer Research UK (2021*a*), 'Breast cancer statistics'. Available at: `http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer` [Accessed 7 Mar. 2021].

Cancer Research UK (2021*b*), 'About breast cancer'. Available at: `http://www.cancerresearchuk.org/about-cancer/breast-cancer/about` [Accessed 7 Mar. 2021].

Carney, P. A., Bogart, T. A., Geller, B. M., Haneuse, S., Kerlikowske, K., Buist, D. S. M., Smith, R., Rosenberg, R., Yankaskas, B. C., Onega, T. and Miglioretti, D. L. (2012), 'Association between time spent interpreting, level of confidence, and accuracy of screening mammography', *American Journal of Roentgenology* **198**(4), 970–978.

Chakraborty, D. P. and Berbaum, K. S. (2004), 'Observer studies involving detection and localization: Modeling, analysis, and validation', *Medical Physics* **31**(8), 2313–2330.

Chakraborty, D. P. and Yoon, H.-J. (2009), JAFROC analysis revisited: figure-of-merit considerations for human observer studies, *in* B. Sahiner and D. J. Manning, eds, 'Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment', SPIE.

Chan, H., Doi, K., Galhotra, S., Vyborny, C., MacMahon, H. and Jokich, P. (1987), 'Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography', *Medical Physics* **14**(4), 538–548.

Chan, H., Doi, K., Vybrony, C., Schmidt, R., Metz, C., Lam, K., Ogura, T., Wu, Y. and MacMahon, H. (1990), 'Improvement in Radiologists' Detection of Clustered Microcalcifications on Mammograms', *Investigative Radiology* **25**(10), 1102–1110.

Chan, L. and Hayward, W. (2013), 'Visual search', *Wiley Interdisciplinary Reviews: Cognitive Science* **4**(4), 415–429.

Chen, Y. and Gale, A. (2010*a*), Intelligent Computing Applications Based on Eye Gaze: Their Role in Medical Image Interpretation, *in* D. Huang, M. McGinnity,

L. Heutte and X. Zhang, eds, ‘Advanced Intelligent Computing Theories and Applications. ICIC 2010. Communications in Computer and Information Science, vol 93’, Springer, Berlin, Heidelberg, pp. 320–325.

Chen, Y. and Gale, A. (2010*b*), ‘Using eye gaze in intelligent interactive imaging training’, *Proceedings of the 2010 workshop on Eye gaze in intelligent human machine interaction - EGIHMI ’10* .

Chougrad, H., Zouaki, H. and Alheyane, O. (2018), ‘Deep convolutional neural networks for breast cancer screening’, *Computer Methods and Programs in Biomedicine* **157**, 19–30.

Cole, E. B., Zhang, Z., Marques, H. S., Nishikawa, R., Hendrick, R. E., Yaffe, M. J., Padungchaichote, W., Kuzmiak, C., Chayakulkheeree, J., Conant, E. F., Fajardo, L. L., Baum, J., Gatsonis, C. and Pisano, E. (2012), ‘Assessing the stand-alone sensitivity of computer-aided detection with cancer cases from the digital mammographic imaging screening trial’, *American Journal of Roentgenology* **199**(3), W392–W401.

Comstock, C. E., Gatsonis, C., Newstead, G. M., Snyder, B. S., Gareen, I. F., Bergin, J. T., Rahbar, H., Sung, J. S., Jacobs, C., Harvey, J. A., Nicholson, M. H., Ward, R. C., Holt, J., Prather, A., Miller, K. D., Schnall, M. D. and Kuhl, C. K. (2020), ‘Comparison of abbreviated breast MRI vs digital breast tomosynthesis for breast cancer detection among women with dense breasts undergoing screening’, *JAMA* **323**(8), 746.

Cousineau, D. (2005), ‘Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method’, *Tutorials in Quantitative Methods for Psychology* **1**(1), 42–45.

Cowan, N. (2001), ‘The magical number 4 in short-term memory: A reconsideration of mental storage capacity’, *Behavioral and Brain Sciences* **24**(1), 87–114.

Cowell, C., Weigelt, B., Sakr, R., Ng, C., Hicks, J., King, T. and Reis-Filho, J. (2013), 'Progression from ductal carcinomain situto invasive breast cancer: Revisited', *Molecular Oncology* **7**(5), 859–869.

Cunningham, C., Drew, T. and Wolfe, J. (2016), 'Analog Computer-Aided Detection (CAD) information can be more effective than binary marks', *Attention, Perception, & Psychophysics* **79**(2), 679–690.

Dalmaijer, E., Mathôt, S. and Van der Stigchel, S. (2014), 'PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eye tracking experiments', *Behavior Research Methods* **46**, 913–921.

Dean, J. and Ilvento, C. (2006), 'Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers', *American Journal of Roentgenology* **187**(1), 20–28.

Delabarre, E. B. (1898), 'A method of recording eye-movements', *The American Journal of Psychology* **9**(4), 572.

DeMartini, W. and Lehman, C. (2008), 'A Review of Current Evidence-Based Clinical Applications for Breast Magnetic Resonance Imaging', *Topics in Magnetic Resonance Imaging* **19**(3), 143–150.

Dhungel, N., Carneiro, G. and Bradley, A. (2017), 'A deep learning approach for the analysis of masses in mammograms with minimal user intervention', *Medical Image Analysis* **37**, 114–128.

Dierkes, K., Kassner, M. and Bulling, A. (2019), A fast approach to refraction-aware eye-model fitting and gaze prediction, *in* 'Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications', ACM.

DM Challenge (2017), 'The Digital Mammography DREAM Challenge'. Available at: `https://www.synapse.org/#!Synapse:syn4224222/wiki/` [Accessed 9 Mar. 2021].

Dong, L., Chen, Y., Gale, A. and Phillips, P. (2016), 'Eye Tracking Method Compatible with Dual-screen Mammography Workstation', *Procedia Computer Science* **90**, 206–211.

Dong, L., Tang, Q., Gale, A., Bernardi, D. and Chen, Y. (2018), Analysis of visual search behaviour from experienced radiologists interpreting digital breast tomosynthesis (DBT) images: a pilot study, *in* R. M. Nishikawa and F. W. Samuelson, eds, 'Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment', SPIE.

Drew, T., Aizenman, A., Thompson, M., Kovacs, M., Trambert, M., Reicher, M. and Wolfe, J. (2015), 'Image toggling saves time in mammography', *Journal of Medical Imaging* **3**(1), 011003.

Drew, T., Cunningham, C. and Wolfe, J. (2012), 'When and Why Might a Computer-aided Detection (CAD) System Interfere with Visual Search? An Eye-tracking Study', *Academic Radiology* **19**(10), 1260–1267.

Drew, T., Guthrie, J. and 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.

Drew, T., Võ, M. and Wolfe, J. (2013), 'The Invisible Gorilla Strikes Again', *Psychological Science* **24**(9), 1848–1853.

Drew, T. and Williams, L. H. (2017), 'Simple eye-movement feedback during visual search is not helpful', *Cognitive Research: Principles and Implications* **2**(1).

Drew, T., Williams, L. H., Aldred, B., Heilbrun, M. E. and Minoshima, S. (2018), ‘Quantifying the costs of interruption during diagnostic radiology interpretation using mobile eye-tracking glasses’, *Journal of Medical Imaging* **5**(03), 1.

Du-Crow, E., Astley, S. M. and Hulleman, J. (2019), ‘Is there a safety-net effect with computer-aided detection?’, *Journal of Medical Imaging* **7**(02), 1.

Du-Crow, E., Astley, S. M. and Hulleman, J. (2020), Suspicious minds: effect of using a lesion likelihood score on reader behaviour with interactive mammographic CAD, *in* C. V. Ongeval, N. Marshall and H. Bosmans, eds, ‘15th International Workshop on Breast Imaging (IWBI2020)’, SPIE.

Du, T., Zhu, L., Levine, K. M., Tasdemir, N., Lee, A. V., Vignali, D. A. A., Houten, B. V., Tseng, G. C. and Oesterreich, S. (2018), ‘Invasive lobular and ductal breast carcinoma differ in immune response, protein translation efficiency and metabolism’, *Scientific Reports* **8**(1).

Duchowski, A. (2017), ‘Eye Tracking Techniques’, *Eye Tracking Methodology* pp. 49–57.

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.

Duncan, J. and Humphreys, G. (1992), ‘Beyond the search surface: Visual search and attentional engagement', *Journal of Experimental Psychology: Human Perception and Performance* **18**, 578–588.

Duncan, J. and Humphreys, G. W. (1989), 'Visual search and stimulus similarity.', **96**(3), 433–458.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G. and Beck, H. P. (2003), 'The role of trust in automation reliance', *International Journal of Human-Computer Studies* **58**(6), 697–718.

Efron, B. and Tibshirani, R. J. (1993), *An introduction to the bootstrap*, Chapman & Hall/CRC.

Ehinger, B. V., Groß, K., Ibs, I. and König, P. (2019), 'A new comprehensive eye-tracking test battery concurrently evaluating the pupil labs glasses and the EyeLink 1000', *PeerJ* **7**, e7086.

Elmore, J. G. (1997), 'The impact of clinical history on mammographic interpretations', *JAMA: The Journal of the American Medical Association* **277**(1), 49.

Elmore, J., Jackson, S., Abraham, L., Miglioretti, D., Carney, P., Geller, B., Yankaskas, B., Kerlikowske, K., Onega, T., Rosenberg, R., Sickles, E. and Buist, D. (2009), 'Variability in Interpretive Performance at Screening Mammography and Radiologists' Characteristics Associated with Accuracy', *Radiology* **253**(3), 641–651.

Elmore, J., Wells, C., Lee, C., Howard, D. and Feinstein, A. (1994), 'Variability in radiologists' interpretations of mammograms', *New England Journal of Medicine* **331**(22), 1493—-1499.

Eng, H. Y., Chen, D. and Jiang, Y. (2005), 'Visual working memory for simple and complex visual stimuli', **12**(6), 1127–1133.

Evans, K., Haygood, T., Cooper, J., Culpan, A. and Wolfe, J. (2016), 'A half-second glimpse often lets radiologists identify breast cancer cases even when viewing the mammogram of the opposite breast', *Proceedings of the National Academy of Sciences* **113**(37), 10292–10297.

Evans, K. K., Birdwell, R. L. and 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), 1–6.

Fauci, F., Bagnasco, S., Bellotti, R., Cascio, D., Cheran, S., Carlo, F. D., Nunzio, G. D., Fantacci, M., Forni, G., Lauria, A., Torres, E. L., Magro, R., Masala, G., Oliva, P., Quarta, M., Raso, G., Retico, A. and Tangaro, S. (2004), Mammogram segmentation by contour searching and massive lesion classification with neural network, *in* 'IEEE Symposium Conference Record Nuclear Science 2004.', IEEE.

FDA (2008), 'FDA Radiological Devices Panel Meeting'. March 2008, Briefing Package.

FDA (2018), '510(k) summary transpara'. Available at: `https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K181704` [Accessed 3 Jun. 2021].

Filice, R. W. and Ratwani, R. M. (2020), 'The case for user-centered artificial intelligence in radiology', **2**(3), e190095.

Fouad, T. M., Kogawa, T., Shen, D. D. L., Masuda, H., Woodward, R. E.-Z. A., Chavez-MacGregor, M., Alvarez, R. H., Arun, B., Lucci, A., Krishnamurthy, S., Babiera, G., Buchholz, T. A., Valero, V. and Ueno, N. T. (2015), 'Overall survival differences between patients with inflammatory and noninflammatory breast cancer presenting with distant metastasis at diagnosis', *Breast Cancer Research and Treatment* **152**(2), 407—-416.

Freer, P. (2015), ‘Mammographic Breast Density: Impact on Breast Cancer Risk and Implications for Screening’, *RadioGraphics* **35**(2), 302–315.

Freer, T. and Ulissey, M. (2001), ‘Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center’, *Radiology* **220**(3), 781–786.

Gale, A. (2010), ‘Maintaining quality in the UK breast screening program’, *Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment* .

Gandomkar, Z. and Mello-Thoms, C. (2019), ‘Visual search in breast imaging’, *The British Journal of Radiology* **92**(1102), 20190057.

Gandomkar, Z., Tay, K., Brennan, P. C., Kozuch, E. and Mello-Thoms, C. (2018), ‘Can eye-tracking metrics be used to better pair radiologists in a mammogram reading task?’, *Medical Physics* **45**(11), 4844–4856.

Gandomkar, Z., Tay, K., Ryder, W., Brennan, P. C. and Mello-Thoms, C. (2017), ‘iCAP: An Individualized Model Combining Gaze Parameters and Image-Based Features to Predict Radiologists’ Decisions While Reading Mammograms’, *IEEE Transactions on Medical Imaging* **36**(5), 1066–1075.

Gao, Y., Geras, K. J., Lewin, A. A. and 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.

García-Manso, A., García-Orellana, C. J., González-Velasco, H., Gallardo-Caballero, R. and Macías, M. M. (2013), ‘Consistent performance measurement of a system to detect masses in mammograms based on blind feature extraction’, *BioMedical Engineering OnLine* **12**(1), 2.

Georgian-Smith, D., Moore, R., Halpern, E., Yeh, E., Rafferty, E., D'Alessandro, H., Staffa, M., Hall, D., McCarthy, K. and Kopans, D. (2007), 'Blinded Comparison of Computer-Aided Detection with Human Second Reading in Screening Mammography', *American Journal of Roentgenology* **189**(5), 1135–1141.

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

Gibaldi, A. and Sabatini, S. P. (2020), 'The saccade main sequence revised: A fast and repeatable tool for oculomotor analysis', **53**(1), 167–187.

Giger, M. L., Huo, Z., Vyborny, C. J., Lan, L., Bonta, I. R., Horsch, K., Nishikawa, R. M. and Rosenborough, I. (2002), Intelligent CAD workstation for breast imaging using similarity to known lesions and multiple visual prompt aids, *in* M. Sonka and J. M. Fitzpatrick, eds, 'Medical Imaging 2002: Image Processing', SPIE.

Gilbert, F., Astley, S., Boggis, C., McGee, M., Griffiths, P., Duffy, S., Agbaje, O., Gillan, M., Wilson, M., Jain, A., Barr, N., Beetles, U., Griffiths, M., Johnson, J., Roberts, R., Deans, H., Duncan, K. and Iyengar, G. (2008a), 'Variable size computer-aided detection prompts and mammography film reader decisions', *Breast Cancer Research* **10**(4).

Gilbert, F., Astley, S., Gillan, M., Agbaje, O., Wallis, M., James, J., Boggis, C. and Duffy, S. (2008b), 'Single Reading with Computer-Aided Detection for Screening Mammography', *New England Journal of Medicine* **359**(16), 1675–1684.

Gilbert, F., Tucker, L., Gillan, M., Willsher, P., Cooke, J., Duncan, K., Michell, M., Dobson, H., Lim, Y., Purushothaman, H., Strudley, C., Astley, S., Morrish, O., Young, K. and Duffy, S. (2015), 'The TOMMY trial: a comparison of TOMosynthesis with digital MammographY in the UK NHS Breast Screening Programme –

a multicentre retrospective reading study comparing the diagnostic performance of digital breast tomosynthesis and digital mammography with digital mammography alone', *Health Technology Assessment* **19**(4), 1–136.

Godwin, H. J., Hout, M. C., Alexdóttir, K. J., Walenchok, S. C. and Barnhart, A. S. (2021), 'Avoiding potential pitfalls in visual search and eye-movement experiments: A tutorial review', *Attention, Perception, & Psychophysics* .

Gøtzsche, P. and Jørgensen, K. (2013), 'Screening for breast cancer with mammography'.

Green, D. and Swets, J. (1966), *Signal Detection Theory and Psychophysics*, Wiley, New York.

Gur, D., Bandos, A., Fuhrman, C., Klym, A., King, J. and Rockette, H. (2007), 'The prevalence effect in a laboratory environment', *Academic Radiology* **14**(1), 49–53.

Gur, D., Rockette, H. E., Armfield, D. R., Blachar, A., Bogan, J. K., Brancatelli, G., Britton, C. A., Brown, M. L., Davis, P. L., Ferris, J. V., Fuhrman, C. R., Golla, S. K., Katyal, S., Lacomis, J. M., McCook, B. M., Thaete, F. L. and Warfel, T. E. (2003), 'Prevalence effect in a laboratory environment', *Radiology* **228**(1), 10–14.

Halling-Brown, M. D., Warren, L. M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M. G., Wilkinson, L., Given-Wilson, R. M., McAvinchey, R. and Young, K. C. (2020), 'Optimam mammography image database: a large scale resource of mammography images and clinical data'.

Hanley, J. A. and McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (ROC) curve.', *Radiology* **143**(1), 29–36.

Harkness, E. F., Lim, Y. Y., Wilson, M. W., Haq, R., Zhou, J., Tate, C., Maxwell, A. J., Astley, S. M. and Gilbert, F. J. (2015), Initial experience with computer aided detection for microcalcification in digital breast tomosynthesis, *in* L. M. Hadjiiski and G. D. Tourassi, eds, 'Proc. SPIE 9414, Medical Imaging 2015: Computer-Aided Diagnosis', SPIE.

Hatton, J., Wooding, D., Gale, A. and Scott, H. (2004), 'The effect of novel prompts upon radiologists' visual search of mammograms', *Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment* .

Helbren, E., Fanshawe, T., Phillips, P., Mallett, S., Boone, D., Gale, A., Altman, D., Taylor, S., Manning, D. and Halligan, S. (2015), 'The effect of computer-aided detection markers on visual search and reader performance during concurrent reading of CT colonography', *European Radiology* **25**(6), 1570–1578.

Helvie, M., Hadjiiski, L., Makariou, E., Chan, H., Petrick, N., Sahiner, B., Lo, S., Freedman, M., Adler, D., Bailey, J., Blane, C., Hoff, D., Hunt, K., Joynt, L., Klein, K., Paramagul, C., Patterson, S. and Roubidoux, M. (2004), 'Sensitivity of Noncommercial Computer-aided Detection System for Mammographic Breast Cancer Detection: Pilot Clinical Trial', *Radiology* **231**(1), 208–214.

Hendrick, R. (2010), 'Radiation Doses and Cancer Risks from Breast Imaging Studies', *Radiology* **257**(1), 246–253.

Henriksen, E. L., Carlsen, J. F., Vejborg, I. M., Nielsen, M. B. and Lauridsen, C. A. (2018), '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.

Henrot, P., Leroux, A., Barlier, C. and Geninb, P. (2014), 'Breast microcalcifications: The lesions in anatomical pathology', *Diagnostic and Interventional Imaging* **95**(2), 141——152.

Heywang-Köbrunner, S. H., Hacker, A. and Sedlacek, S. (2011), ‘Advantages and Disadvantages of Mammography Screening’, *Breast Care* **6**(3), 199–207.

Hologic Inc. (2014), ‘Understanding ImageChecker CAD 10.0 User Guide’. Available at: `http://www.hologic.com/sites/default/files/package%20inserts/Understanding%20ImageChecker%20CAD%2010.0.%20English.pdf` [Accessed 18 Dec. 2017].

Hooge, I. T. C. and Erkelens, C. J. (1996), ‘Control of fixation duration in a simple search task’, **58**(7), 969–976.

Horowitz, T. (2017), ‘Prevalence in Visual Search: From the Clinic to the Lab and Back Again’, *Japanese Psychological Research* **59**(2), 65–108.

Horowitz, T. S. and Wolfe, J. M. (1998), ‘Visual search has no memory’, **394**(6693), 575–577.

Huey, E. B. (1898), ‘Experiments in the physiology and psychology of reading’, *The American Journal of Psychology* **9**(4), 575–586.

Hulleman, J., Lund, K. and Skarratt, P. A. (2019), ‘Medium versus difficult visual search: How a quantitative change in the functional visual field leads to a qualitative difference in performance’, *Attention, Perception, & Psychophysics* **82**(1), 118–139.

Hulleman, J. and Olivers, C. (2017), ‘The impending demise of the item in visual search’, *Behavioral and Brain Sciences* **40**, 1–69.

Hupse, R., Samulski, M., Lobbes, M., Mann, R., Mus, R., den Heeten, G., Beijerinck, D., Pijnappel, R., Boetes, C. and Karssemeijer, N. (2013), ‘Computer-aided Detection of Masses at Mammography: Interactive Decision Support versus Prompts’, *Radiology* **266**(1), 123–129.

iCAD (2016), 'Mammography: Benefit of computer aided detection'. Clinical Case Study.

Ionescu, G., Harkness, E., Hulleman, J. and Astley, S. (2018), 'A citizen science approach to optimising computer aided detection (cad) in mammography', *Proc. SPIE 10577* .

Ionescu, G. V., Fergie, M., Berks, M., Harkness, E. F., Hulleman, J., Brentnall, A. R., Cuzick, J., Evans, D. G. and Astley, S. M. (2019), 'Prediction of reader estimates of mammographic density using convolutional neural networks', *Journal of Medical Imaging* **6**(03), 1.

Jenkins, J., Sellars, S. J. and Wheaton, M. (2013), 'Guidelines on organising the surveillance of women at higher risk of developing breast cancer in an NHS Breast Screening Programme'. NHSBSP Publication No 73.

Jesinger, R. A. (2014), 'Breast Anatomy for the Interventionalist', *Techniques in Vascular and Interventional Radiology* **17**(1), 3–9.

Jian, W., Sun, X. and Luo, S. (2012), 'Computer-aided diagnosis of breast microcalcifications based on dual-tree complex wavelet transform', *BioMedical Engineering OnLine* **11**(1).

Jiang, Y., Nishikawa, R. M., Wolverton, D. E., Metz, C. E., Giger, M. L., Schmidt, R. A., Vyborny, C. J. and Doi, K. (1996), 'Malignant and benign clustered microcalcifications: automated feature analysis and classification.', *Radiology* **198**(3), 671–678.

Jiang, Y., Nishikawa, R., Schmidt, R., Toledano, A. and Doi, K. (2001), 'Potential of Computer-aided Diagnosis to Reduce Variability in Radiologists' Interpretations of Mammograms Depicting Microcalcifications', *Radiology* **220**(3), 787–794.

Jiang, Z., Das, M. and Gifford, H. C. (2017), ‘Analyzing visual-search observers using eye-tracking data for digital breast tomosynthesis images’, *Journal of the Optical Society of America A* **34**(6), 838.

Jiménez-Gaona, Y., Rodríguez-Álvarez, M. and Lakshminarayanan, V. (2020), ‘Deep-learning-based computer-aided systems for breast cancer imaging: A critical review’, *Applied Sciences* **10**(22), 8298.

Jing, H., Naqa, I. E. and Yang, Y. (2015), Detection and Diagnosis of Microcalcications in Mammography, *in* Q. Li and R. M. Nishikawa, eds, ‘Computer-Aided Detection and Diagnosis in Medical Imaging’, CRC Press, Boca Raton, Florida, pp. 42–53.

Jing, H., Yang, Y. and Nishikawa, R. M. (2010), ‘Detection of clustered microcalcifications using spatial point process modeling’, *Physics in Medicine and Biology* **56**(1), 1–17.

Jørgensen, K. and Gøtzsche, P. (2010), ‘Who evaluates public health programmes? A review of the NHS Breast Screening Programme’, *Journal of the Royal Society of Medicine* **103**(1), 14–20.

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

Karssemeijer, N. (1992), ‘Stochastic model for automated detection of calcifications in digital mammograms’, *Image and Vision Computing* **10**(6), 369–375.

Karssemeijer, N. (1993), ‘Adaptive noise equalization and recognition of microcalcification clusters in mammograms’, *International Journal of Pattern Recognition and Artificial Intelligence* **07**(06), 1357–1376.

Karssemeijer, N. (2015), Detection and Diagnosis of Breast Masses in Mammography, *in* Q. Li and R. M. Nishikawa, eds, ‘Computer-Aided Detection and Diagnosis in Medical Imaging’, CRC Press, Boca Raton, Florida, pp. 21–36.

Karssemeijer, N. and te Brake, G. (1996), ‘Detection of stellate distortions in mammograms’, *IEEE Transactions on Medical Imaging* **15**(5), 611–619.

Kassner, M., Patera, W. and Bulling, A. (2014), ‘Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction’.

Ke, E., Liu, W., Xu, W., Li, L., Zheng, B., Zhang, J. and Zhang, L. (2012), ‘Perceptual mass segmentation using eye-tracking and seed-growing’, *Proc. SPIE 8315, Medical Imaging 2012: Computer-Aided Diagnosis, 831520* .

Keen, J., Keen, J. and Keen, J. (2017), ‘Utilization of Computer-Aided Detection for Digital Screening Mammography in the United States, 2008 to 2016’, *Journal of the American College of Radiology* .

Kelly, K., Dean, J., Comulada, W. and Lee, S. (2010), ‘Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts’, *European Radiology* **20**, 734–742.

Kim, S. J., Moon, W. K., Cho, N., Cha, J. H., Kim, S. M. and Im, J.-G. (2008), ‘Computer-aided detection in full-field digital mammography: Sensitivity and reproducibility in serial examinations’, **246**(1), 71–80.

Kim, S. J., Moon, W. K., Seong, M. H., Cho, N. and Chang, J. M. (2009), ‘Computer-aided detection in digital mammography: false-positive marks and their reproducibility in negative mammograms’, **50**(9), 999–1004.

Kim, Y. and Mansfield, L. (2014), ‘Fool Me Twice: Delayed Diagnoses in Radiology With Emphasis on Perpetuated Errors’, *American Journal of Roentgenology* **202**(3), 465–470.

Kingstone, A. and Klein, R. M. (1993), 'What are human express saccades?', **54**(2), 260–273.

Klein, C. and Ettinger, U., eds (2019), *Eye Movement Research*, Springer International Publishing.

Kneusel, R. and Mozer, M. (2017), 'Improving Human-Machine Cooperative Visual Search With Soft Highlighting', *ACM Transactions on Applied Perception* **15**(1), 1–21.

Ko, J., Nicholas, M., Mendel, J. and Slanetz, P. (2006), 'Prospective Assessment of Computer-Aided Detection in Interpretation of Screening Mammography', *American Journal of Roentgenology* **187**(6), 1483–1491.

Kohli, A. and Jha, S. (2018), 'Why CAD failed in mammography', *Journal of the American College of Radiology* **15**(3), 535–537.

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C., Mann, R., den Heeten, A. and Karssemeijer, N. (2017), 'Large scale deep learning for computer aided detection of mammographic lesions', *Medical Image Analysis* **35**, 303–312.

Krupinski, E. (2016), 'Diagnostic Accuracy and Visual Search Efficiency: Single 8 MP vs. Dual 5 MP Displays', *Journal of Digital Imaging* **30**(2), 144–147.

Krupinski, E. A. (1996), 'Visual scanning patterns of radiologists searching mammograms', *Academic Radiology* **3**(2), 137–144.

Krupinski, E. A. (2010), 'Current perspectives in medical image perception', *Attention, Perception, & Psychophysics* **72**(5), 1205–1217.

Kunar, M., Watson, D., Taylor-Phillips, S. and 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.

Kundel, H. L. (2004), 'Reader error, object recognition, and visual search', *Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment* .

Kundel, H. L., Nodine, C. F., Krupinski, E. A. and Mello-Thoms, C. (2008), 'Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms', *Academic Radiology* **15**(7), 881–886.

Kundel, H. and Nodine, C. (2004), 'Modeling visual search during mammogram viewing', *Journal of Experimental Psychology: Human Perception and Performance, Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment* .

Kundel, H., Nodine, C., Conant, E. and Weinstein, S. (2007), 'Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study', *Radiology* **242**(2), 396–402.

Kundel, H. and Nodine, C. F. (2010), A short history of image perception in medical radiology, *in* E. K. E Samei, ed., 'The Handbook of Medical Image Perception and Techniques', Cambridge University Press, pp. 9–20.

Lång, K., Hofvind, S., Rodríguez-Ruiz, A. and Andersson, I. (2021), 'Can artificial intelligence reduce the interval cancer rate in mammography screening?', *European Radiology* **31**(8), 5940–5947.

LeCun, Y., Bengio, Y. and Hinton, G. (2015), 'Deep learning', *Nature* **521**(7553), 436–444.

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

Lesniak, J. M., Hupse, R., Blanc, R., Karssemeijer, N. and Székely, G. (2012), 'Comparative evaluation of support vector machine classification for computer aided detection of breast masses in mammography', *Physics in Medicine and Biology* **57**(16), 5295–5307.

Lévêque, L., Berg, B. V., Bosmans, H., Cockmartin, L., Keupers, M., Ongeval, C. V. and Liu, H. (2019), 'A statistical evaluation of eye-tracking data of screening mammography: Effects of expertise and experience on image reading', *Signal Processing: Image Communication* **78**, 86–93.

Leveque, L., Young, P. and Liu, H. (2021), Studying the gaze patterns of expert radiologists in screening mammography: A case study with breast test wales, *in* '2020 28th European Signal Processing Conference (EUSIPCO)', IEEE.

Li, H., Giger, M. L., Yuan, Y., Lan, L., Suzuki, K., Jamieson, A., Yarusso, L., Nishikawa, R. M. and Sennett, C. (2006), Comparison of computerized image analyses for digitized screen-film mammograms and full-field digital mammography images, *in* 'Digital Mammography', Springer Berlin Heidelberg, pp. 569–575.

Li, Y., Chen, H., Yang, Y., Cheng, L. and Cao, L. (2015), 'A bilateral analysis scheme for false positive reduction in mammogram mass detection', *Computers in Biology and Medicine* **57**, 84–95.

Liberman, L., Morris, E. A., Lee, M. J.-Y., Kaplan, J. B., LaTrenta, L. R., Menell, J. H., Abramson, A. F., Dashnaw, S. M., Ballon, D. J. and Dershaw, D. D. (2002), 'Breast Lesions Detected on MR Imaging: Features and Positive Predictive Value', *American Journal of Roentgenology* **179**(1), 171–178.

Littlefair, S., Brennan, P., Mello-Thoms, C., Dung, P., Pietryzk, M., Talanow, R. and Reed, W. (2016), 'Outcomes knowledge may bias radiological decision-making', *Academic Radiology* **23**(6), 760–767.

Loy, C. T. and Irwig, L. (2004), 'Accuracy of diagnostic tests read with and without clinical information', *JAMA* **292**(13), 1602.

Lynch, P. J. (2006), 'Breast anatomy normal scheme'. Available at: `https://commons.wikimedia.org/wiki/File:Breast_anatomy_normal_scheme.png` [Accessed 3 Mar. 2021].

Ma, H., Bandos, A., Rockette, H. and Gur, D. (2013), 'On use of partial area under the ROC curve for evaluation of diagnostic performance', *Statistics in Medicine* **32**(20), 449–3458.

Mahoney, M. C., Gatsonis, C., Hanna, L., DeMartini, W. B. and Lehman, C. (2012), 'Positive predictive value of BI-RADS MR imaging', *Radiology* **264**(1), 51–58.

Majid, A., de Paredes, E., Doherty, R., Sharma, N. and Salvador, X. (2003), 'Missed Breast Carcinoma: Pitfalls and Pearls', *RadioGraphics* **23**(4), 881–895.

Malich, A., Sauner, D., Marx, C., Facius, M., Boehm, T., Pfleiderer, S. O., Fleck, M. and Kaiser, W. A. (2003), 'Influence of breast lesion size and histologic findings on tumor detection rate of a computer-aided detection system', *Radiology* **228**(3), 851–856.

Mall, S., Brennan, P. C. and Mello-Thoms, C. (2018), ‘Modeling visual search behavior of breast radiologists using a deep convolution neural network’, *Journal of Medical Imaging* **5**(03), 1.

Mall, S., Brennan, P. C. and Mello-Thoms, C. (2019a), ‘Can a machine learn from radiologists’ visual search behaviour and their interpretation of mammograms—a deep-learning study’, *Journal of Digital Imaging* **32**(5), 746–760.

Mall, S., Krupinski, E. A. and Mello-Thoms, C. R. (2019b), Missed cancer and visual search of mammograms: what feature based machine-learning can tell us that deep-convolution learning cannot, *in* R. M. Nishikawa and F. W. Samuelson, eds, ‘Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment’, SPIE.

Marmot, M., Altman, D., Cameron, D., Dewar, J., Thompson, S. and Wilcox, M. (2013), ‘The benefits and harms of breast cancer screening: an independent review’, *British Journal of Cancer* **108**(11), 2205–2240.

Massat, N., Dibden, A., Parmar, D., Cuzick, J., Sasieni, P. and Duffy, S. (2015), ‘Impact of screening on breast cancer mortality: The UK program 20 years on’, *Cancer Epidemiology Biomarkers & Prevention* **25**(3), 455–462.

Mayo, R., Kent, D., Sen, L., Kapoor, M., Leung, J. and Watanabe, A. (2019), ‘Reduction of false-positive markings on mammograms: a retrospective comparison study using an artificial intelligence-based CAD’, *Journal of Digital Imaging* **32**(4), 618–624.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., Fauw, J. D. and Shetty, S. (2020), 'International evaluation of an AI system for breast cancer screening', *Nature* **577**(7788), 89–94.

McLoughlin, K., Bones, P. and Karssemeijer, N. (2004), 'Noise equalization for detection of microcalcification clusters in direct digital mammogram images', *IEEE Transactions on Medical Imaging* **23**(3), 313–320.

MedCalc (2021), 'ROC curve analysis'. Available at: `https://www.medcalc.org/manual/roc-curves.php` [Accessed 22 Mar. 2021].

Mello-Thoms, C. (2003), 'Perception of breast cancer', *Academic Radiology* **10**(1), 4–12.

Mello-Thoms, C. (2008), 'How much agreement is there in the visual search strategy of experts reading mammograms?', *Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment* .

Mello-Thoms, C. (2009), The holistic grail: possible implications of an initial mistake in the reading of digital mammograms, *in* B. Sahiner and D. J. Manning, eds, 'Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment', SPIE.

Mello-Thoms, C. (2010), 'Visual search characteristics in mammogram reading: SFM vs. FFDM', *Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment* .

Mello-Thoms, C. and Gur, D. (2007), 'Remote vs. head-mounted eye-tracking: a comparison using radiologists reading mammograms', *Medical Imaging 2007: Image Perception, Observer Performance, and Technology Assessment* .

Mello-Thoms, C., Trieu, P. and Brennan, P. (2014), Going on with false beliefs: What if satisfaction of search was really suppression of recognition?, *in* C. R. Mello-Thoms and M. A. Kupinski, eds, 'Medical Imaging 2014: Image Perception, Observer Performance, and Technology Assessment', SPIE.

Menneer, T., Barrett, D. J. K., Phillips, L., Donnelly, N. and Cave, K. R. (2007), 'Costs in searching for two targets: dividing search across target types could improve airport security screening', *Applied Cognitive Psychology* **21**(7), 915–932.

Merchant, J., Morrissette, R. and Porterfield, J. L. (1974), 'Remote measurement of eye direction allowing subject motion over one cubic foot of space', *IEEE Transactions on Biomedical Engineering* **BME-21**(4), 309–317.

Mestry, N., Menneer, T., Cave, K. R., Godwin, H. J. and Donnelly, N. (2017), 'Dual-target cost in visual search for multiple unfamiliar faces.', *Journal of Experimental Psychology: Human Perception and Performance* **43**(8), 1504–1519.

Methven, T. and Qi, L. (2018), 'Texturelab edinburgh – resources – scripts'. Available at: `http://www.macs.hw.ac.uk/texturelab/resources/scripts/` [Accessed 25 Jan. 2018].

Morra, L., Sacchetto, D., Durando, M., Agliozzo, S., Carbonaro, L. A., Delsanto, S., Pesce, B., Persano, D., Mariscotti, G., Marra, V., Fonio, P. and Bert, A. (2015), 'Breast cancer: Computer-aided detection with digital breast tomosynthesis', **277**(1), 56–63.

Morton, M., Whaley, D., Brandt, K. and Amrami, K. (2006), 'Screening mammograms: interpretation with computer-aided detection—prospective evaluation', *Radiology* **239**(2), 375–383.

Mousa, D. S. A., Brennan, P. C., Ryan, E. A., Lee, W. B., Tan, J. and Mello-Thoms, C. (2014), ‘How mammographic breast density affects radiologists' visual search patterns’, *Academic Radiology* **21**(11), 1386–1393.

Mowrer, O. H., Ruch, T. C. and Miller, N. E. (1935), ‘THE CORNEO-RETINAL POTENTIAL DIFFERENCE AS THE BASIS OF THE GALVANOMETRIC METHOD OF RECORDING EYE MOVEMENTS’, *American Journal of Physiology-Legacy Content* **114**(2), 423–428.

Muir, B. M. (1987), ‘Trust between humans and machines, and the design of decision aids’, *International Journal of Man-Machine Studies* **27**(5-6), 527–539.

Murakami, R., Kumita, S., Tani, H., Yoshida, T., Sugizaki, K., Kuwako, T., Kiriyama, T., Hakozaki, K., Okazaki, E., Yanagihara, K., Iida, S., Haga, S. and Tsuchiya, S. (2013), ‘Detection of breast cancer with a computer-aided detection applied to full-field digital mammography’, *Journal of Digital Imaging* **26**(4), 768–773.

Muralidhar, G. S., Bovik, A. C., Giese, J. D., Sampat, M. P., Whitman, G. J., Haygood, T. M., Stephens, T. W. and Markey, M. K. (2010), ‘Snakules: A model-based active contour algorithm for the annotation of spicules on mammography’, *IEEE Transactions on Medical Imaging* **29**(10), 1768–1780.

Müller, H. J., Humphreys, G. W. and Donnelly, N. (1994), ‘SEarch via recursive rejection (SERR): Visual search for single and dual form-conjunction targets.’, **20**(2), 235–258.

Narod, S. A., Iqbal, J. and Miller, A. B. (2015), ‘Why have breast cancer mortality rates declined?’, *Journal of Cancer Policy* **5**(1), 8–17.

Nartker, M. S., Alaoui-Soce, A. and Wolfe, J. M. (2020), ‘Visual search errors are persistent in a laboratory analog of the incidental finding problem’, **5**(1).

National Cancer Institute (2012*a*), 'Breast Cancer Screening Programs in 26 ICSN Countries, 2012: Organization, Policies, and Program Reach'. Available at: `https://healthcaredelivery.cancer.gov/icsn/breast/screening.html` [Accessed 13 Dec. 2017].

National Cancer Institute (2012*b*), 'Policies on Number of Views, Double-Reading, and Computer Aided Detection for Breast Cancer Screening Programs in 26 ICSN Countries, 2012'. Available at: `https://healthcaredelivery.cancer.gov/icsn/breast/policies.reading.html` [Accessed 20 Dec. 2017].

National Eye Institute (2019), 'Nei-medialibrary-7079116.png'. Available at: `https://medialibrary.nei.nih.gov/media/1821` [Accessed 02 Sep. 2021].

NICE (2017), 'Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer — Guidance and guidelines — NICE'. Available at: `https://www.nice.org.uk/guidance/cg164/chapter/Recommendations` [Accessed 14 Jul. 2018].

Nikulin, Y., Fillard, P., Clatz, O. and Iannessi, A. (2017), 'DM Challenge Yaroslav Nikulin (Therapixel) Submission'. Available at: `https://www.synapse.org/#!Synapse:syn9773040/wiki/426908` [Accessed 18 Dec. 2017].

Nishikawa, R. (2007), 'Current status and future directions of computer-aided diagnosis in mammography', *Computerized Medical Imaging and Graphics* **31**(4-5), 224–235.

Nishikawa, R. M., Giger, M. L., Doi, K., Vyborny, C. J. and Schmidt, R. A. (1995), 'Computer-aided detection of clustered microcalcifications on digital mammograms', *Medical & Biological Engineering & Computing* **33**(2), 174–178.

Nishikawa, R. M. and Kyongtae, 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), 49–52.

Nishikawa, R. and Pesce, L. (2009), 'Computer-aided Detection Evaluation Methods Are Not Created Equal', *Radiology* **251**(3), 634–636.

Nishikawa, R., Schmidt, R., Linver, M., Edwards, A., Papaioannou, J. and Stull, M. (2012), 'Clinically Missed Cancer: How Effectively Can Radiologists Use Computer-Aided Detection?', *American Journal of Roentgenology* **198**(3), 708–716.

Nodine, C. F., Kundel, H. L., Mello-Thoms, C., Weinstein, S. P., Orel, S. G., Sullivan, D. C. and Conant, E. F. (1999), 'How experience and training influence mammography expertise', *Academic Radiology* **6**(10), 575–585.

Nodine, C. and Kundel, H. (1987), THE COGNITIVE SIDE OF VISUAL SEARCH IN RADIOLOGY, *in* 'Eye Movements from Physiology to Cognition', Elsevier, pp. 573–582.

Nodine, C., Kundel, H., Mello-Thoms, C. and Weinstein, S. (2001), 'Role of computer-assisted visual search in mammographic interpretation', *Medical Imaging 2001: Image Perception and Performance* .

Oberauer, K. (2002), 'Access to information in working memory: Exploring the focus of attention.', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **28**(3), 411–421.

Office for National Statistics (2019), 'Cancer Registration Statistics, England: 2017'.

Office for National Statistics (2020), 'Monthly diagnostics: Waiting times and activity for diagnostic tests and procedures (december 2019)'. Available at: `https://www.`

`england.nhs.uk/statistics/statistical-work-areas/diagnostics-waiting-times-and-activity/monthly-diagnostics-waiting-times-and-activity/monthly-diagnostics-data-2019-20/` [Accessed 21 Apr. 2021].

Ooms, K., Coltekin, A., Maeyer, P. D., Dupont, L., Fabrikant, S., Incoul, A., Kuhn, M., Slabbinck, H., Vansteenkiste, P. and der Haegen, L. V. (2014), ‘Combining user logging with eye tracking for interactive and dynamic applications’, *Behavior Research Methods* **47**(4), 977–993.

Palmer, J. (1995), ‘Attention in visual search: Distinguishing four causes of a set-size effect’, **4**(4), 118–123.

Papenmeier, F. and Huff, M. (2010), ‘DynAOI: A tool for matching eye-movement data with dynamic areas of interest in animations and movies’, *Behavior Research Methods* **42**(1), 179–187.

Parasuraman, R. and Riley, V. (1997), ‘Humans and Automation: Use, Misuse, Disuse, Abuse’, *Human Factors: The Journal of the Human Factors and Ergonomics Society* **39**(2), 230–253.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E. and Lindeløv, J. K. (2019), ‘PsychoPy2: Experiments in behavior made easy’, *Behavior Research Methods* **51**(1), 195–203.

Phelps, A., Callen, A., Marcovici, P., Naeger, D., Mongan, J. and Webb, E. (2017), ‘Can Radiologists Learn From Airport Baggage Screening?’, *Academic Radiology* .

Philpotts, L. (2009), ‘Can Computer-aided Detection Be Detrimental to Mammographic Interpretation?’, *Radiology* **253**(1), 17–22.

Pupil Labs (2021*a*), ‘Technical specs & performance’. Available at: `https://do` `cs.pupil-labs.com/core/software/pupil-capture/#pupil-detection` [Accessed 6 Aug. 2021].

Pupil Labs (2021*b*), ‘User guide – Pupil Capture – Pupil Detection’. Available at: `https://docs.pupil-labs.com/core/software/pupil-capture/#pupil-detection` [Accessed 21 Jul. 2021].

Qian, W., Mao, F., Sun, X., Zhang, Y., Song, D. and Clarke, R. A. (2002), ‘An improved method of region grouping for microcalcification detection in digital mammograms’, *Computerized Medical Imaging and Graphics* **26**(6), 361–368.

Ragab, D. A., Sharkas, M., Marshall, S. and Ren, J. (2019), ‘Breast cancer detection using deep convolutional neural networks and support vector machines’, *PeerJ* **7**, e6201.

Ramadan, S. Z. (2020), ‘Methods used in computer-aided diagnosis for breast cancer detection using mammograms: A review’, *Journal of Healthcare Engineering* **2020**, 1–21.

Rangayyan, R. M., Banik, S. and Desautels, J. E. L. (2010), ‘Computer-aided detection of architectural distortion in prior mammograms of interval cancer’, *Journal of Digital Imaging* **23**(5), 611–631.

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

Reed, W. M., Chow, S. L. C., Chew, L. E. and Brennan, P. C. (2014), ‘Can prevalence expectations drive radiologists' behavior?’, *Academic Radiology* **21**(4), 450–456.

Ribli, D., Horváth, A., Unger, Z., Pollner, P. and Csabai, I. (2018), 'Detecting and classifying lesions in mammograms with deep learning', *Scientific Reports* **8**(1).

Rizzi, M., Matteo, D. and Castagnolo, B. (2012), 'Review: Health Care CAD Systems for Breast Microcalcification Cluster Detection', *Journal of Medical and Biological Engineering* **32**(3), 147.

Robinson, D. A. (1963), 'A method of measuring eye movemnent using a scieral search coil in a magnetic field', *IEEE Transactions on Bio-medical Electronics* **10**(4), 137–145.

Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.-J., Schilling, K., Heywang-Köbrunner, S. H., Sechopoulos, I. and Mann, R. M. (2019a), 'Detection of breast cancer with mammography: Effect of an artificial intelligence support system', *Radiology* **290**(2), 305–314.

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., Helbich, T. H., Chevalier, M., Tan, T., Mertelmeier, T., Wallis, M. G., Andersson, I., Zackrisson, S., Mann, R. M. and Sechopoulos, I. (2019b), 'Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists', *JNCI: Journal of the National Cancer Institute* **111**(9), 916–922.

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Teuwen, J., Broeders, M., Gennaro, G., Clauser, P., Helbich, T. H., Chevalier, M., Mertelmeier, T., Wallis, M. G., Andersson, I., Zackrisson, S., Sechopoulos, I. and Mann, R. M. (2019c), 'Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? a feasibility study', *European Radiology* **29**(9), 4825–4832.

Rojas-Domínguez, A. and Nandi, A. K. (2008), ‘Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection’, *Computerized Medical Imaging and Graphics* **32**(4), 304–315.

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

Sadaf, A., Crystal, P., Scaranelo, A. and Helbich, T. (2009), ‘Performance of computer-aided detection applied to full-field digital mammography in detection of breast cancers’, *European Journal of Radiology* **77**(3), 457–461.

Sakellaropoulos, F., Skiadopoulos, S., Karahaliou, A., Costaridou, L. and Panayiotakis, G. (2006), Using wavelet-based features to identify masses in dense breast parenchyma, *in* ‘Digital Mammography’, Springer Berlin Heidelberg, pp. 557–564.

Sampat, M. P., Bovik, A. C., Whitman, G. J. and Markey, M. K. (2008), ‘A model-based framework for the detection of spiculated masses on mammographya)’, *Medical Physics* **35**(5), 2110–2123.

Sampat, M. P., Markey, M. K. and Bovik, A. C. (2005), Computer-aided detection and diagnosis in mammography, *in* ‘Handbook of Image and Video Processing’, Elsevier, pp. 1195–1217.

Samulski, M., Hupse, R., Boetes, C., Mus, R., den Heeten, G. and Karssemeijer, N. (2010), ‘Using computer-aided detection in mammography as a decision support’, *European Radiology* **20**(10), 2323–2330.

Saunders, R. S. and Samei, E. (2006), ‘Improving mammographic decision accuracy by incorporating observer ratings with interpretation time’, *The British Journal of Radiology* **79**(special_issue_2), S117–S122.

Scaranelo, A. M., Eiada, R., Bukhanov, K. and Crystal, P. (2012), 'Evaluation of breast amorphous calcifications by a computer-aided detection system in full-field digital mammography', *The British Journal of Radiology* **85**(1013), 517–522.

Scott, H. J. and Gale, A. G. (2006), 'Breast screening: PERFORMS identifies key mammographic training needs', *The British Journal of Radiology* **79**(special_issue_2), S127–S133.

ScreenPoint (2021), 'Transpara'. Available at: `https://www.screenpoint-medical.com/transpara` [Accessed 27 Mar. 2021].

Sickles, E., Wolverton, D. and Dee, K. (2002), 'Performance parameters for screening and diagnostic mammography: specialist and general radiologists', *Radiology* **224**(3), 861–869.

Skaane, P., Kshirsagar, A., Stapleton, S., Young, K. and Castellino, R. (2007), 'Effect of Computer-Aided Detection on Independent Double Reading of Paired Screen-Film and Full-Field Digital Screening Mammograms', *American Journal of Roentgenology* **188**(2), 377–384.

SMI (2017*a*), *BeGaze Manual Vesion 3.7*. Available at: `http://www.humre.vu.lt/files/doc/Instrukcijos/SMI/BeGaze2.pdf` [Accessed 06 Sep. 2021].

SMI (2017*b*), 'SMI Eye Tracking Glasses 2 Wireless'. Available at: `https://cpb-eu-w2.wpmucdn.com/blogs.brighton.ac.uk/dist/9/1193/files/2018/04/smi_prod_ETG_120Hz_asgm-1wxag5y.pdf` [Accessed 20 Jun. 2022].

SR Research (2021), 'Eyelink ii: Head-mounted video-based eye tracker'. Available at: `https://www.sr-research.com/wp-content/uploads/2021/03/EyeLink-II-Specs.pdf` [Accessed 20 Jun. 2022].
URL: *https://www.sr-research.com/eyelink-ii/*

SR Research Ltd (2009), *EyeLink® 1000 User Manual, Version 1.5.0*, Mississauga, Ontario, Canada.
URL: *http://www.sr-research.com/*

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F. (2021), 'Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', **71**(3), 209–249.

Swensson, R. G. and Theodore, G. H. (1990), 'Search and nonsearch protocols for radiographic consultation.', *Radiology* **177**(3), 851–856.

Swirski, L. and Dodgson, N. (2013), A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting, *in* 'Proceedings of ECEM 2013'.

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

Taylor, P. and Potts, H. (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.

Tchou, P. M., Haygood, T. M., Atkinson, E. N., Stephens, T. W., Davis, P. L., Arribas, E. M., Geiser, W. R. and Whitman, G. J. (2010), 'Interpretation time of computer-aided detection at screening mammography', *Radiology* **257**(1), 40–46.

The, J., Schilling, K., Hoffmeister, J., Friedmann, E., McGinnis, R. and Holcomb, R. (2009), 'Detection of breast cancer with full-field digital mammography and computer-aided detection', *American Journal of Roentgenology* **192**(2), 337–340.

The Royal College of Radiologists (2013), 'Guidance on screening and symptomatic breast imaging, Third edition'.

The Royal College of Radiologists (2020), ‘Clinical radiology UK workforce census 2019 report’. London: The Royal College of Radiologists, Ref No. BFCR(20)2.

Tiew, S., Astley, S., Dillon, B., Morris, J. and Boggis, C. (2008), Prompting in mammography: Reproducibility, *in* ‘Digital Mammography, 9th International Workshop, IWDM 2008’, Springer Berlin Heidelberg, pp. 137–142.

Tobii Pro (2017), ‘Dark and bright pupil tracking’. Available at: `https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/what-is-dark-and-bright-pupil-tracking/` [Accessed 02 Jan. 2018].

Tobii Pro (2018), ‘Product Description Tobii Pro Glasses 2’. Available at: `https://www.tobiipro.com/siteassets/tobii-pro/product-descriptions/tobii-pro-glasses-2-product-description.pdf` [Accessed 20 Jun. 2022].

Tobii Pro (n.d.), ‘Working with dynamic stimuli in the AOI tool’. Available at: `https://www.tobiipro.com/learn-and-support/learn/steps-in-an-eye-tracking-study/data/analyzing-dynamic-stimuli-with-the-aoi-tool/` [Accessed 21 Jul. 2021].

Tourassi, G. D., Delong, D. M. and Floyd, C. E. (2006), ‘A study on the computerized fractal analysis of architectural distortion in screening mammograms’, *Physics in Medicine and Biology* **51**(5), 1299–1312.

Tourassi, G. D., Mazurowski, M. A. and Krupinski, E. A. (2010), ‘Perception-driven IT-CADe analysis for the detection of masses in screening mammography: initial investigation’, *Proc. SPIE 7624, Medical Imaging 2010: Computer-Aided Diagnosis* **762406**.

Treisman, A. M. and Gelade, G. (1980), ‘A feature-integration theory of attention’, *Cognitive Psychology* **12**, 97–136.

van den Biggelaar, F. J. H. M., Kessels, A. G. H., van Engelshoven, J. M. A., Boetes, C. and Flobbe, K. (2009), 'Computer-aided detection in full-field digital mammography in a clinical population: performance of radiologist and technologists', *Breast Cancer Research and Treatment* **120**(2), 499–506.

van der Gijp, A., Ravesloot, C., Jarodzka, H., van der Schaaf, M., van der Schaaf, I., van Schaik, J. and ten Cate, T. (2016), 'How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology', *Advances in Health Sciences Education* **22**(3), 765–787.

van Zelst, J. C. M., Tan, T., Clauser, P., Domingo, A., Dorrius, M. D., Drieling, D., Golatta, M., Gras, F., de Jong, M., Pijnappel, R., Rutten, M. J. C. M., Karssemeijer, N. and Mann, R. M. (2018), 'Dedicated computer-aided detection software for automated 3d breast ultrasound; an efficient tool for the radiologist in supplemental screening of women with dense breasts', **28**(7), 2996–3006.

Varela, C., Tahoces, P. G., Méndez, A. J., Souto, M. and Vidal, J. J. (2007), 'Computerized detection of breast masses in digitized mammograms', *Computers in Biology and Medicine* **37**(2), 214–226.

Voisin, S., Pinto, F., Morin-Ducote, G., Hudson, K. B. and Tourassi, G. D. (2013a), 'Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography', *Medical Physics* **40**(10), 101906.

Voisin, S., Pinto, F., Xu, S., Morin-Ducote, G., Hudson, K. and Tourassi, G. D. (2013b), Investigating the association of eye gaze pattern and diagnostic error in mammography, *in* C. K. Abbey and C. R. Mello-Thoms, eds, 'Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment', SPIE.

Wade, N. J. (2010), 'Pioneers of eye movement research', *i-Perception* **1**(2), 33–68.

Warren, L., Mackenzie, A., Cooke, J., Given-Wilson, R., Wallis, M., Chakraborty, D., Dance, D., Bosmans, H. and Young, K. (2012), 'Effect of image quality on calcification detection in digital mammography', *Medical Physics* **39**(6), 3202–3213.

Watanabe, A. T., Lim, V., Vu, H. X., Chim, R., Weise, E., Liu, J., Bradley, W. G. and Comstock, C. E. (2019), 'Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography', *Journal of Digital Imaging* **32**(4), 625–637.

WHO (2020), 'Cancer incidence and mortality statistics worldwide and by region'. Available at: `https://gco.iarc.fr/today/fact-sheets-cancers` [Accessed 30 Mar. 2021].

Wilczek, B., Wilczek, H., Rasouliyan, L. and Leifand, K. (2016), 'Adding 3d automated breast ultrasound to mammography screening in women with heterogeneously and extremely dense breasts: report from a hospital-based, high-volume, single-center breast cancer screening program', *European Radiology* **85**, 1554–1563.

Williams, L., Carrigan, A., Auffermann, W., Mills, M., Rich, A., Elmore, J. and Drew, T. (2020), 'The invisible breast cancer: Experience does not protect against inattentional blindness to clinically relevant findings in radiology', *Psychonomic Bulletin & Review* **28**(2), 503–511.

Winsberg, F., Elkin, M., Macy, J., Bordaz, V. and Weymouth, W. (1967), 'Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis', *Radiology* **89**, 211—-215.

Wolfe, J. (1998), Visual search, *in* H. Pashler, ed., 'Attention', London UK: University College London Press, pp. 13–73.

Wolfe, J., Cain, M., Ehinger, K. and Drew, T. (2015b), 'Guided Search 5.0: Meeting the challenge of hybrid search and multiple-target foraging', *Journal of Vision* **15**(12), 1106.

Wolfe, J. M. (1994), 'Guided search 2.0 a revised model of visual search', **1**(2), 202–238.

Wolfe, J. M. (2007), Guided Search 4.0: Current progress with a model of visual search, *in* W. Gray, ed., 'Series on cognitive models and architectures. Integrated models of cognitive systems', Oxford Scholarship Online, pp. 99–119.

Wolfe, J. M. (2012), 'Saved by a log: How do humans perform hybrid visual and memory search?', **23**(7), 698–703.

Wolfe, J. M. (2021a), 'Guided search 6.0: An updated model of visual search', *Psychonomic Bulletin & Review* .

Wolfe, J. M., Aizenman, A. M., Boettcher, S. E. and Cain, M. S. (2016), 'Hybrid foraging search: Searching for multiple instances of multiple types of target', **119**, 50–59.

Wolfe, J. M., Cain, M. S. and Alaoui-Soce, A. (2017b), 'Hybrid value foraging: How the value of targets shapes human foraging behavior', **80**(3), 609–621.

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

Wolfe, J. M., Evans, K. K., Drew, T., Aizenman, A. and Josephs, E. (2015a), 'How Do Radiologists Use The Human Search Engine?', **169**(1-4), 24–31.

Wolfe, J. M., Horowitz, T. S. and Kenner, N. M. (2005), 'Rare items often missed in visual searches', **435**(7041), 439–440.

Wolfe, J. M., Soce, A. A. and Schill, H. M. (2017a), ‘How did i miss that? developing mixed hybrid visual search as a ‘model system’ for incidental finding errors in radiology’, **2**(1).

Wolfe, J. M., Wu, C.-C., Li, J. and Suresh, S. B. (2021b), ‘What do experts look at and what do experts find when reading mammograms?’, *Journal of Medical Imaging* **8**(04).

Yala, A., Schuster, T., Miles, R., Barzilay, R. and Lehman, C. (2019), ‘A deep learning model to triage screening mammograms: A simulation study’, *Radiology* **293**(1), 38–46.

Yang, S. K., Moon, W. K., Cho, N., Park, J. S., Cha, J. H., Kim, S. M., Kim, S. J. and Im, J.-G. (2007), ‘Screening mammography–detected cancers: Sensitivity of a computer-aided detection system applied to full-field digital mammograms’, *Radiology* **244**(1), 104–111.

Yarusso, L. M., Nishikawa, R. M., Papaioannou, J., Nagel, R., Jokich, P. and A., V. L. (2000), Application of computer-aided diagnosis to full-field digital mammography, *in* Y. M. J., ed., ‘Digital Mammography 2000’, Medical Physics Publishing, Madison, WI, p. 421–426.

Yassin, N. I., Omran, S., Houby, E. M. E. and Allam, H. (2018), ‘Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review’, *Computer Methods and Programs in Biomedicine* **156**, 25–45.

Yoshikawa, R., Teramoto, A., Matsubara, T. and Fujita, H. (2014), Automated detection of architectural distortion using improved adaptive gabor filter, *in* ‘Breast Imaging’, Springer International Publishing, pp. 606–611.

Young, A. H. and Hulleman, J. (2013), ‘Eye movements reveal how task difficulty moulds visual search.’, *Journal of Experimental Psychology: Human Perception and Performance* **39**(1), 168–190.

Zhang, X., Homma, N., Goto, S., Kawasumi, Y., Ishibashi, T., Abe, M., Sugita, N. and Yoshizawa, M. (2013), ‘A hybrid image filtering method for computer-aided detection of microcalcification clusters in mammograms’, *Journal of Medical Engineering* **2013**, 1–8.

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