Evaluating Artificial Intelligence in Breast Cancer Screening

Dr Sarah Elizabeth Hickman

Department of Radiology
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare College June 2022

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the Degree Committee for Clinical Medicine and Veterinary Medicine.

Sarah Elizabeth Hickman

June 2022

Abstract

Evaluating Artificial Intelligence in Breast Cancer Screening

Dr Sarah Elizabeth Hickman

This thesis evaluates the application and performance of artificial intelligence (AI) in breast cancer screening.

Breast cancer screening is conducted on a population scale using mammographic imaging for the earlier detection of breast cancer and has been shown to reduce mortality. A shortage of trained radiologists, as well as the demands of double reading, mean an approach to alleviate pressures within the breast screening workflow is sought. In addition, interval cancers occur at an estimated rate of 3.7/1000 women screened in the UK, thus methods to improve the sensitivity of screening and detect cancers earlier are also needed. Advances in AI over the past decade have demonstrated comparable performance to human readers and could provide a method for an adapted screening workflow to improve both efficiency and efficacy of screening. However, the 2021 National Screening Committee (NSC) report concluded that there was insufficient evidence to support the adoption of AI into the UK breast screening programme.

This thesis aims to fill the gaps in evidence highlighted in the NSC report for the performance of AI algorithms within a UK breast cancer screening population, as well as explore the various potential workflow deployment approaches of AI in the screening programme.

I start by conducting a systematic review and meta-analysis of the current literature investigating the performance of stand-alone AI applications in breast cancer screening for detection and diagnosis as well as triage approaches. I then describe the creation of a large scale independent medical imaging database which is used in the studies throughout this thesis. The remainder of the thesis describes the results of three retrospective studies evaluating three different commercial AI algorithms. The first study assesses the ability of AI to detect interval cancers at the previous screen. The second study investigates the performance of AI as a stand-alone screen reader. The third study evaluates the proportion of cases identified for both high sensitivity rule out and high specificity rule in triage, as well as the proportion of cancers missed at these thresholds.

Overall the results of this thesis will inform discussions around the use of AI in the UK breast screening programme as well as the design of future prospective trials.

Acknowledgements

Firstly, a special thank you to my PhD supervisor Professor Fiona Gilbert. The support, encouragement and guidance you have provided over the course of my PhD has been invaluable. It has been a pleasure to be supervised by such an eminent researcher in the field and your expertise has been pivotal to the success of this project. These past three years have provided me with knowledge and skills as well as increased my confidence that will be useful for the rest of my life.

To Richard Black and Dr Nicholas Payne, I continue to be in awe of your endless knowledge of computer related systems. Your help has been immense and I am grateful for everything you have taught me as well as all your contributions to this research.

To Dr Yuan Huang, I am so grateful for all your statistical help in various projects over the course of my PhD, as well as your patience when having to repeatedly explain statistical tests to me.

To Dr Martin Graves, Dr Andrew Priest, Dr Josh Kaggie and Bahman Kasmai, thank you for your unwavering support and guidance.

I would like to thank Dr Ramona Woitek, Dr Iris Allajbeu, Dr Muzna Nanaa, Liana Hough, Dr Angelica I Aviles-Rivero, Dr Lorena Escudero Sánchez, and Sue Hudson for their contributions to various projects over the course of my PhD.

Thank you to the whole BRAID trial team. Especially Jaimie Taylor, thank you for always being keen to discuss fantasy football or an NFL game.

To all the members of the Cambridge Cohort Database Access Committee: Esme Radin, Dr Arne Juette, Kathryn Taylor, Adam Loveday, thank you for your advice and insights. A special thank you to Carolyn Read and Helen Street, I remember clearly sitting in both your offices three years ago trying to work out how to get the database approved, your guidance, proof reading of documents and general support through this process made this possible.

I would like to thank both the Cambridge and Norwich Breast Unit teams, particularly Lisa Tatham Heather Couzens and Mandy Ballantyne who have helped run numerous NBSS queries and KC62 reports for me.

To Dr Mary Kasanicki and Dr Mona Alexander thank you for persevering with our collaboration contracts throughout this work and teaching me what contract is.

To Dr Gaby Baxter and Dr Dimitri Kessler, thank you for always being there for snacks and a chat or to pull me back from that puzzled look on my face and bounce an idea off. Sarah Perkins, thank you for helping me organise everything and keeping me on track.

Thank you to the NIHR Cambridge Biomedical Research Centre (BRC) PPI team, especially Dr Amanda Stranks, as well as the patients and public who have participated in our events.

To the companies who collaborated on this work, thank you for generosity in making your algorithms available and the giving of your time.

I would like to thank CRUK and the NIHR Cambridge BRC for funding my PhD.

Thank you to all the women whose de-identified data was used in this research. I hope this work will lead to improvements in breast cancer screening overtime. It has been a privilege to work in this field and witness the incredible work the NHS Breast Screening Programme carries out.

To the Cambridge University Association Football Club second team (aka the Eagles), thank you for giving me the opportunity to play in such a wonderful and welcoming team, and three out of three varsity wins against Oxford is not bad!

To the absolute legends that are my closest friends, thank you for keeping me smiling the whole way. To my kind and wonderful sister Katy, thank you for supporting me, listening to ramblings and always knowing how to make me feel better. To my brother Rob, thank you for providing me with the two most adventurous cats to keep me entertained whilst writing. Lastly, to my kind, generous and extraordinary parents, everything I have been able to achieve has been due to your love and support, thank you.

Publications

Publications arising from this thesis

S E Hickman, R Woitek, E P V Le *et al*. Machine learning algorithms for workflow applications in screening mammography: a systematic review and meta-analysis. Radiology. 2021. https://doi.org/10.1148/radiol.2021210391

S Hickman, G Baxter, F J Gilbert. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. British Journal of Cancer. 2021. https://doi.org/10.1038/s41416-021-01333-w

Publications arising from work unrelated to this thesis

F J Gilbert, **S E Hickman**, G C Baxter *et al*. Opportunities in cancer imaging: risk-adapted breast imaging in screening. Clinical Radiology. 2021. https://doi.org/10.1016/j.crad.2021.02.013

I Allajbeu, **S E Hickman**, N Payne *et al.* Automated Breast Ultrasound: Technical Aspects, Impact on Breast Screening, and Future Perspectives. Breast Cancer Screening and Imaging. 2021. https://doi.org/10.1007/s12609-021-00423-1

I Dayan, H R Roth, A Zhong … **S E Hickman** *et al*. Federated learning for predicting clinical outcomes in patients with COVID-19. Nature Medicine. 2021. https://doi.org/10.1038/s41591-021-01506-3

E P V Le, Y Wang, Y Huang, **S Hickman** *et al*. Artificial Intelligence in Breast Imaging. Clinical Radiology. 2019. https://doi.org/10.1016/j.crad.2019.02.006

Presentations

Oral

S E Hickman, N R Payne, Y Huang *et al*. A benchmarking study to evaluate the performance of two artificial intelligence algorithms for interval cancer detection in a UK breast screening setting. Radiological Society of North America Conference, Chicago, USA, December 2021.

S Hickman, J G Mainprize, R Black *et al*. Mammographic case conspicuity, a comparison between a radiologist's assessment and a Masking Index. European Congress of Radiology Conference, virtual, July 2020.

S Hickman, Pantelidou M, Black R *et al*. Masking Risk Index: an evaluation to guide supplemental imaging for breast screening. British Society of Breast Radiology Annual Scientific Conference, Bristol, UK, November 2019. (Prize: best oral presentation)

Poster

S Hickman, S Hudson, N R Payne *et al*. The creation of a breast screening image database – The Cambridge Cohort – Mammography East Anglia Digital Imaging Archive (CC-MEDIA). British Society of Breast Radiology Annual Scientific Conference, virtual, November 2021.

S Hickman, R A Woitek, Y R Im *et al*. Independent machine learning algorithms for workflow adaptation in breast screening mammography: a systematic review and meta-analysis. European Congress of Radiology, virtual, March 2021.

Key collaborator contributions

Code for the analysis in the systematic review and meta-analysis was written by Dr Gaby Baxter.

Code written by Sue Hudson, Dr Andrew Priest, Dr Nicholas Payne, and Dr Lorena Escudero Sánchez was used in the creation of the CC-MEDIA database.

The set-up of the algorithm testing infrastructure at the University of Cambridge was made possible by work from Richard Black and Dr Nicholas Payne.

Three commercial companies provided their artificial intelligence algorithms for analysis as part of this thesis.

Dr Yuan Huang provided statical supervision for the analysis performed in Chapters 5-7.

Table of contents

Declaration *2*

Abstract *3*

Acknowledgements *4*

Publications *6*

Presentations *7*

Key collaborator contributions *8*

Table of contents *9*

List of figures *13*

List of tables *16*

Commonly used abbreviations *19*

Chapter 1 – Introduction *21*

1.1 Breast cancer **21**

1.1.1 Breast cancer overview 21

1.1.2 Breast cancer classification 21

1.2 Breast cancer screening **24**

1.2.1 Breast cancer screening programmes 24

1.2.2 Mammography 26

1.2.3 Mammographic breast density 28

1.2.4 Risk prediction 30

1.2.5 Interval cancers 31

1.3 Artificial intelligence in breast cancer screening **31**

1.3.1 Introduction to artificial intelligence 31

1.3.2 History of computer aided detection systems in breast cancer screening 34

1.3.3 Deep Learning applications to breast cancer screening 35

1.4 Thesis aims and outline **37**

Chapter 2 – Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations *40*

2.1 Introduction **40**

2.2 Evaluation of artificial intelligence in breast imaging **41**

2.2.1 Retrospective evaluation 41

2.2.2 Prospective evaluation 43

2.2.3 Key considerations for clinical evaluation 44

2.3 The breast imaging pathway and AI **45**

2.3.1 Screening 45

2.3.2 Risk stratification 46

2.3.3 Monitoring and prognostication 46

2.4 Ethical and legal constraints **47**

2.4.1 Guidance level 47
2.4.2 Algorithm level 47
2.4.3 Who controls the data? 48
2.4.4 Clinical level 49

2.5 Practical challenges and limitations 50
2.5.1 Technical level 50
2.5.2 Clinical level 51
2.5.3 Governance level 51

2.6 Conclusion 52

Chapter 3 – Machine learning for workflow applications in screening mammography: systematic review and meta-analysis 53

3.1 Introduction 53

3.2 Materials and methods 55
3.2.1 Literature search 55
3.2.2 Study selection 55
3.2.3 Data extraction 55
3.2.4 Meta-analysis 56
3.2.5 Quality assessment 56
3.2.6 Statistical analysis 56

3.3 Results 57
3.3.1 Statistical selection and data extraction 57
3.3.2 Quality assessment 68
3.3.3 Statistical analysis 69

3.4 Discussion 70
3.4.1 Limitations 72

3.5 Conclusion 72

Chapter 4 – Developing a mammographic imaging database – The Cambridge Cohort – Mammography East Anglia Digital Imaging Archive 74

4.1 Aims 74

4.2 Introduction 74

4.3 Methods 76
4.3.1 Database approval 76
4.3.2 Database governance 76
4.3.3 Patient and public involvement work 77
4.3.4 Database sites 80
4.3.5 Database creation 81

4.4 Results 85
4.4.1 Database image content 85
4.4.2 Database content - Interval cancers 86
4.4.3 Database content - Screen detected cancers 88
4.4.4 Database content - Ethnicity 89
4.4.5 Database content - Mammographic breast density 91
4.4.6 Database content - Histopathological information 92

4.5 Technical setup of an AI algorithm testing environment 92

4.6 Uses of the database 92

4.7 Discussion 93
4.7.1 Overall discussion 93

4.7.2 Limitations 93
4.8 Conclusion 94
Chapter 5 – Performance of artificial intelligence algorithms for interval cancer detection 95
5.1 Aims 95
5.2 Introduction 95
5.3 Methods 96
5.3.1 Sample size 96
5.3.2 Data 96
5.3.3 Ground truth 98
5.3.4 AI tools 98
5.3.5 Thresholds 99
5.3.6 Statistical analysis 100
5.3.7 Reporting 101
5.4 Results 101
5.4.1 Data 101
5.4.2 Algorithm results 103
5.4.3 Combined algorithm results 107
5.4.4 Sub-group analysis 108
5.4.5 Failure analysis 111
5.5 Discussion 112
5.6 Conclusion 114
Chapter 6 – Performance of stand-alone deep learning algorithms in a UK screening cohort for detection and diagnosis 116
6.1 Aims 116
6.2 Introduction 116
6.3 Methods 117
6.3.1 Sample size 117
6.3.2 Data 117
6.3.3 Ground truth 119
6.3.4 AI tools 120
6.3.5 Thresholds 120
6.3.6 Statistical analysis 122
6.3.7 Reporting 123
6.4 Results 123
6.4.1 Data 123
6.4.2 Algorithm results 126
6.4.3 Scenario D 99.0% specificity auto recall threshold 131
6.4.4 Combined algorithm results 132
6.4.5 Sub-group analysis 134
6.4.6 Failure analysis 138
6.5 Discussion 140
6.5.1 Overall performance 140
6.5.2 Further analysis 141
6.5.3 Limitations 142
6.5.4 Future work 142
6.6 Conclusion 143

Chapter 7 - Performance of stand-alone artificial intelligence algorithms in a UK screening cohort for high sensitivity and high specificity triage 144

7.1 Aims 144

7.2 Introduction 144

7.3 Methods 146
7.3.1 Data 146
7.3.2 Ground truth 147
7.3.3 AI tools 148
7.3.4 Thresholds 148
7.3.5 Statistical analysis 150
7.3.6 Reporting 151

7.4 Results 151
7.4.1 Data 151
7.4.2 Rule-out triage – Threshold 1 and 2 155
7.4.3 Rule-out triage – Threshold 3 and 4 159
7.4.4 Rule-in triage 163
7.4.5 Combined approach 166
7.4.6 Sub-group analysis 168
7.4.7 Failure analysis 173

7.5 Discussion 174
7.5.1 Overall performance 174
7.5.2 Further analysis 175
7.5.3 Limitations 175

7.6 Conclusion 176

Chapter 8 – Contributions, Future Work and Conclusions 177

8.1 Contributions to knowledge 177

8.2 Future work 178
8.2.1 AI in the NHS 178
8.2.2 Retrospective studies 179
8.2.3 Prospective studies 179
8.2.4 Future work - AI research questions 181

8.3 Conclusions 182

References 184

List of figures

Figure 1-1 – Breast cancer anatomy 22
Figure 1-2 – Production of x-rays diagram 27
Figure 1-3 – Example of a two-view full field digital mammogram (FFDM) 28
Figure 1-4 – Examples of breast imaging-reporting and data system (BI-RADS) 5th edition mammographic breast density categories 29
Figure 1-5 – Artificial intelligence (AI) hierarchy of terms 32
Figure 1-6 – Overview of the architecture of a Convolutional Neural Network (CNN) 33
Figure 1-7 – Application of Deep Learning (DL) Computer Aided Detection (CAD) algorithms to breast cancer screening 36

Figure 2-1 – Broad and narrow artificial intelligence (AI) applications to breast imaging 40

Figure 3-1 – Multi-time (left) and multi-view (right) point data that are produced by 2D standard-view mammography and can be analysed at different levels 53
Figure 3-2 – Preferred Reporting Items for Systematic Reviews and Meta-analysis for Diagnostic Test Accuracy (PRISMA-DTA) flow diagram 58
Figure 3-3 – (a) Prediction model Risk Of Bias ASsessment Tool (PROBAST) and (b) Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) assessment 68
Figure 3-4 – Checklist for Artificial Intelligence in Medical Imaging (CLAIM) assessment 69
Figure 3-5 – Summary Receiver Operating Characteristic (ROC) curves 70

Figure 4-1 – Cambridge Science Festival Event questions 78
Figure 4-2 – National patient survey question regarding acceptability of data fields 79
Figure 4-3 – National patient survey questions regarding commercial involvement 79
Figure 4-4 – Data flow of the CC-MEDIA data collection 81
Figure 4-5 – Nomenclature of case de-identification within the CC-MEDIA database 83
Figure 4-6 – Timeline of mammography data changes over time at Cambridge and Norwich National Health Service Breast Screening Programme (NHSBSP) sites 84
Figure 4-7 – Time to diagnosis (months) for interval cancers (IC) at a) Cambridge and b) Norwich 87
Figure 4-8 – Ethnicity data distribution at Cambridge using National Breast Screening System (NBSS) and Electronic Health Record (EHR) EPIC data 91

Figure 4-9 – Breast imaging-reporting and data system (BI-RADS) 5th edition mammographic density distribution for cases in one year (2017) of data at Cambridge with both raw and processed four views mammograms available [n = 18246] 91

Figure 5-1 – Standards for Reporting of Diagnostic Accuracy Studies (STARD) flow diagram of cases included and excluded in this study..... 97
Figure 5-2 – Example of cases included in the study 98
Figure 5-3 – Proposed workflow image for testing the artificial intelligence (AI) systems as stand-alone readers for interval cancer (IC) detection 100
Figure 5-4 – Receiver operating characteristic (ROC) curves for all three artificial intelligence (AI) algorithms at each site..... 104
Figure 5-5 – Cambridge data testing density plots for each artificial intelligence (AI) algorithm..... 105
Figure 5-6 – Norwich data testing density plots for each artificial intelligence (AI) algorithm..... 107
Figure 5-7 – Combined model receiver operating characteristic (ROC) curve on Cambridge data compared to individual artificial intelligence (AI) algorithms (DL-1, DL-2, DL-3) performance 107
Figure 5-8 – Combined model receiver operating characteristic (ROC) curve on Norwich data compared to individual artificial intelligence (AI) algorithms (DL-1, DL-2, DL-3) performance 108
Figure 5-9 – Proportional Euler diagram of each artificial intelligence (AI) algorithms interval cancer (IC) detection 111
Figure 5-10 – False negative case, which was not detected by all three commercial artificial intelligence (AI) algorithms..... 111
Figure 5-11 – True positive case, which was detected by all three commercial artificial intelligence (AI) algorithms 112

Figure 6-1 – Mediolateral oblique (MLO) views of mammogram artefacts removed from the study 118
Figure 6-2 – Standards for Reporting of Diagnostic Accuracy Studies (STARD) flow diagram of cases included and excluded in this study..... 119
Figure 6-3 – Proposed workflow deployment of a stand-alone computer aided detection and diagnosis (CADe+x) artificial intelligence (AI) algorithm..... 121
Figure 6-4 – Receiver operating characteristic (ROC) curves per artificial intelligence (AI) algorithm 126
Figure 6-5 – Receiver operating characteristic (ROC) curves per site..... 127
Figure 6-6 – Precision recall curves (PRC)..... 128

Figure 6-7 – Individual artificial intelligence (AI) algorithm score distributions normalised from 0-10 130
Figure 6-8 – Combined model receiver operating characteristic (ROC) curves on Cambridge data . 133
Figure 6-9 – Combined model receiver operating characteristic (ROC) curves on Norwich data 134
Figure 6-10 – Venn diagram – not proportional 138
Figure 6-11 – Missing case analysis, case missed by both artificial intelligence (AI) and human readers 138
Figure 6-12 – Missing case analysis, case missed by all human readers and detected by all artificial intelligence (AI) algorithms 139
Figure 6-13 – Missing case analysis, case missed by all artificial intelligence (AI) algorithms 139

Figure 7-1 – Standards for Reporting of Diagnostic Accuracy Studies (STARD) flow diagram of cases included and excluded in this study 147
Figure 7-2 – Cancer outcomes for study cohort 148
Figure 7-3 – Proposed workflow deployment approaches for stand-alone artificial intelligence (AI) systems as triage tools 150
Figure 7-4 – Receiver operating characteristic (ROC) curves for screen detected cancers (SDCs) as cases 156
Figure 7-5 – Receiver operating characteristic (ROC) curves for screen detected cancers (SDCs), next round cancers (NRCs) and interval cancers (ICs) as cases 160
Figure 7-6 – Plots for rule out triage thresholds 163
Figure 7-7 – Plots for rule in triage thresholds – Screen detected cancers (SDCs), next round cancers (NRCs) and interval cancers (ICs) 166
Figure 7-8 – Violin plots for the combined approach of Scenario C and E for both rule in and rule out triage by an artificial intelligence (AI) algorithm 166
Figure 7-9 – Partial receiver characteristic (pROC) curves 168
Figure 7-10 – Venn diagram – not proportional, for screen detected cancers (SDCs) missed at threshold 1, Scenario B 173
Figure 7-11 – Missing case analysis, case missed by artificial intelligence (AI) 173
Figure 7-12 – Venn diagram – not proportional, for a) interval cancers (ICs) and b) next round cancers (NRCs) detected at the 94.0% specificity threshold Scenario E 174

List of tables

Table 1-1 – Breast cancer molecular subtypes classification 23
Table 1-2 – Breast cancer screening programmes and committee recommendations 25

Table 2-1 – Datasets publicly and privately available for breast imaging 42
Table 2-2 – Prospective studies for the use of artificial intelligence (AI) in breast imaging 43
Table 2-3 – Reporting criteria adapted for artificial intelligence (AI) studies 44

Table 3-1 – Computer aided triage (CADt) algorithm details and results. Algorithm performance compared to reader performance for all included studies 60
Table 3-2 – Computer aided triage (CADt) test set data characteristics of all included studies 62
Table 3-3 – Computer aided detection (CADe) and Computer aided diagnosis (CADx) algorithm details and results. Algorithm performance compared to reader performance for all included studies 65
Table 3-4 – Computer aided detection (CADe) and Computer aided diagnosis (CADx) test set data characteristics of all included studies 67

Table 4-1 – Mammographic imaging database characteristics 75
Table 4-2 – Number of exams per site available with images currently held in the CC-MEDIA database 85
Table 4-3 – Cambridge and Norwich CC-MEDIA database 2011-2020 compared to the KC62 report at both sites 86
Table 4-4 – Interval cancers (ICs) at Cambridge and Norwich with imaging data 2011-2020 in CC-MEDIA 87
Table 4-5 – Screen detected cancers (SDCs) at Cambridge and Norwich with imaging data 2011-2020 in CC-MEDIA 89
Table 4-6 – Ethnicity information from National Breast Screening System (NBSS) and Electronic Health Record (EHR) EPIC data at Cambridge 90

Table 5-1 – Artificial intelligence (AI) algorithm characteristics 99
Table 5-2 – Summary of testing dataset characteristics. 102
Table 5-3 – Interval cancer (IC) characteristics by case 103
Table 5-4 – Interval cancer (IC) characteristics by lesions 103

Table 5-5 – Cambridge data testing of three artificial intelligence (AI) algorithms 105
Table 5-6 – Norwich data testing of three artificial intelligence (AI) algorithms 106
Table 5-7 – Subgroup analysis of cases using all interval cancer (IC) data from both Cambridge and Norwich sites 110
Table 5-8 – Subgroup analysis of lesions using all interval cancer (IC) data from both Cambridge and Norwich sites 110

Table 6-1 – Summary of testing dataset characteristics 123
Table 6-2 – Cancer characteristics by lesions and cases 124
Table 6-3 – Interval cancer (IC) characteristics by lesions and cases 125
Table 6-4 – Stand-alone artificial intelligence (AI) algorithm application compared to the single first reader – threshold 1 129
Table 6-5 – Stand-alone artificial intelligence (AI) algorithm application compared to the single first reader – threshold 2 129
Table 6-6 – Artificial intelligence (AI) algorithm (at threshold 2) combined with the single first reader (+/- arbitration where discordance) compared to double reading performance 130
Table 6-7 – Perturbation analysis when adjusting the specificity threshold for the artificial intelligence (AI) algorithm, then combining with the first reader and final action arbitration decision if there is discordance 131
Table 6-8 – Artificial intelligence (AI) algorithm (at threshold 2) combined with the single first reader (+/- arbitration where discordance below 99.0% specificity for the algorithm and above 96.6% specificity) with cases auto recalled above the 99.0% specificity threshold (threshold 3) compared to double reading performance 132
Table 6-9 – DeLong's test comparison results for DL-1, DL-2, DL-3 compared to the Combined model performance on Cambridge data 133
Table 6-10 – DeLong's test comparison results for DL-1, DL-2, DL-3 compared to the Combined model performance on Norwich data 134
Table 6-11 – Sub group analysis of DL-1, DL-2, DL-3 set at the first reader specificity threshold of 96.6% (threshold 1) for screen detected cancers (SDCs) 135
Table 6-12 – Sub group analysis of DL-1, DL-2, DL-3 set at the first reader specificity threshold of 96.6% (threshold 1) for interval cancers (IC) 136
Table 6-13 – Sub group analysis of DL-1, DL-2, DL-3 set at the first reader specificity threshold of 96.6% (threshold 1) for interval cancer (IC) specific categories 137

Table 7-1 – Summary of testing dataset characteristics 152
Table 7-2 – Screen detected (SDC) and next round cancer (NRC) characteristics by lesions and cases 153
Table 7-3 – Interval cancer (IC) characteristics by lesions and cases 154
Table 7-4 – Double and single first reader performance at both Cambridge and Norwich 155
Table 7-5 – Results at 1) 99.0% sensitivity threshold 1 and 2) 99.9% sensitivity threshold 2 157
Table 7-6 – Results for DL-1, DL-2 and DL-3 at the 99.0% sensitivity (threshold 1) Scenario B 158
Table 7-7 – Results for DL-1, DL-2 and DL-3 at the 99.9% sensitivity (threshold 2) Scenario B 158
Table 7-8 – Results for DL-1, DL-2 and DL-3 at the 99.0% sensitivity (threshold 1) Scenario C 159
Table 7-9 – Results at 1) 85.0% sensitivity (threshold 3) and 2) results at 70.0% specificity (threshold 4) 161
Table 7-10 – Results for DL-1, DL-2 and DL-3 at the 85.0% sensitivity (threshold 3) Scenario C 161
Table 7-11 – Results for DL-1, DL-2 and DL-3 at the 70.0% specificity (threshold 4) Scenario C 162
Table 7-12 – Scenario D perturbations of specificity with screen detected cancers (SDCs), interval cancers (ICs) and next round cancers (NRCs) as cases 164
Table 7-13 – Scenario D perturbations of specificity with screen detected cancers (SDCs), interval cancers (ICs) and next round cancers (NRCs) as cases – additional cancers detected. 164
Table 7-14 – Scenario E perturbations of specificity with screen detected cancers (SDCs), interval cancers (ICs) and next round cancers (NRCs) as cases 165
Table 7-15 – Scenario E perturbations of specificity with screen detected cancers (SDCs), interval cancers (ICs) and next round cancers (NRCs) as cases – additional cancers detected. 165
Table 7-16 – Combined approach of Scenario C and E for both rule in and rule out triage by an artificial intelligence (AI) algorithm 167
Table 7-17 – Partial area under the receiver operator characteristic (pAUROC) curve results 167
Table 7-18 – Sub group analysis of DL-1, DL-2, DL-3 set at the threshold of 99.0% sensitivity (threshold 1) using Scenario B for the screen detected cancers (SDCs) missed 169
Table 7-19 – Sub group analysis of DL-1, DL-2, DL-3 set at 96.0% specificity threshold, using Scenario E for the interval cancers (ICs) detected 171
Table 7-20 – Sub group analysis of DL-1, DL-2, DL-3 set at 96.0% specificity threshold, using Scenario E for the next round cancers (NRCs) detected 172

Commonly used abbreviations

AI	Artificial Intelligence
AUROC	Area Under the Receiver Operating Characteristic Curve
AUPRC	Area Under the Precision Recall Curve
BI-RADS	Breast Imaging-Reporting and Data System
BSP	Breast Screening Programme
BRAID	Breast Screening – Risk Adaptive Imaging for Density
CC	Craniocaudal
CADe	Computer Aided Detection
CADx	Computer Aided Diagnosis
CADt	Computer Aided triage
CC-MEDIA	The Cambridge Cohort – Mammography East Anglia Digital Imaging Archive
CLAIM	Checklist for Artificial Intelligence in Medical Imaging
CNN	Convolutional Neural Network
CI	Confidence Interval
DAC	Database Access Committee
DBT	Digital Breast Tomosynthesis
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
EHR	Electronic Health Records
FFDM	Full Field Digital Mammography
FRC	Future Round Cancer
IC	Interval Cancer
MRI	Magnetic Resonance Imaging
ML	Machine Learning
MLO	Mediolateral Oblique
NHS	National Health Service
NRC	Next Round Cancer
NRIC	Next Round Interval Cancer
NSC	National Screening Committee
OMI-DB	The Optimam Mammography Image Database

pAUC	Partial Area Under the Receiver Operating Characteristic Curve
PACS	Picture Archiving Communication Systems
PPI	Patient and Public Involvement
PROBAST	Prediction model Risk Of Bias ASsessment Tool
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies 2
ROC	Receiver Operating Characteristics
TC	Tyrer-Cuzick
SDC	Screen Detected Cancer
STARD	Standards for Reporting Diagnostic Accuracy Studies
VBD	Volumetric Breast Density

Chapter 1 – Introduction

1.1 Breast cancer

1.1.1 Breast cancer overview

Breast cancer is the most common malignancy diagnosed in women with 2.3 million new diagnoses globally each year[1]. Approximately 1 in every 8 women will be diagnosed with breast cancer in their lifetime, such that it accounts for 15.2% of all new cancers diagnosed in the UK, with 45,000 women diagnosed each year[2,3]. It is the leading cause of cancer related death amongst women as well as the fifth leading cause of cancer death world-wide, equating to 685,000 deaths in 2020[1,3–5]. Risk factors for the development of breast cancer include; female sex, age, lifestyle (e.g. alcohol and smoking), family history, genetic mutations (e.g. BReast CAncer gene (BRCA)), increased breast density, history of breast disease, hormone exposure and expression, and radiation exposure [1,6,7]. Breast cancer can be detected through two routes. The first is through symptomatic presentation where a woman presents with a painless lump, skin changes or nipple discharge[1]. The second route is asymptomatic detection as part of a screening programme using imaging, most commonly mammography, leading to the earlier detection of cancer before the onset of symptoms[3].

Breast cancer is a heterogeneous disease with a diverse range of morphological imaging features as well as biological and molecular tumour sub-types[8]. The complexity of this disease means a targeted response at each stage of the care pathway, in which imaging has key role, is required to achieve the best outcomes. Survival rates continue to improve with advances in screening, imaging techniques for diagnosis and monitoring as well as the development of novel and targeted therapies for treatment[9]. The five year survival rate is around 85.0% in the UK, however it is only 26.2% for women diagnosed with stage four disease[10,11]. The early detection of breast cancer, through methods such as mammographic screening is proven to reduce both morbidity and mortality[1,7,12,13].

1.1.2 Breast cancer classification

Breast cancer can be characterised by histopathological type, grade, immunohistochemical profile and gene expression, as well as anatomical extent / staging[3,7]. Correct classification allows for the prediction of response to treatment as well as overall prognostication of survival, by using tools such as Adjuvant online, PREDICT and the Nottingham prognostic index, thus, allowing for the targeted selection of treatment which can range from radiotherapy, chemotherapy, hormone therapy, targeted biological therapy, and surgery[3,14].

Most breast cancers arise from the epithelial lining of the terminal ductal lobular unit (TDLU) and are invasive cancers, either invasive ductal carcinoma or invasive lobular carcinoma, Figure 1-1. Invasive carcinomas extend beyond the basement membrane into the surrounding tissues and can potentially metastasise. Invasive ductal carcinomas (IDC), or as it is otherwise known invasive carcinoma of no special type (NST), is the most common type (70-80%) of breast cancer, whereas 5-15% of breast cancers are invasive lobular carcinomas[14–17]. The precursor to invasive carcinoma is non-invasive carcinoma where the cancer cells have not spread beyond the originating structure and remain 'in-situ'[3,6].

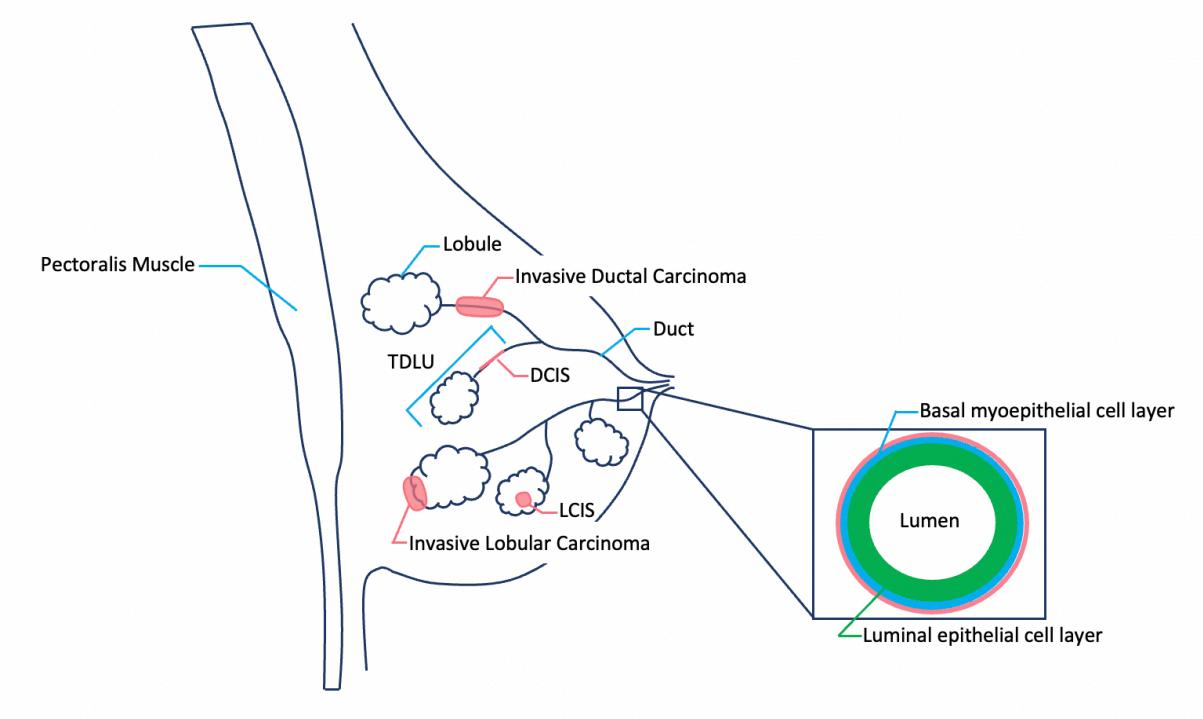


Figure 1-1 – Breast cancer anatomy. *Showing the different anatomical and histological structures, as well as the types of cancers that arise from these structures. DCIS: Ductal carcinoma in situ, LCIS: Lobular carcinoma in situ, TDLU: Terminal ductal lobular unit. Adapted from Harbeck et al[3] and Feng et al[6].*

Histological type classification is made using the tumour cell type, architectural features and the immunohistochemical profile[14]. The World Health Organization (WHO) classification of tumours series fifth edition, published in 2019, divides breast cancers into the following two main histological type categories; invasive breast carcinoma of NST and invasive breast carcinomas of special type[18,19]. Invasive breast carcinoma of NST includes; pleomorphic, oncocytic, lipid-rich, glycogen-rich clear cell, sebaceous, osteoclast-like stromal giant cells, or carcinomas with choriocarcinomatous or melanotic patterns, as well as medullary features which are considered tumour-infiltrating lymphocyte rich invasive breast carcinoma of NST (TIL-rich IBC-NST). Invasive breast carcinomas of special type includes: lobular, tubular, cribriform, metaplastic, apocrine, mucinous, papillary, and micropapillary[18,20]. Tumours can also be of mixed type, such that they contain multiple subtypes and

the proportion of each should be reported[16]. Histological type alone does not provide enough information regarding the true heterogeneity seen in breast cancers and thus the additional classification categories are used for prognostication to determine treatments and predict survival. Tumour grade provides information regarding the degree of differentiation of the tumour cells from normal breast epithelial cells[17]. The histological grading system used by breast pathologists is the Nottingham Grading System (NGS), which was developed by Elston and Ellis and modified from the Scarff-Bloom Richardson grading system[14,21]. The NGS grading system assesses three components of tumour morphology (in invasive cancers only): tubule formation, nuclear pleomorphism, and mitotic count[17]. Each component is scored from 1-3 and adding these scores together gives the total count which relates to the overall grade (grade 1 = scores 3-5, grade 2 = scores 6-7, and grade 3 = scores 8-9)[22]. Grade 3 tumours are often larger and grow more rapidly leading to a worse prognosis[17].

St. Gallen International Expert Consensus molecular subtype definitions, classify breast cancer into five categories based on results of immunohistochemistry for the following markers: oestrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor 2 receptor (HER2 / ERBB2) as well as Ki-67 which is a proliferation marker protein[3,23–26]. The five categories are outlined in Table 1-1. Luminal A is the most common subtype and has the best prognosis out of the five categories.

Subtype	ER	PR	HER2	Ki-67	Prognosis
Luminal A	+	+	-	Low	Good
Luminal B (HER2 -)	+	+	-	High	Intermediate
Luminal B (HER2 +)	+	+	+	High	Poor
Triple negative (Basal)	-	-	-	High	Poor
HER2-enriched	-	-	+	Moderate-High	Poor
Normal-like	+	+	-	Low	Intermediate

Table 1-1 – Breast cancer molecular subtypes classification. *+: Positive, - : Negative. Adapted from Dai et al[23], Harbeck et al[3] and Feng et al[6].*

An anatomical staging classification system published by the American Joint Committee on Cancer (AJCC) uses the extent of the primary tumour (Tis to T4), regional lymph nodes (N0 to N3) and metastases (M0 or M1), resulting in a TMN status which relates to the five stage categories (0-IV)[27]. The eighth edition, published in 2017 of the AJCC classification system incorporated changes to account for the prognostic stage, which includes; tumour grade using the NGS, biomarkers, and multigene panels (e.g. Oncotype DX)[27–29]. The incorporation of this prognostic information can result in a 'stage migration', such that triple negative cancers (ER -, PR -, HER2 -) or grade 3 cancers would be 'upstaged' and HER2 positive or grade 1 cancers would be 'downstaged'.

Future classification, using techniques such as next generation sequencing, will allow for the identification of additional breast cancer sub-types to further tailor treatment approaches[30].

1.2 Breast cancer screening

1.2.1 Breast cancer screening programmes

Population based screening programmes are designed around the ten Wilson and Jungner principles, published by the WHO in 1968 and subsequent modifications of these principles[31]. National screening programmes are conducted to identify certain diseases in an asymptomatic population, for earlier diagnosis and to facilitate prompt treatment[32]. Breast cancer screening aims to "maximise the success of treatment, reducing mortality from breast cancer"[33]. Most European countries as well as many countries across the globe have either a national or regional population breast cancer screening programme[24]. Though it is recognised that these types of organised screening programmes are limited to high-income countries[34]. The 'Marmot review' 2012 and recent meta-analyses of randomised control trials estimate a 15-30% reduction in mortality due to mammographic screening[35,36]. A study published by Duffy *et al* in 2020 found in a Swedish screening population a 34% reduction in 10-year mortality through participation in screening which was independent to changes in treatment[37]. However, the risk of harm from screening such as false positives resulting in subsequent increase in patient anxiety, overdiagnosis (estimated at 11-19%), and overtreatment, whilst difficult to truly gauge, needs to be balanced against the benefits of screening[36,38,39].

This research thesis primarily considers the UK National Health Service Breast Screening Programme (NHSBSP). Following the recommendations of the 'Forrest Report' 1986, the NHSBSP commenced in 1988, with women aged 50-64 years old screened every three years with a one view mediolateral oblique (MLO) screen film mammogram read by a single reader[40–42]. The NHSBSP has evolved over time and now uses two-view full field digital mammograms (FFDM) as well as double reading of each mammogram. Women aged 50-70 years old are invited to attend every three years at one of the seventy-five screening units across the UK[43]. Women can also self-refer after the age of 70 years old and continue with screening every three years. This deviates from the current European Commission Initiative on Breast Cancer (ECIBC) guidelines, where screening is recommended for women age 50-69 years old every two years, women aged 70-74 every three years and every two to three years for women aged 45-49 years old[38,44,45]. Guidance from the American College of Radiology (ACR) and Society of Breast Imaging (SBI) differs too. ACR SBI recommend annual mammography is to start at the age of 40 years old[46]. A summary of the variations between different screening programmes and committee recommendations are shown in Table 1-2.

	Interval	Age	Modality
UK[43]	Triennial	50-70+	FFDM
Sweden[47]	Biennial	40-74	FFDM
Netherlands[48]	Biennial	50-75	FFDM
Norway[49]	Biennial	50-69	FFDM
Australia[50]	Opt-in	40-49	FFDM
	Biennial	50-74	FFDM
	Opt-in	74+	FFDM
China[51]	Annual-Triennial	20-39	Examination (self-exam monthly)
	Annual-Biennial Annual	40-69	FFDM + *USS Examination (self-exam monthly)
	Annual	70+	Examination (self-exam monthly)
ECIBC[44,45]	Biennial / Triennial	45-49	FFDM
	Biennial	50-69	FFDM
	Triennial	70-74	FFDM
ACR[46]	Annual	40+	FFDM
USPSTF[52]	Individual decision	40-49	FFDM
	Biennial	50-74	FFDM
CTFPHC[53]	Shared decision	40-49	FFDM
	Biennial / Triennial	50-74	FFDM

Table 1-2 – Breast cancer screening programmes and committee recommendations. *Detailing the recommendations for the frequency, age and modality of screening from different screening programmes and screening committee recommendations. Adapted from Clift et al[39]. ACR: American College of Radiology, CTFPHC: Canadian Task Force on Preventive Health Care, ECIBC: European Commission Initiative on Breast Cancer (ECIBC), FFDM: Full field digital mammography, USS: Ultrasound, USPSTF: US Preventive Services Task Force. *In patients with dense breasts only.*

Approximately 2.5 million women aged 50-70 years old are invited for screening as part of the NHSBSP each year, with 1.8 million women attending screening (~71% uptake) and 66,000 (~3.7%) being recalled to attend an assessment clinic[33,54]. The NHSBSP has set a standard to ensure the number of women recalled to assessment is not too high through maximising specificity. The recall rate is set at an acceptable level of 10% for prevalent (first round of screening) and 7% for incident (not first round of screening) screens, as well as an achievable level of 7% for prevalent and 5% for incident screens[33]. Each year ~15,000 cancers are diagnosed through screening (0.8% women screened) and the NHSBSP has set an age standardised detection ratio for invasive cancer of ≥ 1.00 for acceptable and ≥ 1.40 for achievable, in women age 50-70 years old[33,54]. The screening programme aims to detect clinically significant cancers to reduce mortality, such that it is important screening programmes detect small cancers, which are defined by the NHSBSP as an invasive cancer < 15mm. The NHSBSP standard is set at an age standardised detection ratio of ≥ 1.00 for acceptable

and ≥ 1.40 for achievable regarding the detection of small invasive cancers. Approximately 51.9% of invasive cancers detected through screening are small[33,54].

In the NHSBSP each mammogram is read on 5-megapixel monitors by two expert readers[55]. Every reader must report 5,000 mammograms each year and also undertake the PERsonal perFORmance in Mammographic Screening (PERFORMS) assessment to monitor reader standards[56]. Double reading can either be blinded (independent) such that the readers are unable to see the other reader's decision or unblinded (dependent) so that the readers are able to see each other's decision. Where there is discordance between readers arbitration / consensus reading takes place by a third reader or group of readers. Double reading is shown to pick up an additional 9% of cancers compared to single reading[55,57]. However, there is a national shortage of screen readers in the UK with this shortage expected to increase over the next five years[58].

A women's life time risk of developing breast cancer is ~11%[59]. In the UK breast screening is currently adapted for those with a high risk (> 30% lifetime risk) of breast cancer, where women are offered annual magnetic resonance imaging (MRI) and / or mammography depending on their age[38]. Moderate risk (17-30% lifetime risk) women are also offered annual mammography[59]. Further risk adaptation using breast density, polygenic risk scores as well as risk prediction models is currently being investigated and discussed later in this chapter.

1.2.2 Mammography

A mammogram is an image of the breast acquired using low energy x-rays and mammography is the primary imaging technique used world-wide for breast cancer screening. The process of x-ray production in mammography is as follows; electrons are released from a filament via a process of thermionic emission and are then accelerated away from the negatively charged cathode across the vacuum tube towards the positively charged anode. The tube voltage is the potential difference between the anode and cathode which causes the acceleration of the electrons across the tube towards the anode when they are emitted[60]. The electrons then hit the rotating anode target (which can be made out of a number of materials e.g. tungsten, molybdenum, or rhodium) which causes the emission of x-rays through Bremsstrahlung radiation[61]. The anode is rotated to dissipate the heat generated by the bombardment of electrons. These x-rays are then directed towards the patient via a window in the lead shielding. Low and high energy photons are removed via a filter as well as the beam is focused using a collimator, as shown in Figure 1-2. The x-ray beam is attenuated differently by different tissues in the body giving contrast in the resulting image. X-rays are attenuated to a greater degree by denser fibroglandular tissue than by fatty tissue such that dense fibroglandular tissue appears whiter in the x-ray image. The same is true of many lesions and microcalcifications which is why mammography is an appropriate imaging modality for screening. The target / filter

combination and the exposure factors (tube voltage, tube current and exposure time) can be adjusted to increase the image quality whilst maintaining an acceptable level of dose for each woman. The recommended dose per mammogram according to the National Diagnostic Reference Levels (NRLs) is 2.5 mGy / mean glandular dose[62].

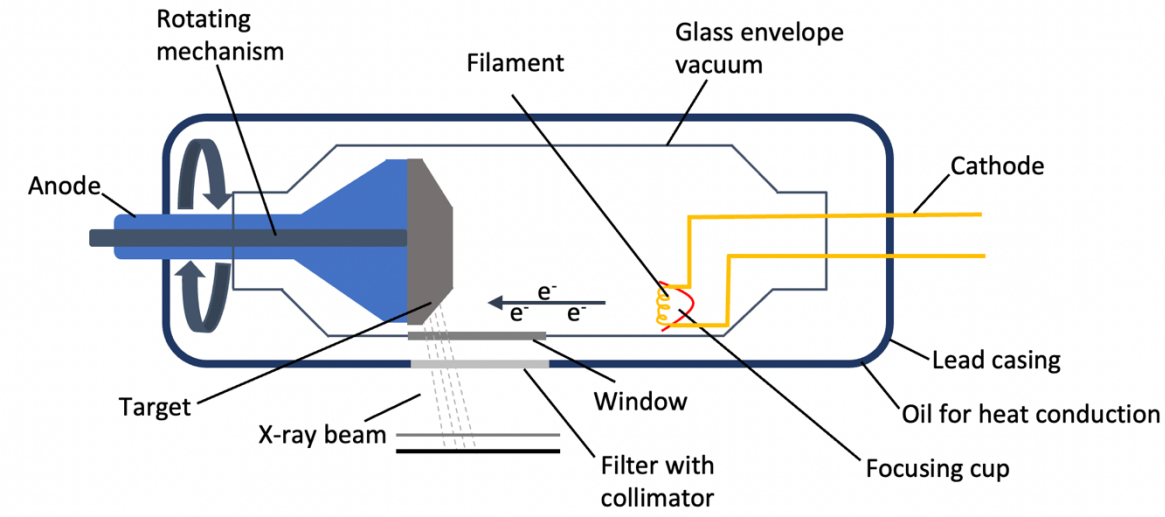


Figure 1-2 – Production of x-rays diagram. *Showing the different components of an x-ray machine for mammography. Adapted from Radiology Cafe[60].*

A mammogram is made up of the craniocaudal (CC) and MLO views of the right and left breast, creating a two-view mammogram[63]. Mammography is held to high technical quality standards to ensure the images are of adequate quality to allow for interpretation and minimising errors[64]. These standards include ensuring there are no skin folds, blurring or artefacts included in the image as well as ensuring that the whole breast is included and the nipple is in profile in at least one view. Methods to ensure the whole breast is included entail using the posterior nipple line (PNL) and making sure the inframammary angle is included in the MLO view, as shown in Figure 1-3. Common mammographic signs of cancer are a; irregular ill-defined mass, spiculated mass, microlobulated mass, asymmetry, fine linear pleomorphic microcalcification, and architectural distortion.

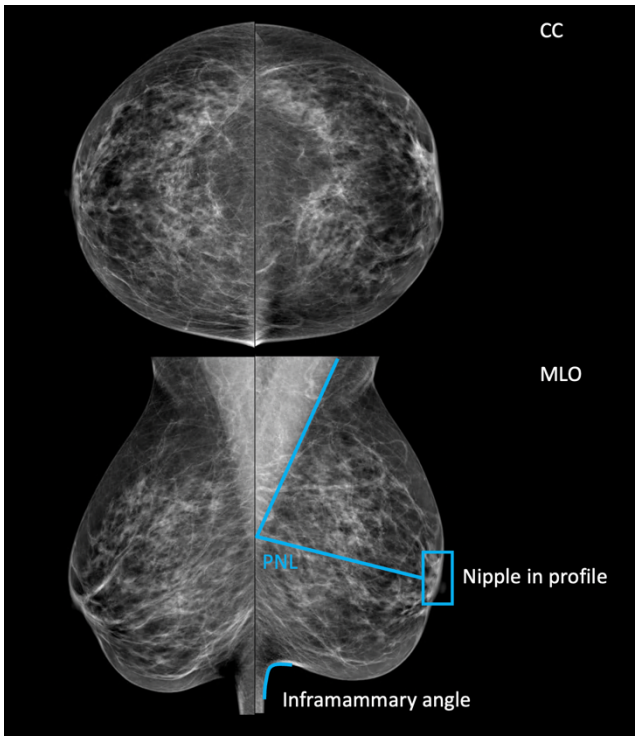


Figure 1-3 – Example of a two-view full field digital mammogram (FFDM). *Image quality markers are shown for the posterior nipple line, inframammary angle and the nipple in profile. CC: craniocaudal, MLO: mediolateral oblique, PNL: posterior nipple line.*

1.2.3 Mammographic breast density

Mammographic breast density is a radiographic representation of fibroglandular tissue (fibrous connective tissue (stroma), and glandular tissue (terminal ductal lobular units)) to fatty tissue proportion in the breast, where density is represented by areas that are radiopaque[38,65]. Variations in density between individuals occur due to genetic predisposition, ethnicity as well as due to changes in weight, nutrition, hormone exposure / expression, and age[66,67]. It is not only the amount and distribution, but also the heterogeneity, and texture of the fibroglandular tissue that is important[68]. Breast density can be measured from mammographic images using visual methods (ACR Breast Imaging-Reporting and Data System (BI-RADS) 5th edition lexicon (2013) / Visual Analogue Scale (VAS) / Wolfe classification (1976) / Boyd classification (1995) / Tabar classification (1997)), which are subjective, thus there is inter and intra-reader variability in reporting[68–71]. When using the BI-RADS lexicon the greatest discordance is reported in the middle two categories (b and c)[72,73]. The greatest proportion of the screening women's mammographic breast density is also represented in these middle two categories of BI-RADS breast density (b and c)[74,75]. Four different FFDM images demonstrating the four BI-RADS breast density categories are shown in Figure 1-4.

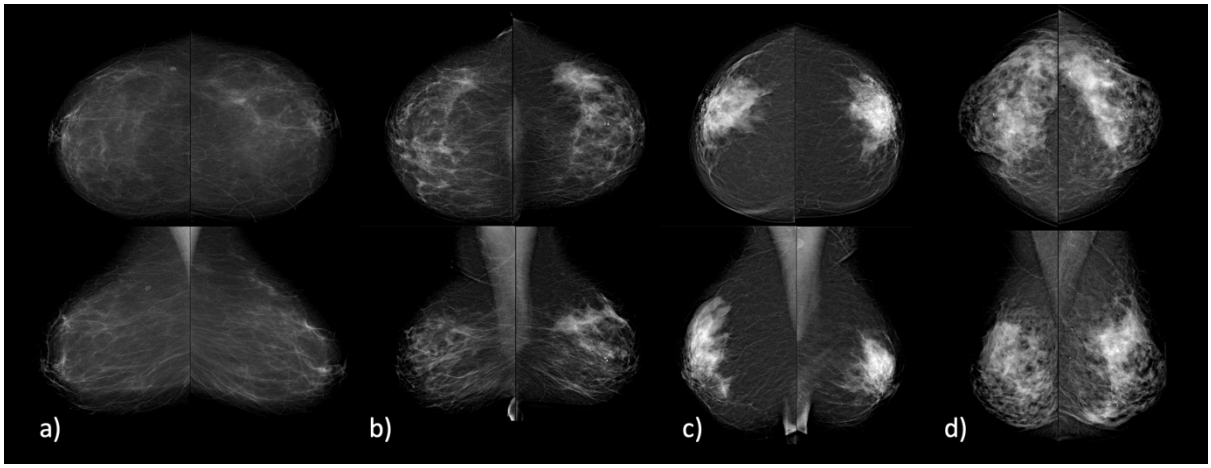


Figure 1-4 – Examples of breast imaging-reporting and data system (BI-RADS) 5th edition mammographic breast density categories. *a) Almost entirely fatty, b) scattered areas of fibroglandular density, c) heterogeneously dense, d) extremely dense.*

Alternatively, breast density can be measured using semi-automated or fully automated systems, which provide a more consistent output[76]. Planimetry, semi-automated thresholding techniques (Cumulus and Medena) and fully automated systems (Volpara™ and Quantra™) can be used with varying levels of human interaction to provide quantitative measures of mammographic breast density[68,77]. Fully automated systems, such as Volpara™ and Quantra™, provide density as either an area or volumetric breast density (VBD) measure. These quantitative measures can then be mapped into the BI-RADS density categories. Previous studies have shown variable agreement between automated systems and radiologists density assessment (κ = 0.46-0.57)[78]. In addition, these measurements can be affected by positioning, radiographic factors such as kVp (tube voltage) and mAs (tube current and exposure time) as well as the incorporation of nonstandard views[38,76]. Most automated systems require raw ("for processing") FFDM data, which is not kept routinely due to storage space requirements. A raw image is proportional to the x-ray attenuation detector signal. The raw FFDM is then processed to create "for presentation" images by the mammography vendors algorithm[68]. New deep learning (DL) density systems have started to use the processed FFDM images to calculate mammographic breast density[79–82]. Lehman *et al* demonstrated good agreement (κ = 0.85; 95% CI 0.84-0.86) and acceptance (90%) with radiologists when implementing a new DL density algorithm in clinical practice, reviewing > 10,000 mammograms[82,83]. When this model was then externally validated there was a high rate of agreement by both academic (94.9%) and community (90.7%) radiologists[84].

Increased breast density reduces the sensitivity of mammography as overlapping fibroglandular tissue can obscure the detection of a breast cancer (known as "masking"). Sensitivity reduces from 75.0-98.0% to 30.4-66.0% from the highest to lowest BI-RADS categories (a to d)[85,86]. Breast density

is also an independent risk factor for developing breast cancer[87,88]. A systematic review and meta-analysis of forty-two studies demonstrated the relative risk of developing breast cancer was 2.9 and 4.6 in women with a Percentage Mammographic Density (PMD) of 50-74% and ≥ 75% respectively, relative to women with PMD < 5%[87,89]. VAS (OR 4.4 (95% CI 2.7-7.0)), Densitas % (OR 2.17 (95% CI 1.41-3.33)), Volpara % (OR 2.42 (95% CI 1.56-3.78)) and BI-RADS (OR 2.3 (95% CI 1.9-2.8)) measures have shown to be strong predictors of breast cancer risk[78,79].

Legislation passed by the USA Congress in 2019 requires the mandatory reporting of density as part of the USA breast screening programme[90,91]. Women classified as having dense breasts (BI-RADS c or d) in USA screening are recommended to discuss with their doctor if they should undergo additional imaging, as a cancer could have potentially been obscured by the dense breast tissue[38]. In 2022 the European Society of Breast Imaging (EUSOBI) recommended that women "should be informed about their breast density" and that women aged 50-70 with "extremely dense breasts" should be offered screening breast MRI "every 2 to 4 years"[92].

1.2.4 Risk prediction

Opportunities for risk adapted screening by using different measures for risk stratification as well as different imaging modalities or screening frequencies for cancer detection are currently under investigation[38,93]. Methods for risk stratification include using breast density alone, which involves triaging the women in the highest density categories (c and d) for supplemental imaging with more sensitive imaging modalities (e.g. MRI or ultrasound)[94]. Alternatively mammographic breast density can be incorporated into in to risk prediction models (e.g. Tyrer-Cuzick (TC)) to increase the predictive power; TC + BI-RADS OR 1.55 (95% CI 1.33-1.80), TC + Volpara VBD OR 1.40 (95% CI 1.21-1.61) vs TC alone 1.27 (95% CI 1.14-1.40))[95,96]. Yala *et al* demonstrated their DL model (Mirai), which uses the mammographic imaging data only and no additional risk prediction fields, achieved a one-year cancer risk prediction C-index of 0.75-0.84 at seven separate sites in four continents when used alone compared to the TC C-index of 0.62, which was tested on data from only one USA site[93]. Furthermore, using the data from one USA site and thresholding Mirai at the TC specificity of 85.2%, Mirai achieved a sensitivity of 39.7% (95% CI 32.9%-46.5%) compared to TC which achieved a sensitivity of 22.9% (95% CI 15.9%-29.6%)[93]. However, Mirai was both developed and tested only on Hologic mammograms and so further generalisability testing using different mammographic vendors is required to account for variability of post-processing[93].

Polygenic risk scores can be calculated through sequencing a pre-defined panel of single nucleotide polymorphisms[97]. The incorporation of polygenic risk scores into risk prediction models (e.g. TC or Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA)) resulted in improved risk stratification accuracy (area under the receiver operating characteristic

curve (AUROC) 0.691 to 0.697)[98], however it can also lead to overestimation of risk in certain risk categories (e.g. high risk expected to observed number of cases (E/O) 1.54 (0.81-2.29)).
A collaborative approach using breast density, polygenic risk scores, and risk prediction models / DL models is likely to further increase the accuracy of screening risk statification[38,94,96,97,99]. With increasing accuracy the feasibility and cost-effectiveness of risk stratified screening also improves[38].

1.2.5 Interval cancers

Interval cancers (ICs) are defined as those occurring between the screening round ("a breast cancer diagnosed in the interval between scheduled screening episodes in women who have been screened and issued with a normal screening result")[38,100,101] . An estimated 6,000 ICs occur in the NHSBSP each year with the average time to diagnosis of 644 days, such that the highest proportions are diagnosed in the second (42.0%) and third years (39.0%) after screening[102,103]. Updated Public Health England (PHE) guidance in 2017, requires all ICs to be reviewed in order to ascertain whether or not a cancer had been missed at the original screen read[33]. A score is applied (1) normal / benign (77.0%), (2) uncertain (16.0%), and (3) suspicious (7.0%), with a 'duty of candour' requiring all patients to be informed of a suspicious finding[33]. ICs are often of higher grade and lesion size compared to screen detected cancers and therefore have a worse prognosis[103]. Increased mammographic breast density is associated with increased risk of IC development[101,104–106]. ICs are a key measure of the performance of a screening programme and the NHSBSP has set an acceptable IC rate target of 0.65/ per 1000 women screened in the first 12 months, 1.40/ per 1000 12-24 months and 1.65/ per 1000 24-36 months, which has increased to reflect the changes in incidence overtime[33,100,101]. Therefore, in total a rate of 3.7/ per 1000 ICs is expected in the NHSBSP.

1.3 Artificial intelligence in breast cancer screening

1.3.1 Introduction to artificial intelligence

Alan Turing first proposed the question, "Can machines think?", in 1950[107]. The term Artificial Intelligence (AI) is thought to have first been used in 1956 at the Dartmouth Summer Research Project on Artificial Intelligence[108]. Progression in the field of AI has recently accelerated due to the availability of large datasets, sufficient computing power as well as a growing interest and funding to develop algorithms to automate everyday tasks. AI is an umbrella term for Machine Learning (ML) and DL disciplines as show in Figure 1-5. AI is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception and decision-making. AI is strictly defined by ISO/IEC TR 24028:2020 as the "capability of an engineered system to acquire, process and apply knowledge and skills"[109,110]. ML is a 'sub-field' of AI, where algorithms learn and improve autonomously through the provision of data, without 'explicit programming'[111].

Examples of traditional ML techniques include; support vector machines, k nearest neighbours, principal component analysis, and decision trees. DL is a subset of ML that uses multiple algorithms working in a neural network architecture with many layers to extract high level features from data and carry out hierarchical learning[112].

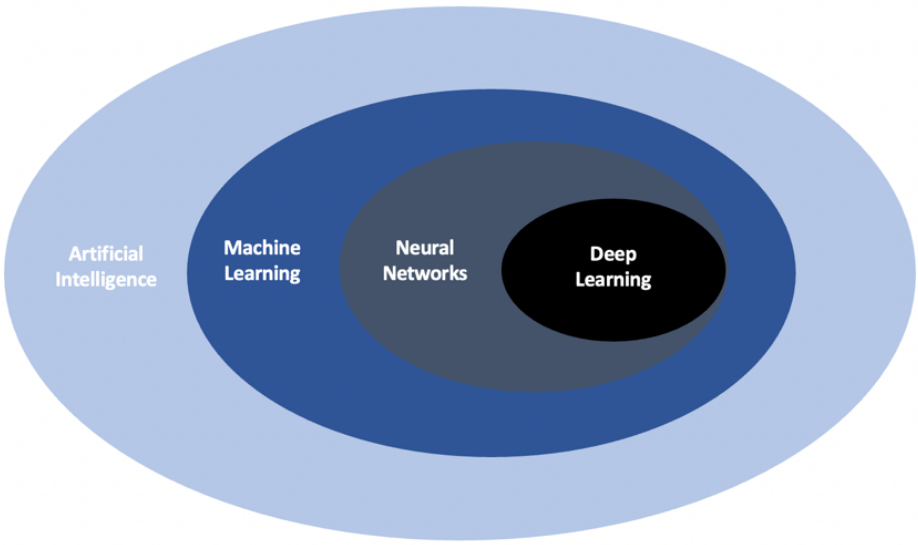


Figure 1-5 – Artificial intelligence (AI) hierarchy of terms. *Image used with permission of the National Breast Imaging Academy e-LfH programme.*

DL has been applied to computer vision tasks using Convolutional Neural Networks (CNNs) achieving good performance. The 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is seen as the catalyst for this innovation where the AlexNet algorithm reduced the error rate in an image recognition task to 15.3%, improving the error rate compared to the team that came second, whose achieved an error rate was 26.2%[113]. CNN algorithms are increasingly being applied to every day image recognition tasks such as self-driving cars, facial recognition, and automated text translation. A typical CNN consists of multiple layers. The first layer is the input layer. Next there are the convolutional, detector, pooling and fully connected layers, which are the hidden layers. Lastly, there is the output layer. In the input layer an image is provided to the algorithm. Next in the convolution layers a kernel passes over the image to extract high level features[111]. An activation function can then be applied in the detector layer, such as a rectified linear unit (ReLU) where all negative values in the feature map are replaced with a zero-value adding nonlinearity to the data. The feature map is then passed to the pooling operation (e.g. max pooling, sum pooling or average pooling) for feature reduction / down sampling, providing translation invariance. This map is then passed to the fully connected layer where a classifier function (e.g. softmax activation function) is applied. There can be hundreds of hidden layers in a CNN[112,114]. The output from the algorithm is then provided in the output layer as a classification result. Iterative adjustments of the algorithm

take place through a loss error function and back-propagation processes in training. An overview of a CNN architecture is shown in Figure 1-6.

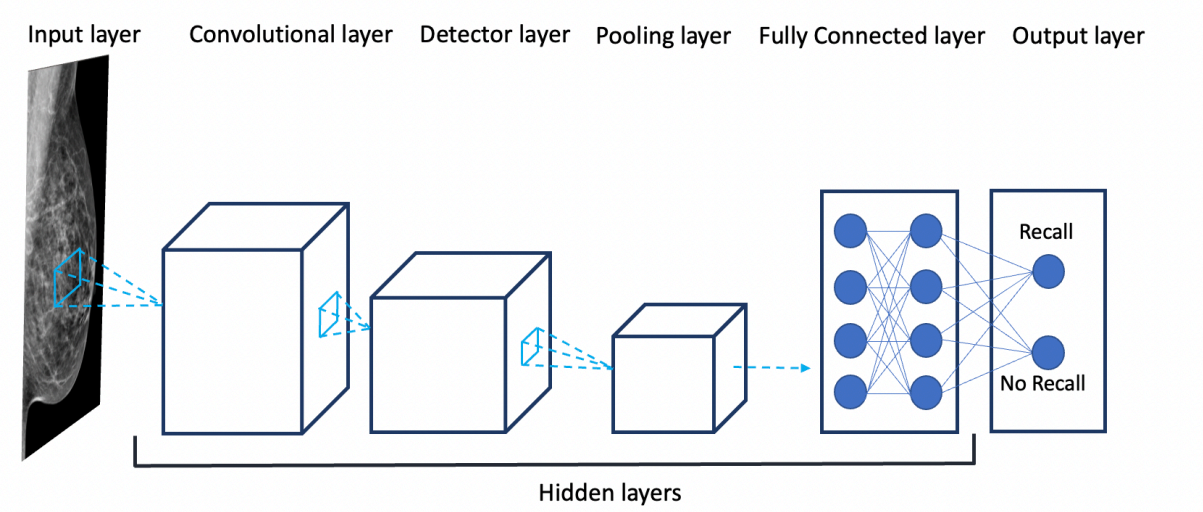


Figure 1-6 – Overview of the architecture of a Convolutional Neural Network (CNN). *Adapted from National Breast Imaging Academy – Computer-Aided Detection (CAD) and Artificial Intelligence (AI) module.*

Radiology is a digitally advanced field of medicine as well as being a mainly visual-based specialty, with ~45 million radiological images reported each year in England (the most common of which is plain film x-rays with ~1.86 million reported annually)[115]. Thus, radiological image interpretation is a prime candidate for the application of DL to aid in automating radiology tasks. CNNs have been used in detection, diagnosis, and segmentation-based tasks in radiology[111]. Medical images, such as mammograms, do differ from everyday images. They are of higher resolution, for example mammograms contain between 2600 x 2000 pixels, and the area to be detected (disease) in medical images is a relatively small area of the total image. Moreover, mammograms are more complex than natural images due to the high variability in patterns, the difference in features and task requirements, making this a challenging task[116,117]. Imaging data for training is becoming increasingly available, such as the ImageNet database which contains over 15 million labelled natural images, however there is still a limitation in the availability of ethically approved curated medical image datasets[118]. To overcome this limitation algorithms can first be pre-trained on datasets such as the ImageNet, or other publicly available imaging datasets, and then re-trained on representative medical imaging data through transfer learning. Algorithm training can take place by supervised, semi-supervised, or unsupervised approaches requiring varying levels of data annotation. In supervised learning, detailed lesion level annotations and labels are provided, whereas in unsupervised learning no annotations or labels are provided and the algorithm itself identifies the pertinent image features from which to classify the image. Semi-supervised learning provides a

hybrid approach[111,114]. Such labels also provide the "ground truth" when testing algorithms performance. This ground truth is seen as the absolute outcome of a case, and can consist of expert radiologist annotation, time follow-up or histopathological outcome.

van Leeuwen *et al* reported that there are over 100 Conformité Européenne (CE) marked AI products for a radiological application in 2021, however only 36 had peer-review evidence, and of the available publications 49% of studies were performed "independently from the vendor"[109]. Kim *et al* reported that only 6% of published studies relating to the evaluation of AI algorithms were performed by external validation, such that the AI was tested on data from a separate institution (geographical) or time period (temporal) from the training data[119]. Thus, further unbiased evidence provided from large external studies conducted independently of the commercial vendor are needed across the radiology AI field.

1.3.2 History of computer aided detection systems in breast cancer screening

Research into the use of Computer Aided Detection (CAD), also known as CADe systems, in medical imaging commenced in the 1960s, with the first CAD mammography system (Hologic R2 (Image Checker M1000)) receiving FDA clearance in 1998[120]. Traditional CAD systems, based on hand-crafted features, provided prompts for radiologist such as a Δ symbol for calcification and a $*$ symbol for mass to mark areas of increased suspicion in order to reduce reader oversight, acting as a "second-look". These initial systems had high sensitivity for calcifications (98.0-100%) and could also detect masses (88.0-92.0% sensitivity), but few could detect features of asymmetry or distortion[121,122]. A high number of false positive prompts, due to low specificity, resulted in reader fatigue, distraction, and loss in confidence of CAD systems. Thus, overall performance when using a CAD system is dependent on the decision-making process of the reader, the accuracy of the system, and the interaction between the two[123,124]. There is also the possibility of over reliance on the system leading to a loss in synergy between the computer system and human reader required to maintain a high level of sensitivity, as lesions could be overlooked if not marked[125]. In 2008 between 74.0-91.4% of USA mammograms were read using a CAD based system in conjunction with a single reader, and an updated survey in 2016 found 92.3% of screening centres used CAD systems[126,127]. However, the adoption of CAD systems has been limited across other screening programmes in the world, including the NHSBSP, due to the effect of increasing recall rates and thus deemed lack of cost effectiveness[128]. Lehman *et al* investigated the use of CAD in the USA screening system, Breast Cancer Surveillance Consortium, between 2003 and 2009 and demonstrated a reduction in sensitivity when using CAD, from 87.3% without CAD to 85.3% with CAD as well as an increase in recall rate when 495,818 mammograms were interpreted with CAD and 129,807 without CAD, by 271 radiologists at 66 facilities[125]. A pooled analysis of ten studies (2001-2008), Taylor *et al* (2008),

demonstrated that single reading plus CAD vs single reading resulted in a significant increase in recall rates (OR 1.10, 95% CI 1.09 - 1.12, P < 0.001), and that double reading with consensus / arbitration resulted in a significant improvement in recall rates compared to single reading plus CAD[129]. CAD systems demonstrated a varied performance for cancer detection, and in the pooled analysis as part of the review, no statistically significant difference in cancer detection rate was found (OR 1.04, 95% CI 0.96-1.13, p = 0.35)[129]. CAD systems also face the same reduction in sensitivity due to increased mammographic density that human readers incur and CAD prompts have been shown to increase overall reading time by approximately 10-20 seconds[130–132].

1.3.3 Deep Learning applications to breast cancer screening

The traditional CAD systems are now being superseded by DL CAD algorithms which have improved sensitivity and specificity for the detection of cancers[133]. The majority of DL algorithms with Federal Drug Agency (FDA) or CE mark approval are for clinical decision support system (CDSS) applications, similar to traditional CAD systems where the algorithm supports the reader by providing prompt suggestion to locate the cancer. However, there are multiple other applications of the latest DL CAD systems for mammography interpretation to aid readers and improve the efficiency and efficacy of breast screening, which are shown in Figure 1-7. These include the use of DL CAD systems as stand-alone readers for DL CAD triage (CADt), to prioritise work lists and pre-populate image reports for normal studies to improve programme efficiency. Studies have shown that these DL CADt systems can operate as stand-alone readers to both rule out a high proportion of normal cases (17.0%-60.0%) whilst missing a small proportion of cancer cases (0.0-7.0%), as well as rule in a small proportion of cases (1.0-5.0%) highly suspicious of cancer (13.0%-32.0% next round (NRCs) and ICs) for further assessment[134,135]. In addition, DL CAD detection and diagnosis (CADe+x) algorithms could operate as stand-alone systems to replace a reader in a double reading system. This approach has the potential to improve both efficiency as well as efficacy, as these systems could replace one reader as well as reduce the rate of ICs. A study by Lång *et al* found 11.2% of ICs could be detected at the screening mammogram when the DL CADe+x algorithm was set at a 4.0% recall rate. As outlined earlier in this chapter (section 1.2.3 and 1.2.4), DL density algorithms could potentially be used to risk stratify the population for adapted screening and the application of supplemental imaging. These DL algorithms could either be used alone or in combination with other risk information, such as mammographic breast density, polygenic risk scores and other risk prediction models.

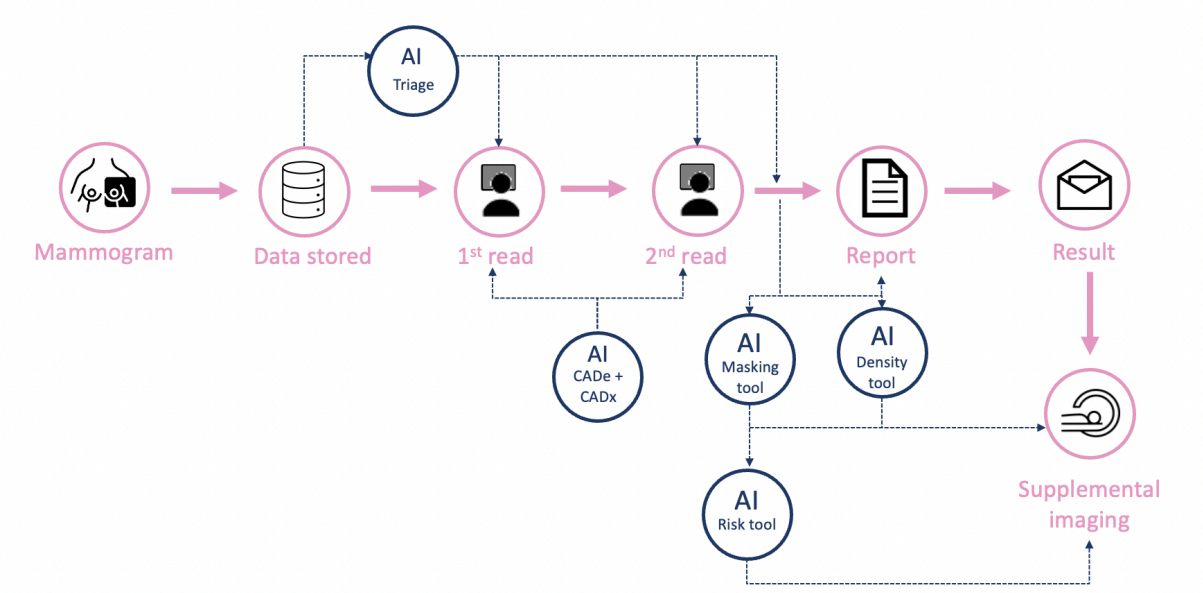


Figure 1-7 – Application of Deep Learning (DL) Computer Aided Detection (CAD) algorithms to breast cancer screening. *Image used with permission of the National Breast Imaging Academy e-LfH programme.*

The UK National Screening Committee (NSC) report 2021 on the "Use of artificial intelligence for mammographic image analysis in breast cancer screening – Rapid review and evidence map" did "not recommend using AI in the NHSBSP"[136]. This was due to: a lack of evidence relating the accuracy of AI in clinical practice, the varying reported performance of AI in different settings, the lack of UK based evidence, lack of quality evidence, and lack of evidence pertaining to AI performance for different types of breast cancer as well as performance in different groups of women (e.g. "different ethnic groups")[136]. All studies included in the report were of high risk of bias using the QUADAS-2 tool for assessment and the authors recommended the importance of external validation using geographically different datasets as well as pre-specified test thresholds to limit bias. All evidence provided in the report was retrospective and the report highlighted that a number of studies used enriched cancer cohorts for testing. Enriched cohorts consist of an increased cancer proportion which is "atypical of a screening population", this was defined in the report as a cancer percentage of more than 3.0%. A repeat review was recommended for 1-3 years' time from the date of publication to review the latest evidence.

The 2016 "Digital Mammography Dialogue on Reverse Engineering Assessment and Methods" (DM DREAM) challenge, although not conducted using UK data, did provide an external validation study on a large representative cohort, from two different countries and screening programmes, for the testing of multiple DL algorithms[137]. The DREAM challenge included 126 different teams, who each developed their own DL CADe+x algorithm for the prediction of cancer development within the next 12 months. The curated datasets consisted of 144,231 and 166,578 screening FFDMs, with prior

exams and clinical information available, from a USA (Kaiser Permanente Washington (KPW)) and Swedish (Karolinska Institute (KI)) screening site respectively[137]. 1.1% of the mammograms in the KPW and KI datasets were cancers and no image level annotations were available. Each image was assigned a binary image label ground truth from histopathology results (cancer) or follow-up of ≥ 12 months (normal)[133]. The KPW dataset as well as other publicly available datasets were used for model training. The top twenty algorithms were evaluated using the KI dataset. The top performing algorithm on the KPW dataset also achieved the top performance on the KI dataset demonstrating the generalisability of this algorithm. No challenge algorithm outperformed reader performance when the threshold was set at single reader sensitivity (77.1%), with the top performing algorithm achieving a specificity of 88.0% compared to the single reader specificity of 96.7%. Hybridisation of the best performing algorithms into the challenge ensemble method, achieved a specificity of 92.5% (at the single reader sensitivity (77.1%))[137].

It is important to identify the limitations of a DL CAD algorithms to enable the radiologist to be vigilant prior to implementation in clinical workflow. The latest systems are still susceptible to the limitations that both traditional CAD and human readers face, such as the reduction in sensitivity with increasing breast density[138]. The true impact of these latest systems on reading time, recall rates, biopsy rates, cost effectiveness and cancer detection are unknown and prospective evaluations are required to fully assess the clinical impact of DL CAD algorithms on breast cancer screening targets. It is also pertinent that DL CAD algorithms do not exacerbate the potential harms of screening such as overdiagnosis, overtreatment and false positive recalls to assessment which lead to patient anxiety. Lastly, the gaps in evidence highlighted by the UK NSC report 2021 are required to be addressed before DL CAD algorithms are introduced into the NHSBSP.

1.4 Thesis aims and outline

The focus of this thesis is the evaluation of AI algorithms (specifically DL CAD algorithms) for breast cancer screening applications. The developments in DL means that AI algorithms have been created for numerous mammography screen reading tasks. However, more evidence is needed to determine the best way to evaluate and monitor these AI algorithms to ensure acceptable performance for deployment into breast screening programmes as well as which applications of AI algorithms are most suitable for breast cancer screening in different countries. In this thesis, three different commercial AI algorithms are evaluated using a large retrospective dataset from two NHSBSP sites for three different proposed AI algorithm applications in breast cancer screening. For consistency each AI algorithm is assigned a unique identifier (DL-ID) for this thesis which remains consistent throughout all chapters.

The main aims of this thesis are:

1. To investigate the performance of AI algorithms for stand-alone reader applications in breast cancer screening, through a systematic review and meta-analysis, to determine the current performance achieved, the datasets used in testing, as well as gaps in reported evidence.
2. To develop a representative UK screening mammographic imaging database that can be used for retrospective benchmark testing of AI algorithms.
3. To investigate the performance of three AI algorithms for the detection of ICs in breast cancer screening.
4. To benchmark the performance of three AI algorithms to be used as a stand-alone reader as well as in collaboration with human readers for breast cancer screening.
5. To evaluate the performance of three AI algorithms to triage low priority cases that do not require human reading as well as high priority cases that can bypass reading to enhanced assessment, whilst maintaining an acceptable sensitivity and specificity.

Chapter 1 provided an overview of breast cancer and breast cancer screening as well as the main image screening technique of mammography. In addition, the history of CAD systems in breast cancer screening programmes and the advances in DL methods that underpin the latest AI algorithm approaches to breast cancer screening were discussed.

Chapter 2 covers the ethical, legal and regulatory challenges surrounding the use of AI algorithms in breast cancer screening and the development of medical imaging databases required to evaluate AI algorithm performance.

Chapter 3 is a systematic review and meta-analysis of the stand-alone applications of AI algorithms in breast cancer screening. The diagnostic performances of different AI algorithms are compared and the databases used for testing, study methodology and reporting quality are evaluated.

Chapter 4 outlines the construction and contents of a mammographic medical imaging database which is subsequently used in Chapters 5-7 for testing the performance of different AI algorithms.

Chapter 5 presents the results from an experiment to evaluate the performance of AI algorithms for IC detection. This chapter also investigates the use of different thresholding methods to identify the operating point for each AI algorithm.

Chapter 6 details the results from an experiment to investigate the performance of AI algorithms for detection and diagnosis as a stand-alone reader compared to human reader performance. It also evaluates the sub-groups of cancers detected by each AI algorithm.

Chapter 7 describes the results from an experiment to use AI algorithms in a triage-based task for both normal and highly suspicious case identification. In addition, the cancers missed by each algorithm are reviewed.

Chapter 8 summarises the research presented in this thesis and its implications. This is followed by a section on the recommended direction of future work.

Chapter 2 – Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations

This chapter explores how artificial intelligence (AI) is being applied and evaluated in breast imaging. Key ethical and legal challenges at the algorithm, data and clinical levels which need to be considered for the implementation of AI in everyday breast screening are discussed. The barriers and limitations currently facing this field from a technical, clinical and governance perspective are also outlined.

Contents of this chapter have been published in British Journal of Cancer[139].

2.1 Introduction

In breast oncology, a multidisciplinary team approach is essential, with imaging playing a key role in the care pathway for the screening, diagnosis, staging, monitoring and follow-up of malignancies. Novel imaging techniques of increasing complexity have resulted in longer reporting times. This, coupled with a shortage of radiologists and exponential growth in imaging requests, has led to an increasing demand on radiology departments. Recently, there has been a huge interest in using Artificial Intelligence (AI) for breast imaging to address these pressures, in a speciality where timing is critical and resources are finite[140].

The term AI covers both machine learning and deep learning[141]. It is the advances in deep learning for image interpretation that have resulted in the massive growth in interest for use in breast imaging[112]. AI applications can be broken down into two categories, Figure 2-1.

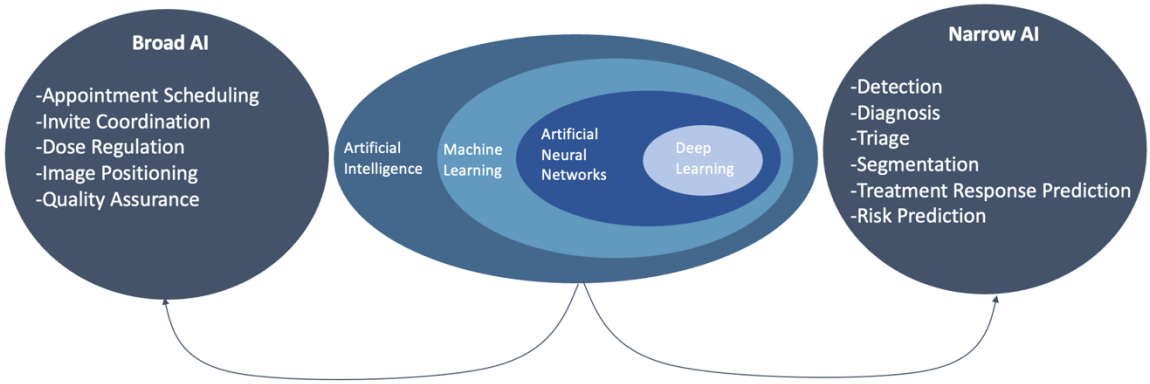


Figure 2-1 – Broad and narrow artificial intelligence (AI) applications to breast imaging.

The first category is “broad AI”, which lends itself to the administrative and organisational tasks within the imaging pathway. These systems can be used to replace repetitive and routine tasks such as appointment booking, contrast adjustment and image quality checks. The second category is

"narrow AI", which covers computer-aided detection (CADe), diagnosis (CADx), and triaging worklists (CADt) as well as predicting treatment response and segmenting lesions[112]. These AI systems can be used as aids for clinicians or be used autonomously. Ultimately these AI solutions aim to improve the patient's outcomes as well as the healthcare system's efficiency. The latest advances in computer processing and the increased availability of data have been pivotal for developing AI-CAD (CADe and CADx) systems[142,138].

It is important to remain vigilant to the potential bias and ethical questions that arise when using this technology as well as the challenges of incorporating such systems into pre-existing workflows[143,144]. These overarching challenges need to be explored in order to facilitate discussion and drive engagement by clinicians, computer scientists, responsible national agencies and National Health Service (NHS) Trusts[145].

This article reviews how AI has been applied and evaluated using breast imaging as an exemplar. We then consider the ethical and legal challenges at the algorithm, data and clinical levels. Lastly, we discuss the barriers and limitations currently facing this field from a technical, clinical and governance perspective.

2.2 Evaluation of artificial intelligence in breast imaging

2.2.1 Retrospective evaluation

Retrospective testing on internal or external datasets is essential when assessing new AI tools for clinical imaging[142,146]. An algorithm is often trained and tested on an internal dataset which has been divided into an 80:20 split[146]. This means that the training data is not used to test the algorithm otherwise this would result in bias and an overestimation in performance[147]. Ideally external datasets consisting of new unseen data which has not been used for algorithm development are used to ascertain the generalisability of an algorithm in different populations with images from different manufacturers (see Ethical and legal constraints – Algorithm level for more information)[146,148]. It is also important to distinguish between testing that is conducted internally (by the AI developers) and externally (by an independent institution). External testing can limit bias and also allow for the comparison of multiple algorithms with similar applications[149].

Data that is representative of the population, structured, annotated and ready to use is limited, existing in only a small number of institutions, Table 2-1[150]. New imaging portals and repositories, such as the Health Data Research Innovation Gateway, have been set-up to try to address these data gaps and are key to developing a data ecosystem to meet the demand[151]. Principles such as FAIR (Findability, Accessibility, Interoperability, and Reusability), aim to guide data extraction as well as long-term management and sharing, in order to obtain the "maximum benefit" from datasets[152].

However, a balance must be found in this ecosystem between the implementation of FAIR principles and the often-strict controls put in place by Information Governance teams and ethics committees when creating imaging repositories.

Dataset	Country	Year of studies	Modality	Number cases	Number images
The Mammographic Image Analysis Society Digital Mammogram Database (MIAS)[153]	UK	1994	SF-MG	161	322
Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM)[154]	USA	1999 (updated 2016)	SF-MG	1566	10239
Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and moLecular Analysis (ISPY1 (ACRIN 6657))[155]	USA	2002-2006	MRI	222	386528
InBreast[156]	Portugal	2008-2010	FFDM	115	410
Cohort of Screen-Aged Women (CSAW)[157]	Sweden	2008-2015	FFDM	499807	>2000000
The OPTIMAM Mammography Image Database (OMI-DB)[150]	UK	2010-2019	FFDM	151403	>2000000
New York University Breast Cancer Screening Dataset (NYU BCSD v1.0)[158]	USA	2010-2017	FFDM	141473	1001093
Breast Cancer Digital Repository (BCDR)[159]	Portugal	NA	SF-MG FFDM	1010724	3703 3612
The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA)[160]	USA	NA	MRI MG	139	230167

Table 2-1 – Datasets publicly and privately available for breast imaging. *FFDM: Full Field Digital Mammography, MG: Mammography, MRI: Magnetic Resonance Imaging, NA: Not Available, SF: Screen Film.*

The performance of an algorithm can be compared against two outcomes, 1) the ground truth and 2) the radiologist's performance[146,147]. The ground truth or "gold standard" is seen as the 'absolute' outcome of a case (for example cancer or no cancer) but variations of the ground truth between healthcare systems occur due to differences in standard of care guidelines, histopathology reporting criteria, imaging procedures conducted (e.g. use of Magnetic Resonance Imaging (MRI) versus ultrasound) and screening frequency (e.g. range from 12-36 months). The radiologist's performance sets a "clinically relevant threshold" for AI performance to be compared against and is essential to understand the potential impact of using such systems in real-time workflows (for example double reading in the UK breast screening programme)[148,161,162]. However, in screening when using the

radiologists assessment as the gold standard, there is potential to introduce bias in favour of the radiologist, where only those patients recalled by the radiologist can be diagnosed by the AI. When trying to prove the superior performance of AI compared to radiologists, interval cancers need to be included in testing sets. Experienced radiologists' reports should also be included to allow for the comparison against representative programme reader performance, and not just prove that the AI is superior to average or non-specialist performance. Algorithms need to meet or exceed these thresholds in order to show a potential benefit before their adoption into healthcare systems is considered.

2.2.2 Prospective evaluation

Whilst testing on retrospective datasets provides a 'snapshot' of possible performance, the nuances of medical pathways cannot be underestimated. Prospective testing in real-time is essential to fully understand the influence of AI on human performance and the interaction between the two[142]. There are few prospective studies on the use of AI in radiology, Table 2-2, with a recent systematic review only reporting one randomised trial registration and two non-randomised prospective studies in radiology[163].

AI	Country	Imaging modality	Stage of care pathway	Estimated completion	Trial ID (ClinicalTrials.gov)
Samsung (Seoul, South Korea) S-Detect™	China	Ultrasound	Diagnosis	February 2020	NCT03887598
Unknown	China	Mammography	Detection & Diagnosis	November 2020	NCT03708978
Unknown	Russia	Mammography (+ others)	Detection	December 2020	NCT04489992
Unknown	China	ABUS	Screening	August 2025	NCT04527510
Kheiron (London, UK) Mia™	UK	Mammography	Screening	Unknown	Unknown – part of the AI Award[144,164]

Table 2-2 – Prospective studies for the use of artificial intelligence (AI) in breast imaging. *ABUS: Automated Breast Ultrasound.*

To ensure the clarity of reporting results from these studies, pre-existing reporting standards have been adapted and include the Consolidated Standards of Reporting Trials-AI (CONSORT-AI), Standard Protocol Items: Recommendations for Interventional Trials-AI (SPIRIT-AI) and the Checklist for Artificial Intelligence in Medical Imaging (CLAIM)[165–167]. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis-Machine Learning (TRIPOD-ML) and

Standards for Reporting Diagnostic Accuracy Studies–AI (STARD-AI) are also currently under development, Table 2-3[168,169].

	Publication date	Application	Number of items	Link
CONSORT -AI[165]	2020	Randomised trials	25 original 14 new	https://www.equator-network.org/reporting-guidelines/consort-artificial-intelligence/
SPIRIT -AI[166]	2020	Clinical trial protocols	51 original 15 new	https://www.equator-network.org/reporting-guidelines/spirit-artificial-intelligence/
CLAIM[167]	2020	AI studies in radiology	42	https://pubs.rsna.org/doi/full/10.1148/ryai.2020200029
TRIPOD -ML[168]	Pending	Clinical prediction model evaluation	-	https://www.tripod-statement.org
STARD -AI[169]	Pending	Diagnostic accuracy studies	-	-

Table 2-3 – Reporting criteria adapted for artificial intelligence (AI) studies.

Performance of AI is often measured in terms of sensitivity, specificity, area under the reciver operating characteristic curve (AUROC) and computation time (time taken to process data). Where AI is used by a radiologist, the effect on performance is measured in the same way (sensitivity, specificity, and AUROC) with the additional measure of reading time by the radiologist. The AUROC provides a summary estimate of diagnostic accuracy, taking into account both the sensitivity and specificity to demonstrate how well the algorithm can differentiate between cancer or not cancer across all thresholds[170]. It provides a measure between 0 and 1, where a higher score means a better classification[146]. However, the AUROC is subject to certain pitfalls. It is not "intuitive" to interpret clinically, and theoretically algorithms with different sensitivities and specificities can have the same AUROC[170]. Therefore, alternative measures such as "net benefit" have been proposed as well as routine reporting of sensitivity and specificity, which allow for direct clinical comparison[170]. Lastly, for both the algorithm alone and when used by the clinician, the effect on nationally reported standards (e.g. cancer detection rate, recall rates, tumour size and lymph node status) should be evaluated as part of prospective studies[146].

2.2.3 Key considerations for clinical evaluation

Screening AI systems could be cost-effective by improving early detection of important "killer" cancers (higher grade) potentially improving long term survival. However, the substantial investment of AI development, IT infrastructure, and continuous monitoring need to be costed, therefore cost-effectiveness requires careful evaluation[147,171]. The ease of integrating AI into pre-existing hospital

systems, such as radiology information systems and Picture Archiving and Communications Systems (PACS), health-records and administrative systems, is a another key consideration[162,171]. Wider measures for clinical evaluation to also include are patient acceptability and effect on uptake of screening programmes as well as the training required for radiologists to be able to use and interpret AI tools[54].

Continuous monitoring to ensure adherence to national standards needs to be in place to observe both static and adaptive ('learning on the fly') AI when used in real-time workflows (see Ethical and legal constraints – Algorithm level for more information)[162,172]. Each hospital could have an infrastructure to evaluate and monitor algorithms, but this is unlikely to be feasible in many hospitals due to the data storage requirements and lack of technical expertise and resources to set up such an environment. A centralised testing system at designated centres using pre-set national standard thresholds for different AI algorithm applications, would be a more sustainable approach.

As outlined above, the steps in the evaluation pathway of AI are clear, requiring retrospective, prospective and continuous real-time testing. However, the caveats of testing such as how to access suitable datasets and defining "clinically relevant thresholds" still need to be agreed. In the UK NHSX has set-up 'AI Labs' to begin conducting centralised and standardised testing procedures[173,174].

2.3 The breast imaging pathway and AI

2.3.1 Screening

AI has been used in radiology since the 1990's with initial CADe tools in mammographic screening prompting readers to look again at areas of concern in the image[128]. More recent AI systems can now meet and exceed the performance of radiologists for stand-alone cancer detection in screening mammography, achieving a sensitivity from 0.562 to 0.819 with a specificity of 0.843 to 0.966 (set at first reader specificity)[138,149]. However, this is not the case for all national screening programmes[137].

In a retrospective international crowdsource competition, the performance of multiple algorithms was compared on a standardised test set from Sweden. An ensemble algorithm was built by concatenating the eight best individual performing algorithms, which was shown to outperform the top single algorithm, but not the clinicians performance[137].

In the UK 2.2 million mammograms are taken each year and read by two radiologists, putting a high demand on an already stretched workforce[54,140]. The majority of screening mammograms are normal[54,175]. A more efficient method is sought whilst maintaining current cancer detection and recall standards. AI can now reliably triage 'normal' mammograms (47% to 60%), which would mean that these would not need to be reviewed by two or possibly even one radiologist[134,135]. Whilst estimated to only miss up to 7% of cancers, the CADt algorithms could drastically improve the

efficiency of breast screening. However, questions remain around what an acceptable miss rate would be for algorithms when used in routine screening.

AI tools previously used for mammography have been adapted for other screen imaging techniques such as Digital Breast Tomosynthesis (DBT), which has longer reading times that can be decreased by around 50% using AI[176]. MRI is used for the screening of high-risk women, particularly those with a familial risk of breast cancer or BRCA1/BRCA2 carriers. Deep learning algorithms can find visual patterns in images and have been used to detect and diagnose breast cancer to produce a fully automated MRI AI-CAD system[177–179].

2.3.2 Risk stratification

Screening can be tailored according to a woman's breast cancer risk. Risk factors for developing breast cancer include breast density, family history, lifestyle factors (e.g., alcohol and smoking), genetic mutations, hormone exposure and expression[180,181]. Breast cancers can also go undetected due to dense breast tissue obscuring the view of a cancer on a mammogram, called 'Masking'[68]. AI density measures can provide quantitative scores or category scores such as BI-RADS, which can provide a more consistent interpretation than a radiologist[68,182]. It may be possible for the latest density tools to detect women who are at the highest risk of 'masking' and more likely to develop a cancer that could progress to later stage disease[68]. Automated breast density can also be incorporated into existing prediction models (BOADICEA and Tyrer-Cuzick) to improve performance and assist in the implementation of targeted screening as well as the use of supplemental imaging[182]. The 'Measurement Challenge' aims to compare automated density measures which have been shown to overcome the inconsistencies in human reporting as well as being able to predict breast cancer risk[183].

2.3.3 Monitoring and prognostication

MRI is routinely used in the monitoring of response to neoadjuvant chemotherapy, with patients imaged before, during, and after treatment. Deep learning algorithms have been implemented to evaluate pathological complete response to chemotherapy using post-treatment MRI with an AUROC of 0.98[184], which could affect the extent of post-treatment surgery, or potentially reduce the need for surgical excision at all. A number of studies have used deep learning to identify features from pre-treatment MRI that are predictive of response in an unsupervised fashion[185–187]. Early prediction of response to different types of chemotherapy could avoid unnecessary toxicity and cost from ineffective treatment as well as enable a more personalised approach to treatment. AI has also been used in prognostication to predict recurrence (Oncotype DX recurrence score) from MRI[188]. However, given the moderate accuracy of these techniques (0.77-0.93), further work is required before their integration into clinical practice.

The evidence base for the performance and possible applications of AI to breast imaging is rapidly evolving. Systems acting as stand-alone readers show promise in decreasing workload, whilst systems to predict treatment response could guide tailored treatment strategies. In addition, systems to identify those at greatest risk of a cancer being missed or developing cancer may aid in the application of a targeted screening approach.

2.4 Ethical and legal constraints

2.4.1 Guidance level

The Department for Health and Social Care, and international collaborations such as the Global Partnership on Artificial Intelligence, have developed guidance for implementing digital technology including AI[189]. They highlight the need for oversight and continued patient involvement to guide the development of "human-centric" AI which is essential to maintain the trust of the public, and avoid a repeat of previous controversies such as inappropriate data sharing[190–192].

2.4.2 Algorithm level

There is a danger of innate latent bias built into certain systems, especially if these have been developed on datasets that underrepresent certain populations (with a lack of diversity in age, ethnicity and socioeconomic background) and therefore lack the ability to generalise[193]. This could be further compounded by the limited diversity within the scientific workforce itself which under represents the "interests and needs of the population as a whole"[194]. Outcomes based on pre-existing inequalities could be exacerbated by the skewed outcome being fed back into the algorithm, creating negative reinforcement, thus limiting the fairness of an algorithm[193]. This can lead to algorithmic decisions that amplify discrimination and health inequalities. The data used in testing should therefore encompass a representative relevant population and the components of the dataset used explicitly reported alongside the results. A recent paper provides an example of such documentation, where an AI-CAD mammography algorithm trained on data from South Korea, USA and UK primarily using data from GE machines, achieved the best performance compared to other algorithms (sensitivity (81.9%) at the reader specificity (96.6%)), when tested on data from Sweden on only Hologic machines, demonstrating generalisability[149]. Algorithms also have the ability to 'learn on the fly', that over time become more biased due to 'performance drift', thus potentially limiting their generalisability[172,194]. 'Learning on the fly' could potentially be beneficial to adjust algorithms to the local systems in which they are being used but this will also require close observation through regular audits to monitor for detrimental 'performance drift'[147,162].
Transparency around how an algorithm reaches a decision, its architecture and source code availability is often limited by intellectual property clauses to protect proprietary information[174]. The

opaqueness of an algorithm's deduction can be clarified by using saliency maps, which highlight (e.g. heatmap) the part of the image which the algorithm has used to make its decision, ensuring that the algorithm is using at the correct part of the image to make its clinical deduction and not "noise" in the image such as a clip, artifact or label[195]. Initial checks built into the algorithm, ensuring the image is of sufficient technical quality from which to deduce an interpretation similar to the checks performed by radiologists, is also an important step for robust interpretation. A reliable algorithm providing consistent, clear and reproducible results, so as not to cause ambiguity in decision making, is key to improving confidence in these systems.

2.4.3 Who controls the data?

In the UK there is an understanding that NHS Trusts will govern, control and use patient data in an anonymised format to conduct research for patient benefit[143,196]. There is also an understanding that patient data will be protected and overseen by Information Governance teams at NHS Trusts[144,173]. Extracting data from the fragmented silos of the NHS remains a challenging task due to the lack of interoperability between systems[197]. Data relating to an individual's health is defined as 'special category' data and requires additional procedures and safeguards including data minimisation, proportionality, and necessity[198–200]. Data from which an individual can be recognised is termed Personal Identifiable Data (PID). This data is often pseudonymised or de-identified for healthcare research to remove identifiers and replace them with a new random identifier (e.g., Trial ID), ensuring privacy is upheld[201].

Where consent from individuals for data use cannot be feasibly obtained, provisions are in place to obtain access to PID in order to create large datasets[202]. Regulation has emphasised the importance of Patient and Public Involvement (PPI) when using patient data for research, especially in the context of unconsented data use[202]. Feedback provided by PPI can be used to enhance the communication between the public and healthcare sector, particularly around the distribution of a data notification and objection mechanism[174,202]. Studies carried out by organisations such as the Welcome Trust show that the public acknowledge a lack of understanding and hesitancy regarding the uses of health data, particularly when data is shared with and accessed by commercial companies[203]. National data opt-outs, proposed as part of the Caldicott Review (2016), give patients the option for their data to not be processed[204]. Recently the National Data Guardian opened a consultation to revisit the seven Caldicott principles that guide the use of PID and to ensure that public 'expectations' should be considered when using confidential information[205]. However, additional steps need to be taken to inform and educate the public around data use in healthcare so they can be empowered to explore these options.

The expected economic trade-off within the NHS in terms of financial payment, shareholding position or fees for product procurement should be outlined as part of a national policies. Allowing for the potential benefits from sharing valuable NHS data when collaborating with the commercial sector to be realised[147,189]. It is important to ensure this benefit is fairly distributed across the whole of the NHS to avoid widening gaps in available resources at different Trusts[145,197].

Linked data across multiple fields such as imaging, genetic and clinical records are of increasing importance for the development of risk prediction models for both prognosis and treatment response. Higher accuracy has been achieved by algorithms when multiple data types are used in training to provide 'rich' risk factor information[206]. Conversely, an understanding of how much data is too much data is required. For example, linking genetics, demographics, home monitoring, smart watch data may mean data is no longer de-identified. In addition, it must be understood that even data collected in large quantities may still be unrepresentative due to a the lack of access to healthcare and ability to participate in research for different populations[194].

Data provenance, whilst currently not at the forefront of discussions, could become an increasingly tangled web to unwind. Individual Trust data that is currently being used for training algorithms could at the same time be incorporated into the development of centralised evaluation datasets, resulting in a concealed overlap. The ability to track data back to the source and see all of its uses since it left the source via a flag-based system is needed. However, such systems do not currently exist and would not be easy to integrate, let alone to apply to data which has already been processed.

2.4.4 Clinical level

Clinical acumen must not be lost. AI and clinicians must work in tandem so that if one system fails (e.g. AI) the safety-net of the other system (e.g. radiologists) is in place to avoid harm. However, when AI systems operate alongside clinicians there is a possibility of the clinician becoming over dependent and automation bias to occur[145,193]. In addition, radiologists might become distracted by prompts from AI, increasing reading time and potentially adversely affecting reader performance[125]. Where these systems are designed to act independently, human supervision via 'pit-stop' analysis of a select cohort of patients, in an audit like fashion, is essential in order to maintain patient safety. The logging and reporting of errors is a potential area of AI automation where human oversight required for the monitoring of AI will necessitate vast amounts of time and resources. Nonetheless in time automation might replace certain aspects of entire jobs. This is juxtaposed against the creation of jobs in the field of healthcare informatics, to create datasets and facilitate the incorporation of AI into hospitals[174,191]. A potential overarching benefit from automation could be

that more time is freed up for clinician interaction with patients and interventions such as image guided biopsies.

A broader question exists around notifying patients when AI is used in making diagnostic and treatment decisions. Will a patient feel worse if a cancer is missed by an AI tool compared to a human reader? Another consideration is that in certain healthcare systems the prediction of cancer risk could impact patient insurance policies as well as patient mental health by causing anxiety. Therefore, prior to calculations such as the risk of developing a disease, should the patient have to approve this analysis following counselling by a healthcare professional, similar to procedures currently provided for genetic testing?

Overall, these ethical and legal dilemmas should not be underestimated and the provision of guidance from national agencies to tackle these, taking into account views from patients, commercial companies and clinicians, is essential.

2.5 Practical challenges and limitations

2.5.1 Technical level

Whilst the NHS has state-of-the-art scanners and treatments, it is also still reliant on certain record systems that are paper-based. Thus, technological advancement is a pivotal challenge facing the NHS to allow for the integration of new technology and the flexibility for exporting data on a mass scale[207]. Modifications to IT capabilities and digitisation of records is vital and should allow for communication and coordination between Trusts[207,208]. The NHS is also a tightly sealed system; however, companies will need access to update and modify their algorithms. Conversely, caution is needed when opening up systems due increasing the vulnerability to "cyber-attacks"[209]. How this external access is overseen and governed is a current technical and logistical challenge.

While the majority of data processing within the NHS at present occurs onsite, 'big data' processing for image analysis requires the procurement of Graphical Processing Units (GPUs) at Trusts or within cloud-based systems, which may entail the processing of data offsite[207]. In addition, capacity for larger data storage is needed for the curation of datasets and the storage of additional image analysis provided by algorithms. A lack of clarity still exists around suitable environments and encryption for data storage as well as the level of de-identification required. When de-identifying imaging data it is necessary to retain data that is essential for image viewing, such as the private Digital Imaging and Communications in Medicine (DICOM) tags, whilst ensuring all PID is removed[210]. As imaging becomes more advanced it is important to ensure that patients cannot be re-identified via the possibilities of image reconstruction, such as reconstructing facial features from Computer Tomography (CT) or MRI head scans.

2.5.2 Clinical level

A new multidisciplinary team will need to be developed and trained including clinical scientists and informaticians to work with clinicians to incorporate AI analysis into care decisions[143,211]. Advancing and generating new technical expertise will require access to training programmes and retention of highly skilled staff who currently re-locate to industry[174,212]. Programmes such as the NHS Digital Academy are designed to upskill healthcare professionals in areas of digital health as well as leadership and management as part of a national learning programme[143,211]. The training of radiologists is also set to change with the recent incorporation of AI into the national curriculum[213]. An openness from commercial companies to disclose the limitations of their algorithms and training radiologists how to interpret these is vital[145,194]. The use of AI itself to train radiologists or even provide continuous performance monitoring of radiologists are possibilities that need further exploration. Conversely whether the adoption of such technology will require radiologists to reach a higher level of performance to keep ahead of AI, is subject to ongoing speculation.

2.5.3 Governance level

Worldwide healthcare systems are moving forward at great pace to try utilise this technology with national funding efforts to develop an AI healthcare 'ecosystem'. In the UK, this has been facilitated by collaborations from the Accelerated Access Collaborative and NHSX with the formation of the NHS AI labs[173,174]. The same two bodies have also partnered with the NIHR (National Institute for Health Research) for the provision of an AI Award, to spur investment into promising commercial companies[164].

The recently published NHSX 'Buyers Guide' provides a much needed resource for Trusts when procuring AI technology[147]. A proposed checklist also published alongside the buyer's guide gives Trusts a procedure to help ensure vital steps of due diligence are taken, such as setting up insurance cover. However, the overall cost benefit of implementing such systems is limited in its evidence base and more robust evidence is needed to ensure systems are cost-effective.

The legal accountability of algorithms has been at the forefront of healthcare professionals' questions, as no clear guidance has been produced[189]. Discussions around the use of AI alongside a radiologist point towards the ultimate responsibility lying with the clinicians, but no specifics have been detailed as to how this would fit with NHS indemnity[144,145]. For both clinical decision support systems working alongside the radiologist and independent stand-alone systems, further guidance as to the accountability of the companies who developed the algorithm and NHS Trusts using the AI is needed. Reviews of "accidents" and "near misses" arising from the use of AI should be included in department discrepancy meetings. How this is then fed back to companies, to facilitate algorithm improvement, needs to be thought through before such events occur.

2.6 Conclusion

There are many steps to be taken by an array of national agencies, professional bodies and individual NHS Trusts before AI will become common place in breast oncological imaging to help mitigate the growing pressures facing radiology. Whilst promise is shown with algorithms across a range of imaging modalities reaching and in certain cases exceeding human performance, and even performing tasks not feasible for an individual, independent prospective testing against national benchmarks is needed.

Technical integration and upskilling the healthcare workforce is essential for AI adoption. The different ethical and legal dilemmas at the algorithm, data and clinical level should continue to be discussed and guidance updated for healthcare professionals to follow. Further research is needed not only to understand the health economic implications and testing required to ensure that systems are working by meeting the required performance thresholds, but also that latent bias is avoided. Lastly, the legal accountability should be clearly stated for companies and healthcare professionals when using such systems.

Chapter 3 – Machine learning for workflow applications in screening mammography: systematic review and meta-analysis

Advances in computer processing and improvements in data availability have led to the development of machine-learning (ML) techniques for mammographic imaging. This chapter systematically evaluates the literature for the performance of stand-alone ML applications for the screening mammography workflow. Retrospective studies demonstrate the performance of stand-alone ML applications in screening mammography can reach reader performance and provide a mechanism for case triage, which merits investigation with prospective studies.
Contents of this chapter have been published in Radiology[133] and presented at the European Congress of Radiology 2021 [abstract number - #C- 14869].

3.1 Introduction

There are now more than five Food and Drug Administration approved algorithms for mammographic interpretation, primarily to be used as clinical decision support systems[214]. Research has demonstrated that these machine-learning (ML) computer-aided detection (CAD) algorithms can reach and even exceed clinician performance, providing an independent definitive output (case level decision) on 2D standard-view mammography (mediolateral oblique and cranial caudal) data, Figure 3-1[112,215]. This could allow for ML stand-alone computer-aided detection (CADe) and computer-aided diagnosis (CADx), or, when ML algorithms are set at a high sensitivity, for the automated case-based computer-aided triage (CADt) of mammograms within the screen reading workflow[216].

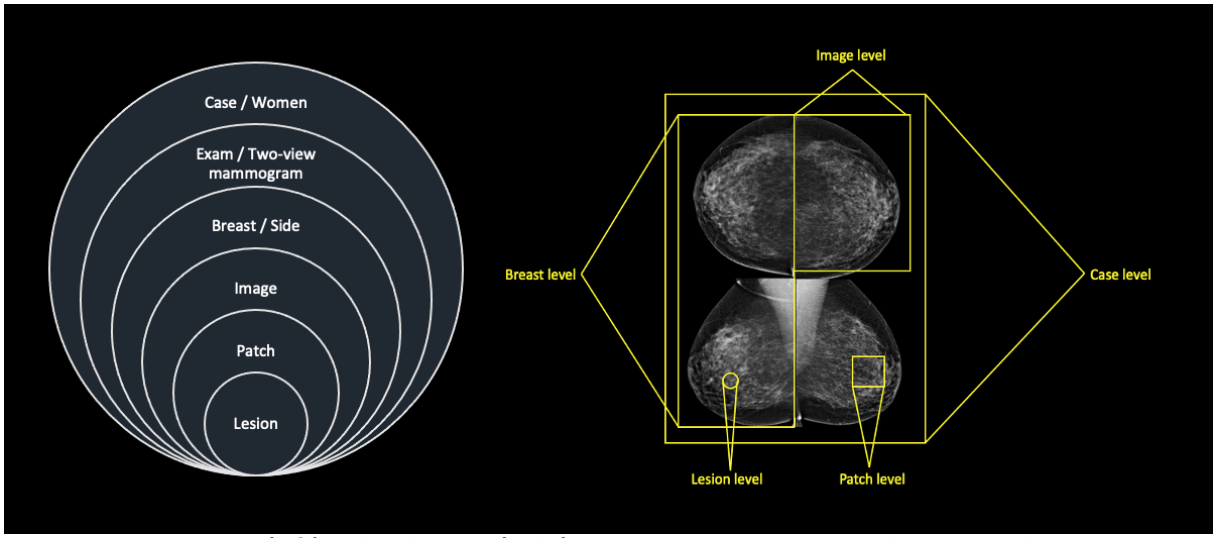


Figure 3-1 – Multi-time (left) and multi-view (right) point data that are produced by 2D standard-view mammography and can be analysed at different levels.

Many countries have implemented breast screening to detect cancer at an earlier stage, albeit with differing screening processes, such as single reading in the USA and double reading in many European countries, with screening starting at varied ages (40-50 years) and differing intervals between screening (annual, biannual and triannual)[45,74,175,217]. Mammography remains the most common imaging modality used, although its cost-effectiveness is debated due to false-positive findings, overdiagnosis, and false—negative findings (interval cancers)[36,218]. Human readers (for example radiologists and reporting radiographers in the UK) are under increasing pressure due to increasing workloads, demands from busy clinics, strict screening program targets as well as staff shortages[140]. Alternatives to double reading of mammograms are being sought to further alleviate pressure, including single reading using CAD prompts, stand-alone ML algorithms with a second reader or CADt triage with various reader combinations[215].

Studies investigating the use of traditional CAD mammography systems demonstrated no significant improvement in reader performance and, although sensitivity was similar to that of double reading, given the increase in recall rates these systems were deemed not cost-effective[125,128]. Additional limitations of traditional CAD systems include; high rates of false-positive prompts, limited reproducibility of prompts, increased reading time as well as a CAD preference for calcification detection over soft-tissue masses and architectural distortion[219,220]. Traditional CAD systems were trained using handcrafted features extracted from human delineations. The latest ML methods can use pre-trained deep learning networks and automatically delineated cancer regions via iterative interactive software to rely upon learned features, and have the potential to overcome the limitations of traditional CAD systems. However, how these new ML systems should be used in real-time workflows is still unclear. One route could be to improve efficiency of the workflow by operating as stand-alone systems. Although the performance expected by such stand-alone ML applications in a screening workflow is yet to be agreed upon, a system should meet a "clinically relevant threshold"[161]. In general, recall rates should not be increased due to the huge impact on workload, thus algorithms with lower specificity would require human intervention to reduce recalls[139,161]. Therefore, making a definitive decision on whether current systems reach the standard required for routine workflow use is challenging.

We conducted a systematic review and meta-analysis to investigate whether or not ML algorithms (CADe and CADx) are as sensitive and specific as radiologists in detecting breast cancer in subjects undergoing screening mammography. In addition, we evaluated the application of stand-alone ML algorithms (CADt) used in breast cancer screening for mammography interpretation and the impact of ML algorithms if adopted into clinical practice. Furthermore, we aimed to identify areas of bias and gaps in the reported evidence. Appendix 1 contains a glossary of terms.

3.2 Materials and methods

This systematic review and meta-analysis was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analysis for Diagnostic Test Accuracy (PRISMA-DTA) guidance[221]. The review protocol was registered on PROSPERO (CRD42019156016), Appendix 2.

3.2.1 Literature search

Digital literature databases including Ovid-EMBASE, Ovid-MEDLINE, Scopus, Web of Science, and CENTRAL were searched from January 2012 to September 2020, with the final search conducted on September 3, 2020 to include the advancements in ML algorithms for medical image interpretation and increased mammographic data availability[142,215]. Hand searches of included article references, a gray literature search of computer science databases (DBLP computer science bibliography, ACM Digital Library, and IEEE Xplore Digital Library), and a search of a pre-print literature database (arXiv) were also conducted for the same time period. The search strategy is detailed in Appendix 3.

3.2.2 Study selection

To limit bias, all publication types and all study designs were included, with no language restriction or dataset age limit applied. Eligibility criteria included women imaged using mammography for screening or diagnosis of breast cancer and a ML algorithm applied as stand-alone workflow application (CADe and CADx or CADt) with sufficient information reported for the performance of stand-alone ML algorithms and reader performance, or the simulated impact on reader performance and workflow to allow for comparison. Any ground truth (e.g., histopathology) was accepted. Because data are available at multiple levels, Figure 3-1, we included algorithms only if they provided an interpretation at the case or exam level to enable comparison with clinician performance as reported in screening programmes.

Two independent reviewers undertook the initial title and abstract screening (SEH., a physician with 2 years' experience, then one of EPVL, CL, YRI., medical students) with discordance arbitration by a third reviewer (EPVL, CL, YRI) with independent full text review (SEH and RW., a radiologist with 11 years' experience) and discordance arbitration by a third reviewer (FJG., a senior radiologist with over 30 years' experience).

3.2.3 Data extraction

A pre-designed data-extraction spreadsheet was used by the reviewers (SEH and RW) and checked by a third reviewer (AIAR., a computer scientist with 4 years' experience), Appendix 4. Results were only extracted for studies where algorithm performance was compared to readers or the impact on reader workflow and performance was reported. If studies reported multiple stand-alone algorithms, results for all algorithms were extracted.

3.2.4 Meta-analysis

For the meta-analysis, CADe and CADx algorithm performance was evaluated by adapting the method described in Liu et al[142]. The primary meta-analysis compared the best performing algorithm of each study, at the test stage using screening mammography data, with the performance of readers. Details of the primary meta-analysis study selection are available in Appendix 5. The secondary meta-analysis extended the primary meta-analysis and compared the performance of all reported algorithms and readers in all stand-alone CADe and CADx studies which used external datasets (for addressing the generalisation capabilities of the techniques), with no limitations of ground truth.

3.2.5 Quality assessment

Risk of bias and quality assessment of all included studies took place using Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2)[222,223] and Prediction model Risk Of Bias ASsessment Tool (PROBAST)[224] by two reviewers (SEH and RW), with discussion between reviewers to resolve discordance. Signalling questions for QUADAS-2 were adapted for ML studies. PROBAST questions were adapted using the technique in Nagendran et al[163]. However, as our review focused on mammography ML, applicability was assessed in all fields except the predictor field.
The Checklist for Artificial Intelligence in Medical Imaging (CLAIM)[167] guide was used by two reviewers (SEH and AIAR), with discussion between reviewers to resolve discordance. An overall reporting score for all parameters was generated as well as for eight key fields identified, and common areas under-reported were documented.

3.2.6 Statistical analysis

All statistical analyses were implemented in R (version 4.0.3; R Project for Statistical Computing, Vienna, Austria)[225] using the 'mada'[226] and 'boot'[227] packages. Normal and benign exams were combined and 2x2 contingency tables were created by calculating true-positive, true-negative, false-positive, and false-negative findings from the reported dataset characteristics and sensitivity and specificity provided, ensuring there was an integer (whole) number of cases. The heterogeneity of the included studies in the quantitative analysis was measured using the I^2 and Cochrane Q test, where high heterogeneity was defined by $I^2 > 50\%$ and $p < 0.05$ for Cochrane Q test. The estimated pooled sensitivity, specificity, and area under the receiver operator characteristic (AUROC) curve were calculated for both readers and ML algorithms using a bivariate random effects model by Reitsma et al[228] with 95.0% CIs. Bootstrapping with 100 iterations was used to generate 95.0% CI for AUROC and a t-test was used to compare the ML algorithm and reader sensitivity and specificity, with a p-value < 0.05 indicating a significant difference. Summary receiver operating characteristic

plots were constructed for both primary and secondary analyses for pooled reader and ML algorithm performance.

3.3 Results

3.3.1 Statistical selection and data extraction

A PRISMA diagram, Figure 3-2, demonstrates the study inclusion process. The search of electronic literature databases and computer science databases returned 7629 records. Removal of duplicates resulted in 4318 records. 4286 records were excluded following the screening of titles and abstracts, the remaining 32 full texts were reviewed, and 14 articles were included in the qualitative review. References of included studies can be found in Appendix 6.

From the included 14 articles, 8 studies reported a stand-alone CADe and CADx algorithm performance, and 7 studies reported the use of a CADt system. 1 article reported on both stand-alone CADe and CADx and CADt. 5 studies for stand-alone CADe and CADx provided enough information to be included in the primary meta-analysis and 6 studies for the secondary meta-analysis, (algorithm [n = 17] and reader [n = 15]).

The included articles were published between 2017 and 2020, with 3/14 (21%) articles published on a pre-print platform (arXiv). A total of 16 algorithms including 12 unique algorithms were included in this review, with 2 algorithms reported multiple times using different versions.

All included studies were conducted retrospectively. Generalizability was demonstrated in 4 studies where algorithms were tested on datasets from a different country to the training dataset. All datasets used for reader comparison testing were private. 8/14 (57%) articles evaluated algorithms on external datasets only, with a further 2/14 (14%) articles using both internal and external datasets. Cancer prevalence within testing datasets varied from 0.6% to 50.0% and the total testing dataset size ranged from 240 exams to 113,663 cases (*cohort size simulated using bootstrapping). The comparator of readers ranged in number (4-101), experience (1-44 years), and specialization (general or breast) for all studies. The algorithms code was available in 9/14 (64%) articles. Commonly used architectures included ResNet, RetinaNet and MobileNet, which are all a type of convolutional neural network. This included algorithms that were commercially available in 6/14 (43%) articles or where code was available on a public repository in 3/14 (21%) articles.

Independent CADt studies reported that between 17%-91% of normal mammograms could be identified, while missing 0%-7% of cancers, Tables 3-1 and 3-2. For CADe and CADx tasks, 8 studies reported the algorithms' AUROCs between 0.69 and 0.96, Tables 3-3 and 3-4.

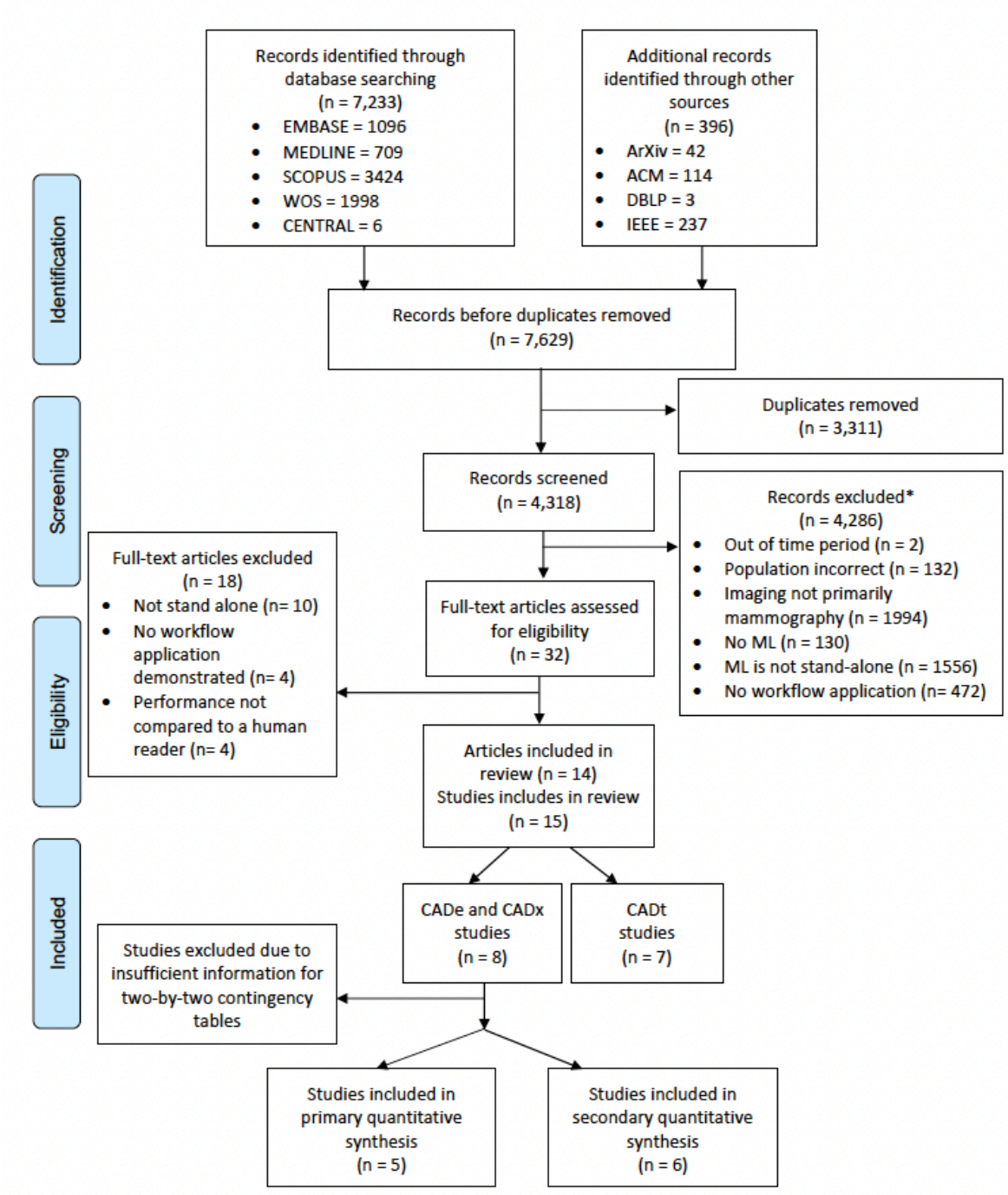


Figure 3-2 – Preferred Reporting Items for Systematic Reviews and Meta-analysis for Diagnostic Test Accuracy (PRISMA-DTA) flow diagram. *For studies included in the identification, de-duplication, screening, and data-extraction stages of this review. CADe: computer-aided detection, CADx: computer-aided diagnosis, CADt: computer-aided triage, ML: Machine Learning, WOS: Web of Science. *Studies could have been excluded for multiple reasons.*

	Reference a)	Journal	Machine Learning Technique	Task	Data Partition Level	Sample Size Development Images (Case) [Exams]	Retrospective / Prospective Testing	Internal / External Testing	Test Threshold	Test N Reader (Experience)	Test Reader Country Format	Test Validation Method	% Normal Triage (Work-load Reduction)	Evaluation (% Missed Cancers)	Code Available (Location)
CADt	Yala 2019[229]	Radiology	DL: ResNet18	Triage normals	Patient Level	Total = 238 117 (63 852) Training = 212 276 (56 831); Validation = 25 841 (7 021)	Retrospective	Internal	"Minimum probability score of a radiologist TP assessment on validation set"	23 (1 - 31 years)	USA single	Hold-out method	19.3%	(1.1%) Sensitivity: 90.1% (172 of 191; 95% CI: 86.0%, 94.3%); Specificity: 94.2% (24 814 of 26 349; 95% CI: 94.0%, 94.6%)	GitHub (https://github.com/yala/OncoNet_public)
	McKinney 2020[138]	Nature	DL: ensemble ResNet-(V2-50 and V1-50), MobileNetV2, RetinaNet	Triage normals	Patient Level	UK: Training = (13 918); Validation = (62 866) USA: Training = (55.0% of 22 225); Validation = (15.0% of 22 225)	Retrospective	*Internal / External	UK NPV 99.9% USA NPV 99.9%	UK 51 (5 – 20+ years) USA (1 – 30 years)	UK double USA single	Hold-out method	UK 41.0% USA 35.0%	ML (vs Reader): ΔAUROC = +0.115 (CI 0.06-0.18, p < 0.001); FP reduction (5.7% and 1.2%); FN reduction (9.4% and 2.7%) (USA and UK)	NA
	Balta 2020[230]	Proceedings of SPIE	*DL Unclear Architecture Commercial System Transpara (v 1.6.0)	*Triage normals to single reading	Patient Level	*Unclear The commercial system was directly used	Retrospective	*Internal / External	7	6	Germany double	External Validation	(32.6%)	(0.0%) ML decreased: recall rate 11.8% (p < 0.001); PPV 10.5% (p < 0.001)	Commercially available (https://screenpoint-medical.com/in-practice/)
	Dembrower 2020[134]	Lancet Digital Health	*DL Unclear Architecture Commercial System Lunit (v 5.5.0)	Triage normals	Patient Level	Training [170 230]	Retrospective	External	NA	NA	Sweden double	External Validation	> 60.0%	Missed cancer at 60.0%, 70.0%, 80.0%: (0.0% 0.3% (CI 0.0-4.3) 2.6% (CI 1.1-5.4))	Commercially available (https://www.lunit.io)

CADt	b)														
	Kyono [2] 2018[231]	arXiv	DL: InceptionResNetV2, Multi-Task Learning	*Triage all cases	Patient Level	(Training 90.0%, Validation 10.0%) (100.0% = 7 162)	Retrospective	Internal	Least patients seen by radiologist without degrading radiologists FPR or FNR	*Detail provided in Kyono [1] 2019	*Single multi-reader	Hold-out method	(42.8%)	Cohen's Kappa = 0.716; F1 - Score = 0.757; TP = 120; TN = 803; FP = 41; FN = 36	NA
	Kyono [1] 2019[232]	Journal of the American College of Radiology	DL: *Inception ResNetV2 Multi-Task Learning	Triage normals	Patient Level	*Unclear Training = (5 060) + 8/10 fold training + 1/10 fold validation out of (2 000)	Retrospective	Internal	NPV > 99.0%	> 30 (> 2 years)	*Single multi-reader	10 – fold CV	34.0% (CI: 25.0%-43.0%) Low prevalence: 91.0% (CI: 88.0%-94.0%)	*NPV < 99.0%	NA
	c)														
	Rodriguez-Ruiz [2] 2019[135]	European Radiology	*DL Unclear Architecture Commercial System Transpara (v 1.4.0)	Triage normals	Patient Level	*Unclear data partition for Training and Validation out of [189 000]	Retrospective	*Internal / External	5 2	101 (52.0% USA, 48.0% Europe) further detail provided in Rodriguez-Ruiz [1] 2019	Single multi-reader	External validation	Threshold of 5 = 47.0% Threshold of 2 = 17.0%	Threshold of 5 = (7.0%) Threshold of 2 = (1.0%)	Commercially available (https://screenpoint-medical.com/in-practice/)

Table 3-1 – Computer aided triage (CADt) algorithm details and results. Algorithm performance compared to reader performance for all included studies. *a) screening mammograms, b) screening mammograms used from screening recalled cases c) screening and diagnostic mammograms. AUROC: Area Under the receiver operating characteristic curve, CV: Cross Validation, DL: Deep Learning, FN: False Negative, FNR: False Negative Rate, FP: False Positive, FPR: False Positive Rate, ML: Machine Learning, N: number, NPV: Negative Predictive Value, NA: information not available, PPV: Positive Predictive Value, TN: True Negative, TP: True Positive, v: Version. * caveat or another reported format.*

	Reference a)	Test Database	Test Data Internal / External	Test Data Country	Test Data N Centres	Test Data Year of Studies	Test Data N Images (Cases) [Exams]	Test Data N Cancer Images (Cases) [Exams]	Test Data Vendor	Test Data SF / FFDM	Test Data Processed	Test Data Screen / Diagnostic Mammograms	Test Data Age of Patients	Test Data Density	Test Data Ground Truth
CADt	Yala 2019[229]	Private	Internal	USA	1	2009 - 2016	26 540 (7 176)	191 (187) (2.6%)	Hologic	FFDM	Processed	Screen	(mean 57.8 years) (SD ± 10.9)	Yes	HP / FU > 1 year
	McKinney 2020[138]	OPTIMAM (Private) + Northwestern Memorial Hospital (Private)	Internal	UK, USA	UK 2 USA 1	UK 2012 - 2015 USA 2001 - 2018	UK: (*25 856) USA: (*3 097)	UK: (*414) (1.6%) USA: (*686) (22.2%)	Hologic, GE, Siemens	FFDM	Processed	Screen	NA	Yes *USA only	HP / FU > 1 year
	Balta 2020[230]	Private	External*	Germany	1	2018	[17 895]	(114) (0.6%)	Hologic, Siemens	FFDM	NA	Screen	NA	NA	HP / no FU
	Dembrower 2020[134]	CSAW (Private)	External	Sweden	1	2009 - 2015	(7 364) (simulated 75 534)	(547) (0.7%)	Hologic	FFDM	NA	Screen	40 – 74 (median 53.6) (IQR 15.4)	Yes	HP / FU > 2 years

CADt	**b)**														
	Kyono [2] 2018[231]	TOMMY (Private)	Internal	UK	6	NA	(1 000)	(156) (15.6%)	NA	FFDM	Processed	Screen (Recalled for assessment and FHx)	40 - 73	Yes	*HP / 3 x reader review of 2D and DBT
	Kyono [1] 2019[232]	TOMMY (Private)	Internal	UK	6	NA	*Unclear 1/10 fold out of (2 000)	(300) (15.0%)	NA	FFDM	Processed	Screen (Recalled for assessment and FHx)	40 - 73	Yes	HP / 3 x reader review of 2D and DBT
	c)														
	Rodriguez-Ruiz [2] 2019[135]	Private	External	Seven countries, further detail provided in Rodriguez-Ruiz [1] 2019	NA	NA	[2 654]	[653] [24.6%]	GE, Hologic, Sectra, Siemens	FFDM	Processed	*Both (50.0% screen, 50.0% clinical)	Detail provided in Rodriguez-Ruiz [1] 2019	NA	HP / FU > 1 year

Table 3-2 – Computer aided triage (CADt) test set data characteristics of all included studies. *a) screening mammograms, b) screening mammograms used from screening recalled cases c) screening and diagnostic mammograms. CSAW: Cohort of Screen Aged Women, DBT: Digital Breast Tomosynthesis, FFDM: Full Field Digital Mammography, FHx: Family History, FU: Follow-up, HP: Histopathology, N: number, NA: information not available, OPTIMAM: OPTIMAM Medical Image Database, SD: Standard Deviation, SF: Screen Film, TOMMY: TOMosynthesis with digital MammographY. *caveat or another reported format.*

	Reference a)	Journal	Machine Learning Technique	Task	Decision	N Development Images (Cases) [Exams]	Retrospective / Prospective Testing	Internal / External Testing	Test N Reader (Experience)	Test Reader Country Format	Test Validation Method	AUROC ML vs Reader	Sensitivity ML vs Reader	Specificity ML vs Reader	Code Available (Location)
CADe and CADx	Geras 2017[116]	arXiv	DL: Customised CNN	SAID	Per case	Training 721 186 [164 224] Validation 108 276 [24 552]	Retrospective	Internal	4	Single - multi-reader	Hold-out method	macAUC 0.688 vs 0.704	NA	NA	GitHub (https://github.com/nyukat/BIRADS_classifier)
	Lotter 2019[233]	arXiv	DL: (ResNet-50 + RetinaNet)	SAID	Per case	(97 769)	Retrospective	*Internal / External	5 (2 - 15 years)	Single - multi-reader	External validation + Bootstrapping	† Test 1 ML: 0.95 (CI 0.92, 0.97) Test 2 ML: 0.77 (CI 0.71, 0.82)	† Test 1.+14.2% (CI 9.2%-18.5%, p < 0.001) ML over Reader Test 2. +17.5% (CI 6.0%-26.2%, p < 0.001) ML over Reader	† Test 1. +24.0% (CI 17.4%-30.4%, p < 0.001) ML over Reader Test 2. +16.2% (CI 7.3%-24.6%, p < 0.001) ML over Reader	NA
	Rodriguez-Ruiz [3] 2019[234]	Radiology	*DL Unclear Architecture Commercial System Transpara (v 1.3.0)	SAID	Per case	*Unclear data partition for Training and Validation out of [18 000]	Retrospective	*Internal / External	14 (11 specialists, 3 – 25 years)	Single - multi-reader	External validation	0.89 vs 0.87 (p = 0.33)	83.0% (CI 81.0%-85.0%) *Reader only	77.0% (CI 75.0%-79.0%) *Reader only	Commercially available (https://screenpoint-medical.com/in-practice/)

CADe and CADx	McKinney 2020[138]	Nature	DL: ensemble ResNet-(V2-50 and V1-50), MobileNetV2, RetinaNet	SAID	Per case	UK: Training (13 918); Validation (62 866) US: Training (55.0% of 22 225); Validation (15.0% of 22 225)	Retrospective	*Internal / External	UK 51 (5 - 20+ years) USA (1 – 30 years) Reader study 6 (4 – 15 years)	UK double USA single Reader study single – multi-reader	Hold-out method + External validation	ML UK: AUROC = 0.89 (CI 0.87 - 0.91) USA (w/training UK+US): AUROC = 0.81 (CI 0.79-0.83) † (UK training only: AUROC = 0.76 (CI 0.73-0.78)) Reader study ML vs Reader: ΔAUROC = +0.115 (CI 0.06-0.18, p < 0.001)	† (+8.1%, p < 0.001) ML improvement % over Reader range [min-max]: [0.0-9.4]	† (+3.5%, p = 0.02) ML improvement % over Reader range [min-max]: [-3.4-5.7]	*NA
	Schaffter 2020[137]	JAMA Open Network	DL: CEM Ensemble (8 networks including VGG, Faster- RCNN) DL: Customised VGG network	SAID	Per case	KPW (59 923) [100 974] + DDSM + Other datasets (e.g. OPTIMAM)	Retrospective	External	USA screen readers Sweden screen readers	USA single Sweden double (*reported single first reader)	Hold-out method + External validation	KPW (CEM) 0.90 (Top performing) 0.86 † KI (CEM) 0.92 (Top Preforming model) 0.90	KPW *Reader sensitivity 85.9% † KI *First reader 77.1%, Reader consensus 83.9%	KPW (CEM) 76.1% (Top performing) 66.3% vs 90.5% † KI (CEM) 92.5% (Top performing model) 88.0%, 81.2% vs * † First reader 96.7%, Reader consensus 98.5%	GitHub (https://github.com/Sage-Bionetworks/DigitalMammographyEnsemble)

CADe and CADx	Salim 2020[149]	JAMA Oncology	DL: (1) ResNet34 (2) MobileNet (3) Unknown	SAID	Per case	(1) 752 000 (2) 239 000 (3) 112 000	Retrospective	External	Sweden screen readers 25 1st reader, 20 2nd reader	Sweden double	External validation + Bootstrapping	(1) 0.96 (2) 0.92 (3) 0.92	ML: (1) † 81.9% (p = 0.03) (p = 0.11) (2) 67.0% (3) 67.4% vs † First reader 77.4%, Reader consensus 85.0%	ML: (1) † 96.6% (2) 96.6% (3) 96.7% vs † First reader 96.6%, Reader consensus 98.5%	NA
	b)														
	Rodriguez-Ruiz [1] 2019[235]	Journal of the National Cancer Institute	*DL Unclear Architecture Commercial System Transpara (v 1.4.0)	SAID	Per case	*Unclear data partition for Training and Validation out of [189 000]	Retrospective	*Internal / External	101 *95 for sensitivity and specificity (1 – 44 years)	Single - multi-reader	External validation	† 0.84 (CI 0.82-0.86) vs 0.81 (CI 0.79 -0.84)	† 75.0%–86.0% vs 76.0%–84.0%	† 49.0% – 79.0% *Clinician specificity	Commercially available (https://screenpoint-medical.com/in-practice/)
	c)														
	Kim 2019[236]	Lancet Digital Health	DL: (ResNet-34) Commercial System Lunit	SAID	Per case	Total [166 968] Training [152 693] Validation [14 275]	Retrospective	*Internal / External	14 (7 specialists, > 6 months)	Single - multi-reader	External validation	0.94 (CI 0.92–0.97) vs 0.81 (CI 0.77–0.85, p < 0.001)	88.8% vs 75.3% (p < 0.001)	81.9% vs 72.0% (p = 0.002)	Commercially available (https://www.lunit.io)

Table 3-3 – Computer aided detection (CADe) and Computer aided diagnosis (CADx) algorithm details and results. Algorithm performance compared to reader performance for all included studies. *a) Screening mammograms, b) screening and diagnostic mammograms c) mammography and ultrasound used for screening. CEM: Challenge Ensemble Method, CI: Confidence Interval, CV: Cross Validation, DL: Deep Learning, DDSM: Digital Database for Screening Mammography, KPW: Kaiser Permanente Washington, N: number, NPV: Negative Predictive Value, NA: information not available, OPTIMAM: OPTIMAM Medical Image Database, SAID: Stand-alone AI Detection, v: Version. * caveat or other reported format. † The results of studies included in the primary meta-analysis.*

	Reference a)	Test Database	Test Data Internal / External	Test Data Country	Test Data N Centres	Test Data Year of Studies	Test Data N Images (Cases) [Exams]	Test Data N Cancer Images (Cases) [Exams]	Test Data Vendor	Test Data SF / FFDM	Test Data Processed	Test Data Screen / Diagnostic Mammograms	Test Data Age of Patients	Test Data Density	Test Data Ground Truth
CADe and CADx	Geras 2017[116]	NYU (Private)	Internal	USA	5	2010 - 2016	[500]	NA	NA	FFDM	Processed	Screen	19 - 99 (mean 57.2) (SD 11.6)	NA	*BIRADS score (0, 1, 2)
	Lotter 2019[233]	Private	External	USA	1	2011 - 2014	Test 1. ("Index") [285] Test 2. ("Pre-index 12-24 month prior") [274]	Test 1. [131] [46.0%] Test 2. [120] [43.8%]	NA	FFDM	Processed	Screen	NA	NA	HP / FU > 1 year
	Rodriguez-Ruiz [3] 2019[234]	Private	External	USA, Europe	USA 1 Europe 1	USA 2013 – 2017 Europe 2014 - 2015	[240]	[100] [41.7%]	Hologic, Siemens	FFDM	Processed	Screen	39 – 89 (mean 61.0)	Yes	HP / FU > 1 year
	McKinney 2020[138]	OPTIMAM (Private) + Northwestern Memorial Hospital (Private)	*Internal / External	USA, UK	UK 2 USA 1	UK 2012 - 2015 USA 2001 - 2018	UK: (25 856) USA: (3 097) Reader study USA (*500)	UK: (414) (1.6%) US: (686) (22.2%) Reader study USA (*125) (25.0%)	GE, Hologic, Siemens	FFDM	Processed	Screen	NA	Yes *USA only	HP / FU > 1 year
	Schaffter 2020[137]	KPW (Private) KI (Private)	*Internal / External	USA, Sweden	KPW 1 KI 2	KPW NA KI 2008 - 2012	KPW (25 657) [43 257] KI (*68 008) [166 578]	KPW (283) (1.1%) KI (*780) (1.1%)	NA	FFDM	Processed	Screen	KPW 40 - 74 (mean 58.4) (SD 9.7) KI 40 - 74 (mean 53.3) (SD 9.4)	NA	HP / FU > 1 year

CADe and CADx	Salim 2020[149]	CSAW (Private)	External	Sweden	1	2008 - 2015	(8 805) (Simulated 113 663)	(739) (Simulated 0.7%)	Hologic	FFDM	Processed	Screen	40 - 74 (median 54.5)	Yes	HP / FU > 2 years
	b)														
	Rodriguez-Ruiz [1] 2019[235]	Private	External	Sweden, UK, Netherlands, USA, Italy, Spain, Austria	NA	NA	[2 652] [*2 389] *for sensitivity and specificity	[653] [24.6%] [*610] [24.6%]	GE, Hologic, Sectra, Siemens	FFDM	Processed	*Both (Some unilateral only)	30 - 92	Yes	HP / FU > 1 year
	c)														
	Kim 2019[236]	Private	*External	South Korea	2	2009 - 2018	[320]	[160] [50.0%]	GE, Hologic	FFDM	NA	Screen (*including US)	(mean 53.2) (SD 10·0)	Yes	*Mammography / USS detected + HP

Table 3-4 – Computer aided detection (CADe) and Computer aided diagnosis (CADx) test set data characteristics of all included studies *a) Screening mammograms, b) screening and diagnostic mammograms c) mammography and ultrasound used for screening. CSAW: Cohort of Screen Aged Women, FFDM: Full Field Digital Mammography, FU: Follow-up, HP: Histopathology, KI: Karolinska Institute, KPW: Kaiser Permanente Washington, N: number, NA: information not available, NYU: New York University, OPTIMAM: OPTIMAM Medical Image Database, SD: Standard Deviation, SF: Screen Film. *caveat or other reported format.*

3.3.2 Quality assessment

The PROBAST and QUADAS-2 tools were applied to all included articles in this review, and summary results of assessments are shown in Figure 3-3 and Appendix 7. Applying both tools identified a high risk of bias for analysis, as well as high bias and applicability concerns for the index test, participants and patient selection, Figure 3-3. Reasons for high bias and applicability include 8/14 (57%) articles with cancer-enriched cohorts, 5/14 (36%) articles that tested the algorithm on an internal dataset, and 3/14 (21%) articles that did not pre-set the algorithm threshold in CADt studies. According to PROBAST assessment, articles were reported to have an overall low (7%), unclear (7%), and high risk (86%) of bias.

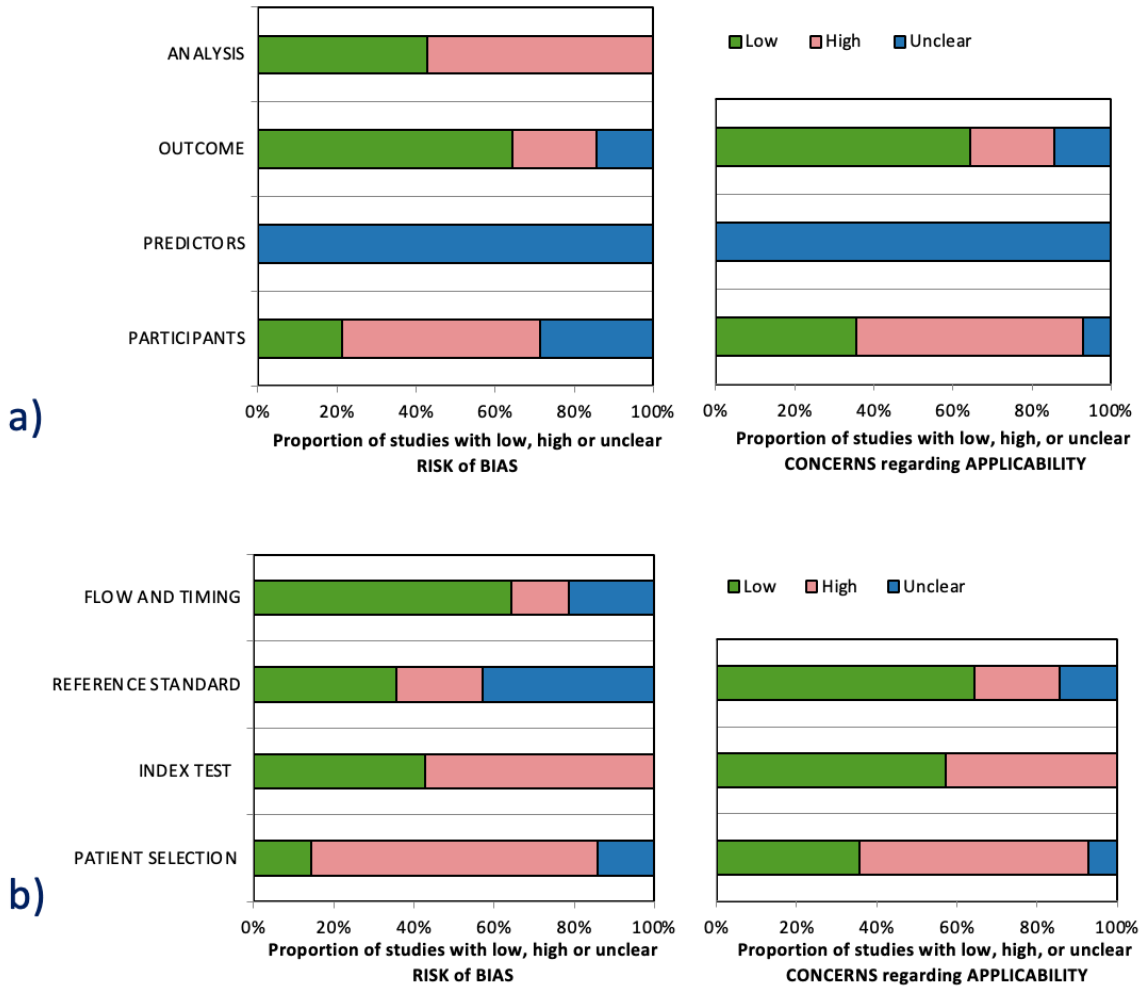


Figure 3-3 – (a) Prediction model Risk Of Bias ASsessment Tool (PROBAST) and (b) Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) assessment. *For 14 included articles, each category is represented as a percentage of the number of articles that have high, low, or unclear levels of bias.*

Critical appraisal of the reporting quality in the 14 included articles using the 42 parameters of CLAIM, found scores ranging from 22 to 34, with an average total score of 30/42 (71%). The points

most commonly under-reported included robustness or sensitivity analysis, methods for explainability or interpretability, and protocol registration. Methods for explainability (e.g., saliency maps) to provide transparency of the algorithm's deduction were reported in 3 articles. Only 50% of articles reported all eight key fields, Figure 3-4.

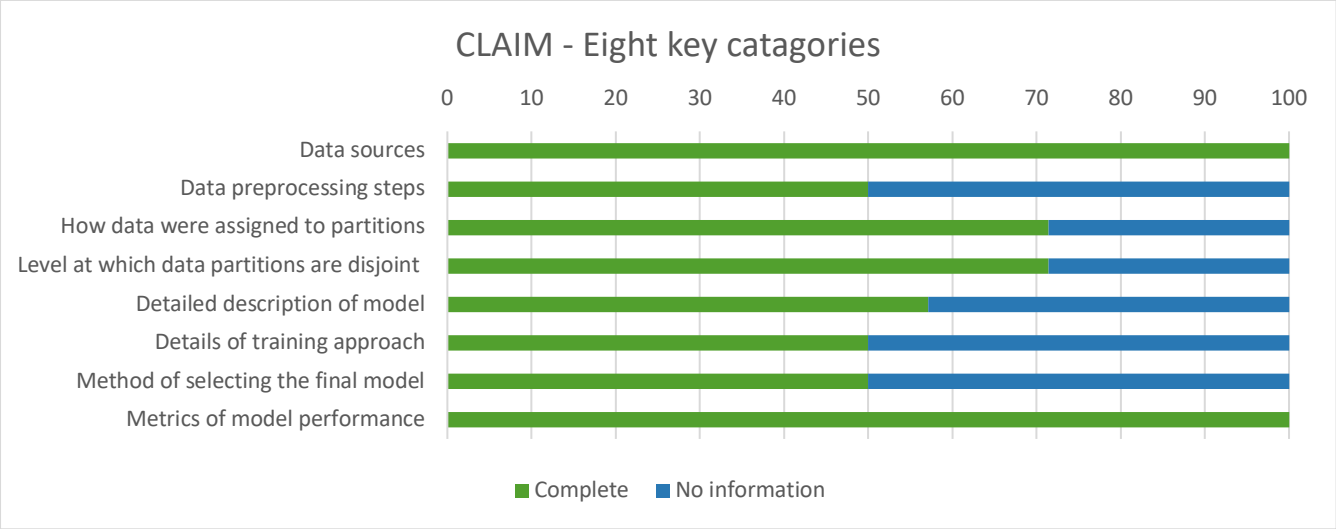


Figure 3-4 – Checklist for Artificial Intelligence in Medical Imaging (CLAIM) assessment. *Results for 14 articles included in this review across 8 key categories identified from the checklist. A score of 1 was provided if complete information was provided, and 0 where no information was provided. The x-axis indicates the percentage of articles in the review which included information about the eight key categories detailed in the y-axis.*

3.3.3 Statistical analysis

Low heterogeneity was found for both algorithms and readers in the primary and the secondary analyses (I^2 0.0%-0.6% and Cochrane Q test p = 0.45-0.78).

An estimated 185,252 cases from 3 countries with > 39 readers were included in the primary meta-analysis. The pooled summary estimates for sensitivity, specificity, and AUROC were 75.4% (95% CI 65.6-83.2), 90.6% (95% CI 82.9-95.0), and 0.89 (95% CI 0.84-0.98), respectively for ML algorithms. For readers, the pooled sensitivity, specificity, and AUROC were 73.0% (95% CI 60.7-82.6), 88.6% (95% CI 72.4-95.8), and 0.85 (95% CI 0.78-0.97), respectively, Figure 3-5. The differences in sensitivity and specificity were not statistically significant, p-value = 0.11 and 0.40 respectively. Algorithms performance thresholds were set at the reported reader sensitivity / specificity in 4 studies.

When including all available results from CADe and CADx studies conducted using external datasets that provided a direct comparison between ML algorithms and readers for a secondary meta-analysis, the pooled sensitivity, specificity, and AUROC was 80.4% (95% CI 75.5-84.6), 82.1% (95% CI 72.7-88.8), and 0.86 (95% CI 0.84-0.90) for algorithms. For readers the pooled sensitivity, specificity, and AUROC was 78.5% (95% CI 73.8-82.5), 82.6% (95% CI 69.2-90.9), and 0.84 (95% CI 0.81-0.88), Figure 3-5. The differences in sensitivity and specificity were not statistically significant, p-value =

0.70 and 0.73 respectively. Summary tables and additional information are available in Appendix 8-11.

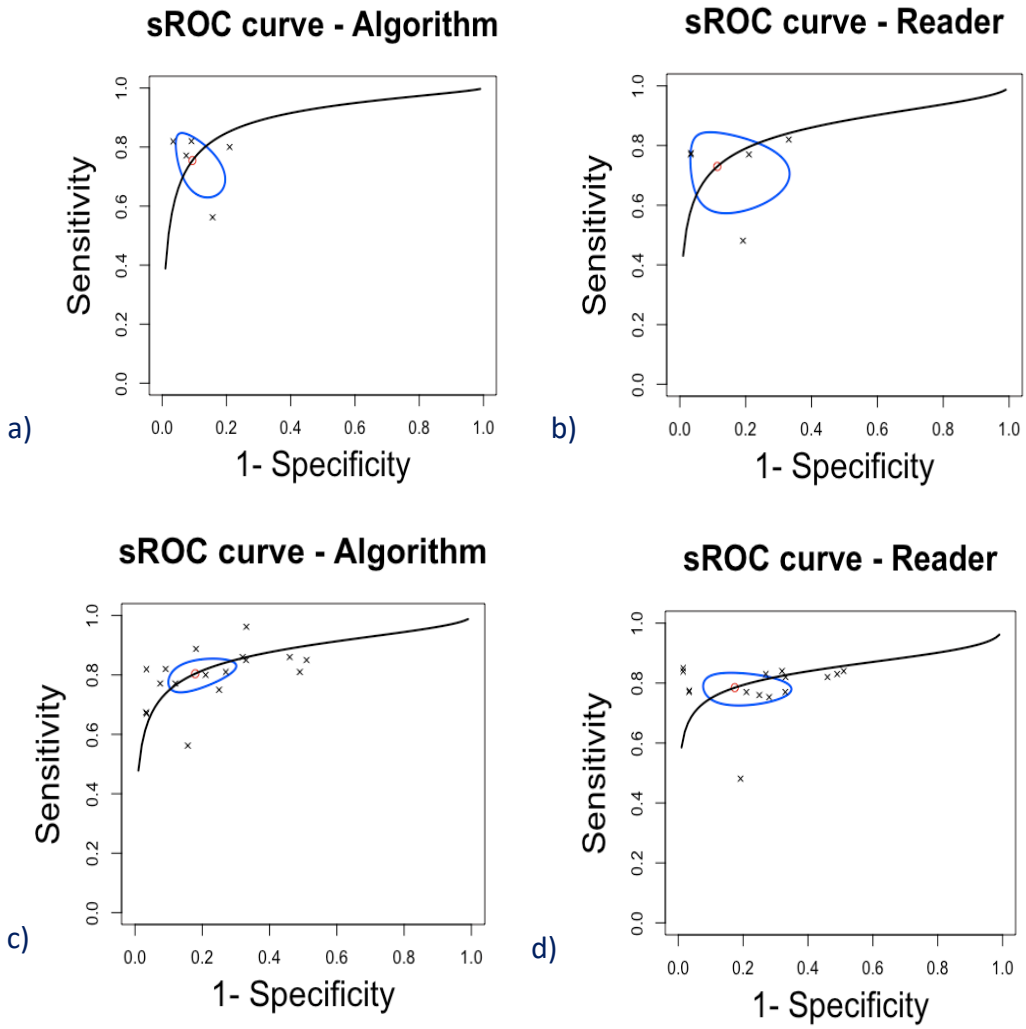


Figure 3-5 – Summary Receiver Operating Characteristic (ROC) curves *(a) in 5 studies for the included algorithm and b) reader results reported for the top performing machine learning (ML) algorithm tested on an external data set, compared to reader performance for computer-aided detection (CADe) and computer-aided diagnosis (CADx) applications, with a ground truth of > 1 year follow-up and histopathology. Primary meta-analysis) (c) For 17 algorithm reported results and d) 15 reader reported results from included studies for CADe and CADx applications tested externally. Seconday meta-analysis). The black line represents sROC, the blue line represents confidence interval, the red dot represents the summary estimate, and the black crosses represent the individual results.*

3.4 Discussion

We found the performance of mammography screening algorithms is reaching equivalence to readers in stand-alone CADe and CADx tasks. Comparing our results to two recently published reader performance studies demonstrated that while the pooled sensitivity of algorithms (75.4%) was higher than that of pooled readers (73.0%) and single reading in Sweden (73.0%)[175], it was inferior to both single reading in USA (86.9%)[74] and double reading with consensus in Sweden (85.0%)[175]. The

pooled specificity of algorithms (90.6%) was superior to pooled readers (88.6%) and single reading in USA (88.9%)[74], but inferior to both single (96.0%) and double reading with consensus in Sweden (98.0%)[175]. Therefore, further improvements are needed to make sure ML systems meet the 'clinically relevant thresholds' of current reader performance and screening programme targets. Our findings are similar to a systematic review and meta-analysis comparing deep learning applications across all medical imaging to "health-care professionals", who came to a similar conclusion and highlighted the importance of continued external testing[142].

Algorithms are also performing tasks not feasible by readers such as high-volume normal case triage, with no detrimental change when reader performance was extrapolated in an adapted screening workflow (using machine only reading of cases assigned to be normal as an alternative to single or double reading)[215]. However, the acceptable "miss" rate for a system, similar to the interval cancer targets, should be agreed and specified for machine only reading of normal mammograms before clinical adoption. The biggest barrier may be public understanding of the concept of acceptable "misses".

No prospective studies have yet been reported, many studies are still conducted with retrospective internal testing, and few studies are conducted by an independent party where multiple algorithms are cross-compared using external datasets[149]. In addition, most of the studies used enriched cancer cohorts for testing, which do not include the class imbalance of cancers to healthy controls in screening. Thus, these datasets may not provide a realistic representation from which to infer model performance in clinical implementation limiting generalisability, clinical applicability and feasibility of workflow translation. Our findings highlight the need for well-designed prospective randomised and non-randomised controlled trials to be conducted across different breast screening programmes. These prospective studies should include representative case proportions, to replicate the class imbalance in screening, with readers of varying experience interacting with ML algorithm outputs within the clinical workflow. This will allow performance to be assessed as well as technological feasibility, reading time, reader acceptability and effect on reader performance[139]. Prospective studies investigating ML applications for mammographic screening are currently underway in the UK, Norway, Sweden, China and Russia with results pending[237–239].

Most articles were from 2019 onwards, reflecting the exponential growth in publications since major milestones such as the ImageNet[118] and DREAM[112,240] challenges. Although the computer codes were available in 64.0% of articles, only 21.0% of code was available on an open-source platform. However, the provision of code alone does not result in a deployable model including training weights as well as the threshold at which the algorithm performance was determined, limiting

reproducibility and transparency[241,242]. Large datasets were used for testing but the majority of these are private, which limits the ability to replicate results.

Two commonly used tools for bias assessment found high risk of bias due to cancer enriched cohorts, use of internal datasets as well as due to the algorithm threshold in triage studies not being pre-set. Therefore, these results may not be applicable and generalisable to all breast screening populations[223]. We applied a specific Artificial Intelligence (AI) medical imaging reporting guideline (CLAIM), to critically appraise AI medical imaging literature. It should be noted that CLAIM was published after more than half of the articles in this review were published. Therefore, we have not presented the results of each individual study but have used this as a foundation to find underreported areas within the current literature, as well as confirm the applicability of CLAIM for ML mammography studies[167].

3.4.1 Limitations

The meta-analysis was limited by both the small number of eligible studies and because the contingency tables were constructed using reported sensitivity, specificity, total cases and malignant cases to provide estimated integers (whole numbers) for calculating true-positives, true-negatives, false-positives, and false-negatives. The primary meta-analysis included studies where reader performance did not reach the level reported in national screening standards, therefore it is possible that the relative improved performance of ML algorithms is overestimated, and the performance of readers is underestimated as part of this analysis. The primary analysis also used only the highest performing (based on test performance) algorithm if multiple algorithms were tested, and therefore may be slightly biased towards the selected algorithms. The secondary meta-analysis incorporated multiple algorithms and readers from the same study, on the same population, which could potentially lead to overrepresentation. Therefore, the results from the meta-analysis should be interpreted with caution. Lastly, for the secondary meta-analysis both screening and diagnostic mammograms were included in studies, as well as in one study women were screened using mammography and ultrasound, both of which would impact on the expected performance metrics.

3.5 Conclusion

There is a growing evidence base that stand-alone ML performance is comparable to reader performance and that ML can undertake triage tasks at a volume and speed not feasible for human readers. Although only retrospective trials have been conducted, the potential for algorithms to perform at the level of or even exceed the performance of a reader within the real-time breast screening workflow is realistic. However, further robust prospective data is critical to understanding where algorithm thresholds are set and are required to examine the interaction between human

readers and algorithms, as well as the effect on reader performance and patient outcomes over time.

Chapter 4 – Developing a mammographic imaging database – The Cambridge Cohort – Mammography East Anglia Digital Imaging Archive

4.1 Aims

In this chapter the design, construction, governance, and content of The Cambridge Cohort – Mammography East Anglia Digital Imaging Archive (CC-MEDIA) is outlined. The CC-MEDIA database aims to provide extensive clinical metadata and multiple imaging episode data, fulfilling a need for a large-scale, external representative UK breast screening dataset for benchmarking. This allows for reproducible, independent testing and feedback of Artificial Intelligence (AI) models being developed for breast cancer screening. As a result of this work setting up the CC-MEDIA database, the methods and procedures developed have informed the University of Cambridge and Cambridge University Hospitals NHS Foundation Trust research ethics procedures regarding medical imaging databases. Contents of this chapter have been presented at the 2021 British Society of Breast Radiology Annual Scientific Conference[243].

4.2 Introduction

Medical imaging has become an integral part of patients care, with 45.2 million imaging tests taking place between September 2018 to September 2019 in the UK[115]. The visual field of radiology lends itself to AI research, for both the visual task at hand as well as due to the abundance of medical imaging data available. High quality medical imaging databases are therefore increasingly important for AI algorithm development and testing. The 2022 Goldacre report highlighted the importance of development of Trusted Research Environments (TRE) to fully utilise the digital data available within the National Health Service (NHS). The report acknowledged the current complex and convoluted ethical approval and governance required as well as the necessary expertise to build large reusable datasets[244]. Initiatives across medical imaging research have led to the development of large imaging databases for chest x-ray (MIMIC-CXR), MRI knee (MRNet) CT head (RSNA 2019 brain haemorrhage challenge), and mammography (CSAW)[157,245–247]. Early mammographic imaging databases date back to 1994 but only contained a very small volume of digitised screen film images, from the USA and UK[248,249]. Recent mammographic imaging databases contain data in Full Field Digital Mammography (FFDM) Digital Imaging and Communications in Medicine (DICOM) format, from countries around the world at a larger scale of > 1,000,000 images[157,158,250]. The increase in data availability in the last ten years has contributed to the development of more accurate algorithms as well as allowing for the reproducibility and generalizability testing of AI algorithms in different screening programmes.

Mammographic imaging databases now contain a diverse range of mammography machine manufacturers, screening programmes (annual, biennial, triennial), and countries (UK, Sweden, USA, Spain, Portugal), as detailed in Table 4-1.

Dataset	Country	Year	SF / FFDM	Cases	Density	Age	Annotations
MIAS[248]	UK	1994	SF	161 (I = 322) (C = ~52)	Y	NA	Y
DDSM[249]	USA	1998-1999	SF	2620 (I = 10480) (C = 914)	Y	NA	Y
CBIS-DDSM [251,252]	USA	2016	SF	1644 (I = 10239) (C = 758)	Y	NA	Y
InBreast [156,253]	Portugal	2008-2010	FFDM	115 (I = 410) (C = NA)	Y	NA	Y
BCDR-FM [254,255]	Portugal + Spain	2009-2013	SF	1010 (I = 2702) (C = NA)	Y	20-90	Y
BCDR-DM [254,255]	Portugal + Spain	2009-2013	FFDM	724 (I = 3612) (C = NA)	Y	27-92	Y
OMI-DB[250]	UK	2011-2020	FFDM + DBT + MRI	172282 (I > 3000000) (C = 8586)	Y	30-84	Y
CSAW[157]	Sweden	2008-2015	FFDM	499807 (I > 2000000) (C = 10582)	NA	40-74	Y
NYU BCSD v1.0[158]	USA	2010-2017	FFDM	141473 (I > 1000000) (C = 1221*)	Y	16-99	Y
EMBED[256]	USA	2013-2020	FFDM + DBT	115910 (I > 3000000) (C = 3733)	Y	> 18	Y

Table 4-1 – Mammographic imaging database characteristics. *BCDR: Breast Cancer Digital Repository, CBIS-DDSM: Curated Breast Imaging Subset of DDSM, CSAW: Cohort of Screen-Aged Women, C: Cancers, DDSM: Digital Database for Screening Mammography, DBT: Digital breast tomosynthesis, DM: Digital mammography, EMBED: EMory BrEast imaging Dataset, FFDM: Full field digital mammography, FM: Film mammography, I: Images, MIAS: The Mammographic Image Analysis Society Digital Mammogram Database, NYU BCSD: New York University Breast Cancer Screening Dataset, NA: Not available, OMI-DB: The Optimam Mammography Image Database, SF: Screen film, Y: Yes. *breasts not cases.*

Different levels of data are available for mammography images (case level, exam level, per breast level, per image level) with certain datasets also providing image level annotations either via bounding box or pixel level regions of interest. The ground truth of the data is an important component of medical imaging databases. Routinely in breast screening AI testing the ground truth

is set at two levels. For a case to be defined as 'normal' there has to be a sufficient time interval with an outcome of a normal screen, and for 'cancers' there should be a histopathological diagnosis outcome.

The NHS Breast Screening Programme (BSP) is a three yearly (triennial) breast screening programme, inviting women aged 50-70 to participate and using a two-view FFDM. Due to NHSBSP age extension trial (AgeX), running from 2011 to 2016, women aged 47-49 years old were also invited to screening. In addition, women aged more than 70 years old can self-refer into the screening programme[54,257].

The NHSBSP is carried out at seventy-five sites across the country within the NHS system that allows for women to be tracked over time and linkage between different data sources using personal identification numbers (NHS number). All mammograms are double read by two expert readers (e.g. radiologists, consultant radiographers, and breast clinicians) either independently or dependently.

The CC-MEDIA database captures the true distribution of the NHSBSP by consecutively collecting screening mammograms for women aged more than 47 years old who attended screening at two NHSBSP sites between 2011 and 2020. Thus, facilitating the independent testing of multiple AI algorithms, for different breast screening applications using large, representative cohorts with extensive follow up to allow for accurate ground truth identification.

4.3 Methods

4.3.1 Database approval

Ethical approval for this database was obtained from the Health Research Authority (HRA) Confidentiality Advisory Committee (CAG), HRA Research Ethics Committee (REC) and Public Health England (PHE) Research Advisory Committee (RAC). IRAS Reference – 258761.

- HRA REC - reference 20/LO/0104 – approval date 03/04/2020
- PHE RAC - reference BSPRAC_090 – approval date 03/04/2020
- HRA CAG - reference 20/CAG/0009 – approval date 11/06/2020

A formal agreement was put in place between Cambridge and Norwich hospital trusts relating to the use of data within this database. Consent was not obtained from individual patients for the creation and use of this database, as the data that is retained within the final Trial Database is in a de-identified format and Section 251 approval was received from the HRA CAG committee. The database has received initial 5-year approval until 2025 and yearly reports are submitted to the ethics committees to maintain support.

4.3.2 Database governance

The database is overseen by The Cambridge Cohort Database Access Committee (DAC). The DAC ensures that the management of the databases is in line with current regulations for data

governance and patient safety. The DAC includes; the principal investigators, representatives from the breast screening units and data managers at both sites, research governance leads at the university and hospital, as well as a lay member. The DAC is responsible for reviewing applications for data access by internal and external sources as well as determining the terms of access given. As the data is unconsented and sensitive, data is not routinely released to external institutions. If requests for processing using a small volume of data is approved by the DAC, e.g. for the specific purpose of company AI tool validation on Philips data, a data sharing agreement is put in place and the small volume of data [n = ~100 cases] is re-anonymised and transferred via a secure process (e.g. secure file transfer protocol (SFTP)).

4.3.3 Patient and public involvement work

Throughout the development of this database extensive patient and public involvement (PPI) work has been undertaken to ensure the views of the patients included in this database and those of the general public are taken into account regarding the management and use of their data. The feedback received from our PPI events has improved the way we explain to people how their patient data is used as part of this research. It also improved explanations regarding how data moves from the hospital to the university, who has access to the data, and how the data is then used in a de-identified format (where all the information that could be used to identify an individual has been removed or amended). Exploring patient acceptability of different aspects of data use has helped ensure we are working both within the public's expectations as well as in line with national ethical requirements. All the PPI work was carried out with the support and guidance of the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (BRC) PPI team.

The first PPI activity was a formal review by the NHIR Cambridge BRC PPI panel of the project lay summary (15/11/2019). The NHIR Cambridge BRC PPI panel is a group of around 60 members of the general public from Cambridge and the surrounding areas who are interested in research. They provide their thoughts and opinions on research projects based on their own personal experiences. Seven panel members reviewed the lay summary and provided feedback. This included clarifying the terms used in our patient facing material to make these more accessible, for example providing an explanation around the term "de-identified". They also raised queries around data flows, and commercial involvement, and how commercial companies will access the data. Following on from this initial activity a Cambridge Science Festival public forum was held on 9th March 2020 to gain insights into the public's views on "Harnessing Big Clinical Data In Medicine. Can AI Improve Breast Cancer Screening?". Thirty-seven members of the public attended, 58.0% were female and 42.0% male, of which 53.0% were aged 18-29, 22.0% aged 30-49, and 26.0% were aged more than 50 years

old. Throughout the session responses to questions were collected from the audience using TurningPoint software (which uses interactive clickers to record anonymised results). This interactive feedback helped to understand how certain terms were phrased and where further explanation should be provided for the complex flow of data in this research. The audience was asked about the acceptability of different organisations having access to the database, such as the university, hospital, and commercial companies. There was a high proportion of agreement for the university and hospital with 63.9% of the audience strongly agreeing and 16.7% agreeing, however a mixed picture for the commercial companies with only 18.8% strongly agreeing and 28.1% agreeing, Figure 4-1.

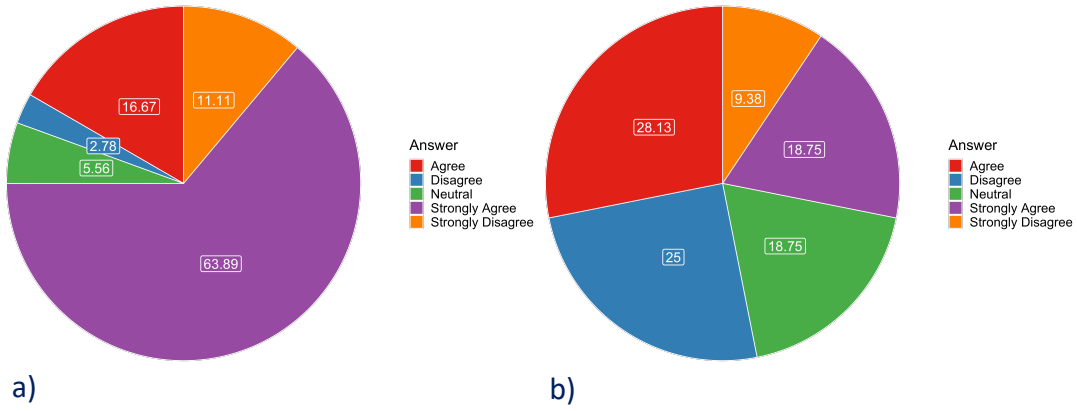


Figure 4-1 – Cambridge Science Festival Event questions. *a) "Would you support the use of your medical images by a hospital or university (in a fully anonymised format, stored in a secure location) to be used for developing algorithms without your consent?", b) "Would you support the use of your medical images by a commercial company (in a fully anonymised format, stored in a secure location) to be used for developing algorithms without your consent?"*

Based on the questions outlined in Figure 4-1, further work was carried out to clarify the role of commercial companies in this research. Such that de-identified data would only be released in small proportions to external companies, and that the data would be held securely within the University of Cambridge so that the algorithms are brought to the data and only those with approved access could see and use the data. A glossary of commonly used terms was developed for our project following this event to be used for future PPI communication as well as an anonymised report which summarised results from our question-and-answer sessions was submitted HRA REC, HRA CAG and PHE RAC for initial ethical approval.

A national patient survey called "The AI Survey - The use of patient data in breast cancer screening artificial intelligence research" was conducted in October 2021. The survey was disseminated through the NIHR Cambridge BRC team, Independent Cancer Patient Voices (ICPV), Breast Cancer Now, Addenbrookes Cancer Patient Partnership Group, and Cancer Research UK, to patients; eligible for breast screening, those who have previously attended breast screening, or have been previously

diagnosed with breast cancer. The survey was hosted using the university Qualtrics platform from 27/10/2020 to 31/01/2021. In total 46 responses were received. The survey highlighted areas for further improvement surrounding the terms used and layout of the patient facing material. We were able to demonstrate the improved clarity of the updated lay summary. In addition, the acceptability of the data fields collected in the database, Figure 4-2. Patients were very likely to accept the use of all fields to be collected in the database. However, 2.0%-6.1% of patients were unlikely or very unlikely to accept the use of family history or additional healthcare information e.g. information relating to other health conditions such as medication.

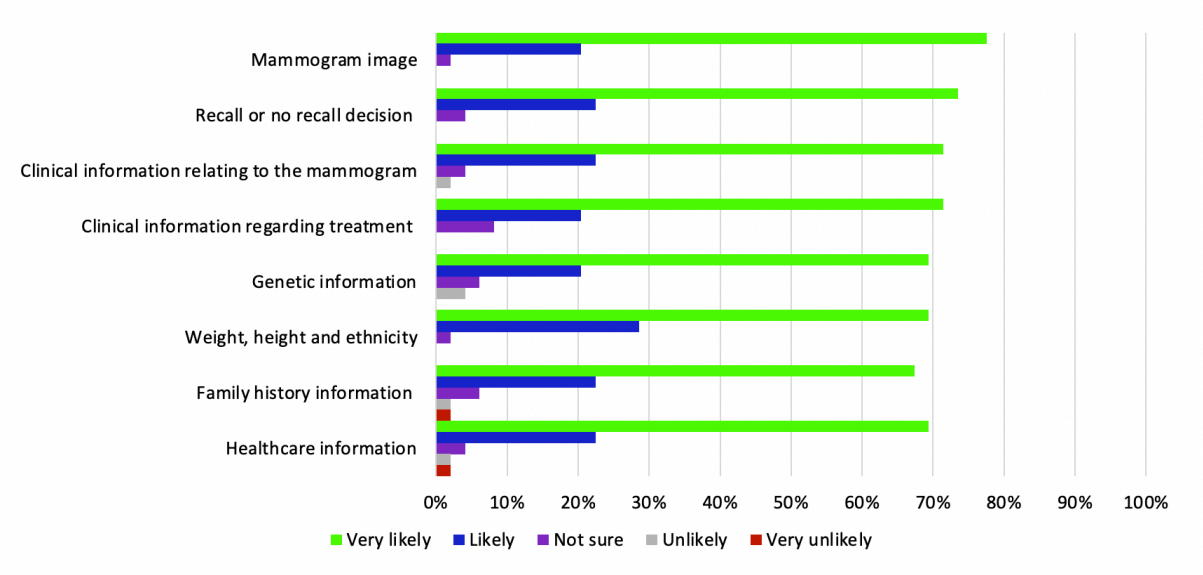


Figure 4-2 – National patient survey question regarding acceptability of data fields. *"How likely are you to accept the secure storage and use for research (algorithm testing and development) of each field of your de-identified healthcare data?".*

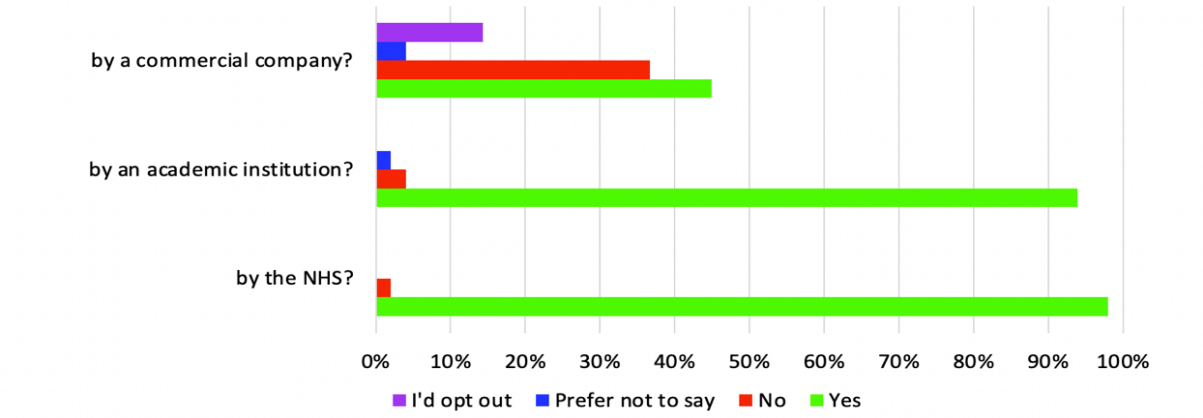


Figure 4-3 – National patient survey questions regarding commercial involvement. *"Would you accept the use of your healthcare data (securely stored in a de-identified format without your consent) for algorithm testing research in the following circumstances:"*

However, we acknowledged the lay summary was felt to be too long and so we produced a shorter version with all the key information still included. Similar results to those found at the science festival public forum were found in the survey regarding "Would you accept the use of your healthcare data (securely stored in a de-identified format without your consent) for algorithm testing research in the following circumstances", Figure 4-3. Further supporting the acceptance of bringing the algorithms to data approach that was taken in this research.

Follow-up small discussion forum groups were held via Zoom in January and February 2021 for people who contacted the research team following the survey. These discussions highlighted that there is hesitancy regarding commercial involvement in this research, mainly regarding concerns over data privacy. However, there was also an understanding of the need to involve commercial companies to enable this type of research to progress, and for the implementation of such technology within the NHS. Panel members noted the increased public awareness and acceptance of commercial collaborations within healthcare research, following the work to develop vaccines during the Covid-19 pandemic. The work involving commercial companies was further clarified in patient facing documents, such as which information commercial companies will have access to and how this access would be controlled. Those who attended the Zoom events kindly helped in further developing the updated versions of patient facing material and all documentation was made available on the University's departmental website.

4.3.4 Database sites

Two sites in East Anglia, England, participated in the creation of the CC-MEDIA imaging database:

- Cambridge (Cambridge University Hospitals NHS Foundation Trust, including Cambridge Breast Unit)
- Norwich (Norfolk and Norwich University Hospitals NHS Foundation Trust, including Norwich Breast Unit)

The average round length for screening at both sites is 34-36 months. Neither site participated in the AgeX trial, however screening was offered to those aged 47 years and older in the region within the study time period.

Cambridge breast screening implements double reading of all mammograms, with the second reader able to see the outcome from the first reader, thus reading is dependant. Arbitration takes place for all cases recalled as well as for cases where there is discordance between the two initial readers, with a panel of up to four readers. During the Covid-19 pandemic Cambridge breast screening was paused twice, once from 23/03/2020 to 16/07/2020, and secondly from 11/01/2021 to 22/02/2021.

Norwich breast screening uses double reading of all mammograms, with the second reader not being able to see the outcome from the first reader, thus reading is independent. Arbitration takes

place only for cases where there is discordance between readers, with an average panel size of five readers. During the Covid-19 pandemic Norwich breast screening was paused from 20/03/2020 to 16/07/2020.

When using the database; the first reader was used as the independent reader to be combined with the algorithm, arbitration was determined if there was a disagreement between readers, and trainee readers were replaced with trained readers to allow comparison using both sites' data.

4.3.5 Database creation

The database consists of cases age greater than or equal to 47 years old, who attended screening between 2011 to 2020. Data was collected at two NHSBSP sites (Cambridge and Norwich) to create a centrally stored database for external testing of multiple AI algorithms. Patients who attended routine screening (triennial) as well as patients who attended high risk screening and subsequently were transferred to routine screening were incorporated into this database. The main sources of information were obtained from; picture archiving and communication system (PACS) for Digital Imaging and Communications in Medicine (DICOM) image data as well as DICOM header and tags, National Breast Screening System (NBSS) for breast screening metadata and Electronic Health Records (EHR) – EPIC / LARDR – for additional clinical metadata, Figure 4-4. A screening episode is defined as the anything that occurs from the time a woman is invited to screening, the screen itself and any assessments, as well as diagnoses and treatments that occur as a consequence of screening. Outcomes for all case episodes were followed up using data from the NBSS until April 2022.

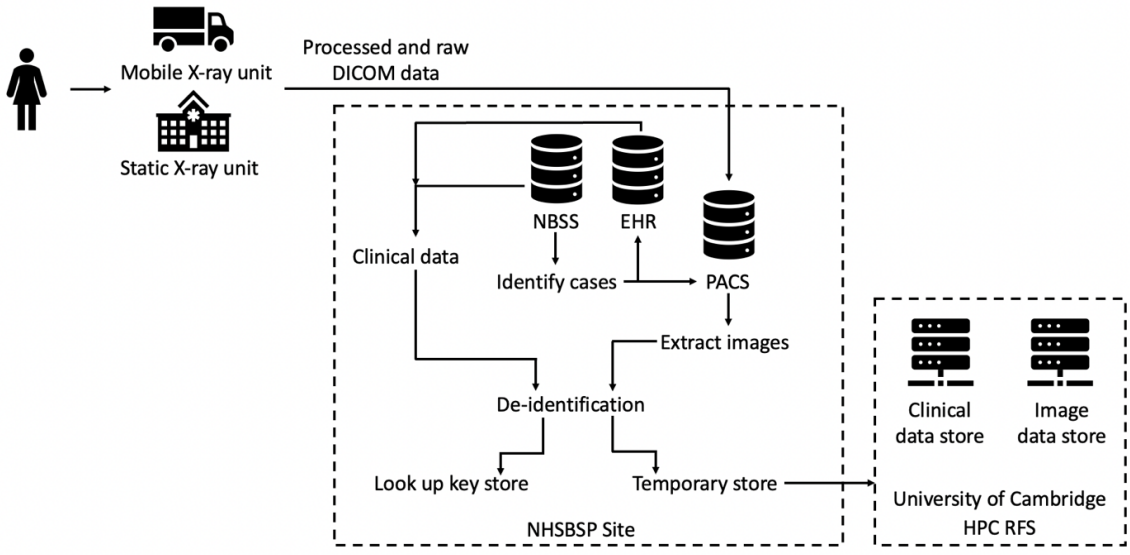


Figure 4-4 – Data flow of the CC-MEDIA data collection. *DICOM: Digital imaging and communications in medicine, EHR: Electronic health records, HPC RFS: High performance computing research file store, NBSS: National breast screening system, NHSBSP: National health service breast screening programme, PACS: Picture archiving and communication system. Adapted from Halling-Brown et al[250].*

An environment was set up within the NHS firewall at each hospital site to facilitate the transfer of images from PACS to a secure store (using code developed by Dr Andrew Priest, Medical Physicist at Cambridge University Hospital NHS Foundation Trust) which entailed; a static IP address, MATLAB (The Mathworks, Inc., Natick, Massachusetts, United States. http://www.mathworks.com, version 2019a) with the Image Processing Toolbox, and the DCMTK tool kit (echoscu, findscu, movescu, storescu) (version 3.6.2 and 3.6.4)[258]. Cases were identified from NBSS using existing and new project specific Crystal Report queries (developed by Sue Hudson, PAS Consulting London). The personal identifiers (NHS number and study date) from NBSS were then used to query PACS and then retrieve the DICOM image data into the on-hospital-site secure store. DICOM imaging data included the standard two-view processed ("for presentation") mammogram screening images as well as available additional views and raw ("for processing") data. All the images were stored in a compressed Joint Photographic Experts Group (JPEG) lossless format. All image data, including the DICOM header and tag information was de-identified by adapting the basic profile provided in DICOM PS3.15, such that all identifiable information was removed (Appendix 12)[259]. Caution was taken when handling latent identifiers, such as date of screening, in order to ensure anonymity was achieved whilst retaining longitudinal information.

Additional NBSS Crystal Report queries were used to extract clinical metadata from NBSS for the whole screening episode. The clinical metadata fields from NBSS provided a ground truth for each case. The clinical metadata was de-identified using Python (Python Software Foundation, http://www/python.org, version 3.8)[260] based scripts (developed by Dr Lorena Escudero Sánchez, Research Associate at University of Cambridge) and Excel (Microsoft Corporation, https://office.microsoft.com/excel) functions within the on-site secure store.

The de-identification processes for both the image data and clinical metadata occurred prior to the transfer of any data from the hospital site. The study nomenclature allows for the easy tracking and re-joining of data for analysis, where each case is assigned a trial ID (case ID), exam ID and image ID, Figure 4-5.

Figure 4-5 – Nomenclature of case de-identification within the CC-MEDIA database. *The site code varied for Cambridge (CC05 or CC06) and Norwich (CC07 or CC08). The trial code was randomly assigned for each case. The exam code increased sequentially for each episode. Series number was taken from the DICOM series tag. MGP: Mammographic processed, MGR: Mammographic raw. LCC: Left craniocaudal, RCC: Right craniocaudal, RMLO: Right mediolateral oblique, LMLO: Left mediolateral oblique.*

All data was then encrypted in the on-hospital-site secure store, using Advance Encryption Standard (AES)-256 encryption, and subsequently transferred to the University of Cambridge high performance computing (HPC) research file store (RFS). A look up key store remained at each site within a separate area in the on-hospital-site secure store to allow for additional data linkage whilst building the database. Following the completion of the database this look up key store will be securely held by the principal investigator at each site.

Once the image data was transferred the DICOM headers and tags were extracted from the image data using the DCMTK dcmdump utility (version 3.6.5)[258]. The DICOM dump data was then adapted into an image metadata file using MATLAB code (developed by Dr Nicholas Payne, Research Associate at University of Cambridge). The DICOM metadata files were then stored in the HPC RFS alongside the de-identified DICOM image data and clinical metadata.

When collecting the image data firstly all interval cancers (ICs) and screen detected cancers (SDCs) from both sites were collected from 2011-2020. Secondly a set of cases that were age and year matched to the ICs in the database were extracted.

All cases from specified year cohorts were consecutively collected, following the completion of the cancer retrieval (ICs and SDCs). The most recent year cohort with a complete follow-up time period

was collected first, hence starting with 2017 from both Cambridge and Norwich. Subsequently year cohorts were collected in a specific order at each site to avoid overlap with existing databases (The Optimam Mammography Image Database (OMI-DB)) where data had previously been collected. Data was collected from Cambridge by the OMI-DB database team between 2012-2016[250].

Thus to date the six sets of data have been extracted are:

- CC05 – Cambridge ICs and year and age matched normal controls 2011-2020
- CC06 – Cambridge SDCs 2011-2020
- CC06 – Cambridge year cohorts 2017-2018
- CC07 - Norwich ICs and year and age matched normal controls 2011-2020
- CC08 – Norwich SDCs 2011-2020
- CC08 – Norwich year cohorts 2014-2018

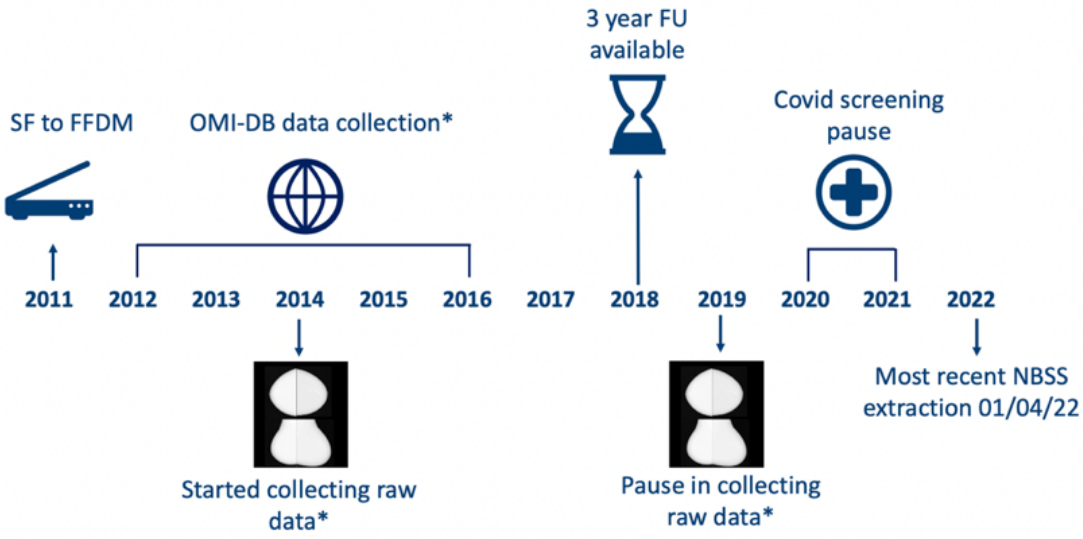


Figure 4-6 – Timeline of mammography data changes over time at Cambridge and Norwich National Health Service Breast Screening Programme (NHSBSP) sites. *SF: Screen film, OMI-DB: The Optimam Mammography Image Database, FU: Follow-up, FFDM: Full field digital mammography, NBSS: National breast screening system. *Only at Cambridge site.*

When using the image cases in studies, first the cohort was identified using the clinical metadata file and then all images for the cases were copied and unencrypted in a separate area on the secure HPC RFS store to retain the completeness of the original data. Figure 4-6 details the important changes at the database sites over the study time period. Due to the change from SF mammograms to FFDM in 2011/2012 at both sites there was limited availability of image data over this time period. In addition, raw data was only collected at Cambridge and only between 2014 and 2019.

4.4 Results

4.4.1 Database image content

Image data collection started on 11/12/2020 and is ongoing. The information reported in this chapter is up to date as of 05/05/2022.

Total	Norwich 2014	Norwich 2015	Norwich 2016	Norwich 2017	Cambridge 2017	Norwich 2018	Cambridge 2018
Exams	27214	28926	25915	22936	18803	26901	21218
Images	116013	122878	107043	94492	151917	110185	171521
Manufacture							
GE	27210 (99.99%)	28926 (100%)	25915 (100%)	22936 (100%)	297 (1.6%)	26901 (100%)	372 (1.8%)
Philips	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	18506 (98.4%)	0 (0.0%)	20846 (98.2%)
Fujifilm	4 (0.01%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
FFDM							
Raw FFDM images	154 (0.13%)	259 (0.21%)	212 (0.20%)	243 (0.26%)	75972 (50.01%)	0 (0.0%)	85735 (49.99%)
Processed FFDM images	115859 (99.87%)	122619 (99.79%)	106831 (99.80%)	94249 (99.74%)	75945 (49.99%)	110185 (100%)	85786 (50.01%)
Breast Implants							
Implant images	768 (0.7%)	1247 (1.0%)	1482 (1.4%)	1169 (1.2%)	1894 (1.3%)	2958 (2.7%)	251 (1.2%)
Age at Screening							
47-49	3033 (11.2%)	2613 (9.0%)	2307 (8.9%)	106 (0.5%)	42 (0.2%)	14 (0.05%)	96 (0.5%)
50-59	10854 (39.9%)	11699 (40.4%)	11553 (44.6%)	10841 (47.2%)	9857 (52.4%)	11887 (44.2%)	10759 (50.7%)
60-69	10211 (37.5%)	12390 (42.8%)	10242 (39.5%)	9441 (41.2%)	7659 (40.7%)	11080 (41.2%)	8235 (38.8%)
70+	3116 (11.5%)	2224 (7.7%)	1813 (7.0%)	2548 (11.1%)	1245 (6.6%)	3920 (14.6%)	2128 (10.0%)
Cancers							
Normal	26882 (98.8%)	28670 (99.1%)	25619 (98.8%)	22662 (98.8%)	18551 (98.7%)	26560 (98.7%)	20947 (98.7%)
SDC	208 (0.8%)	152 (0.5%)	198 (0.8%)	189 (0.8%)	158 (0.8%)	225 (0.8%)	188 (0.9%)
IC	124 (0.5%)	104 (0.4%)	98 (0.4%)	85 (0.4%)	94 (0.5%)	116 (0.4%)	85 (0.4%)

Table 4-2 – Number of exams per site available with images currently held in the CC-MEDIA database. *Interval cancers were diagnosed within 40 months of screening. FFDM: Full field digital mammography, IC: Interval cancer, SDC: Screen detected cancer.*

In total the core database (CC06 and CC08) contains 323,438 images, 40,021 exams, and 39,982 cases from Cambridge, and 550,611 images, 131,892 exams, 87,046 cases from Norwich. Thus in

total the database contains 874,049 images, 171,913 exams, 127,028 cases of which 1,318 are SDC cases, and 706 were IC cases, Table 4-2.

Out of all the exams in the database 82,190 (47.8%) have one instance, 89,606 (52.1%) have two instances, with only a very small proportion having 3 (75 (0.04%)), 4 (32 (0.02%)) and 5 (10 (0.01%)) instances. The age range of cases in the cohort was 47-95 years old (median = 59 years old).

An annual report is created by each screening programme called the KC62 (The NHS Breast Screening Programme Central Return Data Set), which details "information on women invited for Breast Screening, the outcome of the Breast Screening and further information on each cancer detected"[261]. Comparing the volume of cases to the distribution of KC62 data from both sites shows that a similar distribution was collected to the true distribution of cases which attended for screening. Thus the database was representative of the screening carried out at each NHSBSP site, Table 4-3.

	NHSBSP Screened[262]	Cambridge KC62 Screened	Cambridge CC-MEDIA	Norwich KC62 Screened	Norwich CC-MEDIA
2011-2012	1940603	17134	-	25900	-
2012-2013	1970955	17475	-	25798	-
2013-2014	2079271	19590	-	26823	7912*
2014-2015	2105454	21972	-	26070	26133
2015-2016	2161268	19370	-	29150	29065
2016-2017	2199342	18389	4979*	25584	25573
2017-2018	2138434	19035	18970	22471	22286
2018-2019	2234514	20830	16072*	27371	20923*
2019-2020	2123589	15144	-	23675	-
Total exams	18953430	168939	40021	232842	131892

Table 4-3 – Cambridge and Norwich CC-MEDIA database 2011-2020 compared to the KC62 report at both sites. *The KC62 reports programme performance from 01/04/YYYY to 31/03/YYYY at each NHSBSP site. KC62 data is taken from Table-T of the annual KC62 report which reports the sum of tables A-F2; first invite for routine screening, routine invitation to previous non-attenders, return invitation to previous attenders (last screening within 5 years and last screen more than 5 years), short term recall, self / GP referrals for women not previously screened or previously screened (last screen within 5 years or last screen more than 5 years previously). *Fields that have incomplete year data.*

4.4.2 Database content - Interval cancers

ICs are a key measure of screening programme performance. The acceptable IC rate set by the NHSBSP is 3.7/1000 women screened[101,103]. ICs can occur anytime from the last negative screen to 40 months post screen as defined by the NHSBSP. Figure 4-7 shows the time to diagnosis at Cambridge and Norwich by months. IC image data available within the CC-MEDIA database is shown in Table 4-4.

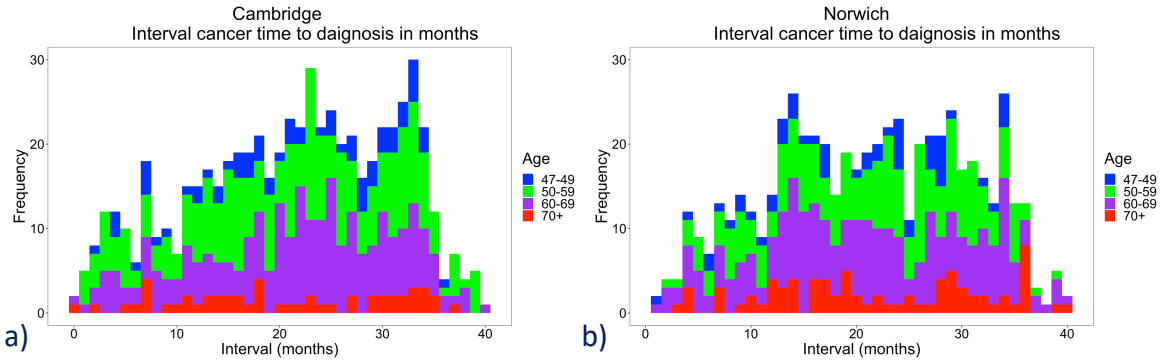


Figure 4-7 – Time to diagnosis (months) for interval cancers (IC) at a) Cambridge and b) Norwich.

	Cambridge n (%)	Norwich n (%)
Total exams n	611	561
Total images n	3937	2350
Year of Screening		
2010-2011	2* (0.3%)	0* (0.0%)
2011-2012	26* (4.3%)	0* (0.0%)
2012-2013	75* (12.3%)	24* (4.3%)
2013-2014	92 (15.1%)	100 (17.8%)
2014-2015	106 (17.3%)	104 (18.5%)
2015-2016	86 (14.1%)	104 (18.5%)
2016-2017	62 (10.1%)	94 (16.8%)
2017-2018	85 (13.9%)	76 (13.6%)
2018-2019	53* (8.7%)	48* (8.6%)
2019-2020	23* (3.8%)	11* (2.0%)
2020-2021	1* (0.2%)	0* (0.0%)
Age at Screening		
47-49	67 (11.0%)	54 (9.6%)
50-59	262 (42.9%)	206 (36.7%)
60-69	230 (37.6%)	229 (40.8%)
70+	52 (8.5%)	72 (12.8%)
Manufacture		
GE	23 (3.8%)	557 (99.3%)
Philips	553 (90.5%)	0 (0.0%)
Hologic	15 (2.5%)	4 (0.7%)
Sectra	20 (3.3%)	0 (0.0%)
FFDM		
Raw FFDM images	1478 (37.5%)	0 (0.0%)
Processed FFDM images	2459 (62.5%)	2350 (100%)
Implants		
Implant images	17 (0.4%)	28 (1.2%)

Table 4-4 – Interval cancers (ICs) at Cambridge and Norwich with imaging data 2011-2020 in CC-MEDIA. *Interval cancers were diagnosed within 40 months of screening, leading to 5 cases excluded from Cambridge and 4 cases from Norwich that were diagnosed > 40 months. FFDM: Full field digital mammography. *Fields that have incomplete year data.*

As shown in the table there is good coverage of image data availability from 2013 to 2018. In addition four different mammographic machine vendors are included over this time period, however the majority are Philips at Cambridge and GE at Norwich. Information regarding IC rates is provided in this database from the NBSS local site data. Ethical approval has been obtained to apply for additional information from the Screening History Information Management system (SHIM) and National Cancer Registry (NCRAS) in the future.

4.4.3 Database content - Screen detected cancers

SDCs that are recalled at the screening episode and diagnosed at the assessment clinic, where a triple assessment is carried out of; clinical examination, further imaging (e.g. ultrasound), and biopsy. It is estimated that in the triennial NHSBSP, SDCs occur at a rate of 8/1000 women screened[54,263]. The SDC image data that is available for each site within the CC-MEDIA database is shown in Table 4-5. As shown in the table there is good coverage of SDC data from 2013 to 2020 at both sites. In addition five different mammographic machine vendors are included over this time period, however the majority are Philips at Cambridge and GE at Norwich.

	Cambridge KC62 n (%)	Cambridge n (%)	Norwich KC62 n (%)	Norwich n (%)
Total exams n	1539	1286	2179	1551
Total images n	-	8327	-	6528
Year of Screening				
2010-2011	123 (8.0%)	1* (0.08%)	202 (9.3%)	0* (0.0%)
2011-2012	148 (9.6%)	52* (4.0%)	204 (9.4%)	0* (0.0%)
2012-2013	131 (8.5%)	129 (10.0%)	186 (8.5%)	52* (3.4%)
2013-2014	148 (9.6%)	143 (11.1%)	180 (8.3%)	168 (10.8%)
2014-2015	162 (10.5%)	168 (13.1%)	201 (9.2%)	203 (13.1%)
2015-2016	166 (10.7%)	163 (12.7%)	188 (8.6%)	187 (12.1%)
2016-2017	144 (9.4%)	148 (11.5%)	201 (9.2%)	197 (12.7%)
2017-2018	162 (10.5%)	161 (12.5%)	198 (9.1%)	200 (12.9%)
2018-2019	184 (12.0%)	196 (15.2%)	249 (11.4%)	247 (15.9%)
2019-2020	97 (6.3%)	96 (7.5%)	202 (9.3%)	207 (13.3%)
2020-2021	74 (4.8%)	29* (2.3%)	186 (8.5%)	90* (5.8%)
Age at Screening	Cambridge n (%) [n = 1286]		Norwich n (%) [n = 1551]	
47-49	74 (5.7%)		69 (4.5%)	
50-59	469 (36.5%)		509 (32.8%)	
60-69	589 (45.8%)		717 (46.2%)	
70+	154 (12.0%)		256 (16.5%)	
Manufacture				
GE	33 (2.6%)		1548 (99.8%)	
Philips	1190 (92.5%)		0 (0.0%)	
Hologic	22 (1.7%)		3 (0.2%)	
Siemens	1 (0.08%)		0 (0.0%)	
Sectra	40 (3.1%)		0 (0.0%)	
FFDM				
Raw FFDM images	3133 (27.6%)		13 (0.2%)	
Processed FFDM images	5194 (62.4%)		6515 (99.8%)	
Implants				
Implant images	45 (0.5%)		25 (0.4%)	

Table 4-5 – Screen detected cancers (SDCs) at Cambridge and Norwich with imaging data 2011-2020 in CC-MEDIA. *The KC62 reports programme performance from 01/04/YYYY to 31/03/YYYY at each NHSBSP site. KC62 data is taken from Table-T of the annual KC62 report which reports the sum of tables A-F2; first invite for routine screening, routine invitation to previous non-attenders, return invitation to previous attenders (last screening within 5 years and last screen more than 5 years), short term recall, self / GP referrals for women not previously screened or previously screened (last screen within 5 years or last screen more than 5 years previously). FFDM: Full field digital mammography. *Fields that have incomplete year data.*

4.4.4 Database content - Ethnicity

Ethnicity information is sparsely available within the NBSS output from Cambridge and no information was available from Norwich, Table 4-6. A similar volume of ethnicity data availability from NBSS was found when searching EPIC the EHR system at Cambridge, Figure 4-8. This limited availability of data meant it was not possible in the studies detailed in Chapters 5-7 to evaluate AI tools for bias relating to ethnicity. In addition, as the data included in this study is only from two

sites in East Anglia, England, it is not representative of the UK population. This is further outlined in the 2011 Census of 25 million households in England and Wales where it was reported "people from the White ethnic group were more likely to live in the South East than any other region"[264].

	Cambridge NBSS [n = 40021]	Cambridge EHR EPIC [n = 83662]
A = White – British	18200 (45.5%)	45424 (54.3%)
B = White – Irish	212 (0.5%)	381 (0.5%)
C = White – Any other White background	770 (1.9%)	2329 (2.8%)
D = Mixed – White and Black Caribbean	27 (0.07%)	32 (0.04%)
E = Mixed – White and Black African	19 (0.05%)	19 (0.02%)
F = Mixed – White and Asian	66 (0.2%)	78 (0.09%)
G = Mixed – Any other Mixed background	35 (0.09%)	129 (0.2%)
H = Asian or Asian British – Indian	125 (0.3%)	350 (0.4%)
J = Asian or Asian British – Pakistani	31 (0.08%)	75 (0.09%)
K = Asian or Asian British – Bangladeshi	21 (0.05%)	55 (0.07%)
L = Asian or Asian British – Any other Asian background	122 (0.3%)	396 (0.5%)
M = Black or Black British - Caribbean	54 (0.1%)	140 (0.2%)
N = Black or Black British - African	59 (0.2%)	181 (0.2%)
P = = Black or Black British – Any other Black background	6 (0.01%)	61 (0.07%)
R = Other ethnic groups – Chinese	176 (0.4%)	420 (0.5%)
S = Other ethnic groups – Any other group	105 (0.3%)	305 (0.4%)
Z = Not stated	257 (0.6%)	5193 (6.2%)
Missing	19736 (49.3%)	28094 (33.6%)

Table 4-6 – Ethnicity information from National Breast Screening System (NBSS) and Electronic Health Record (EHR) EPIC data at Cambridge. *NBSS: National breast screening system, EHR: Electronic health record.*

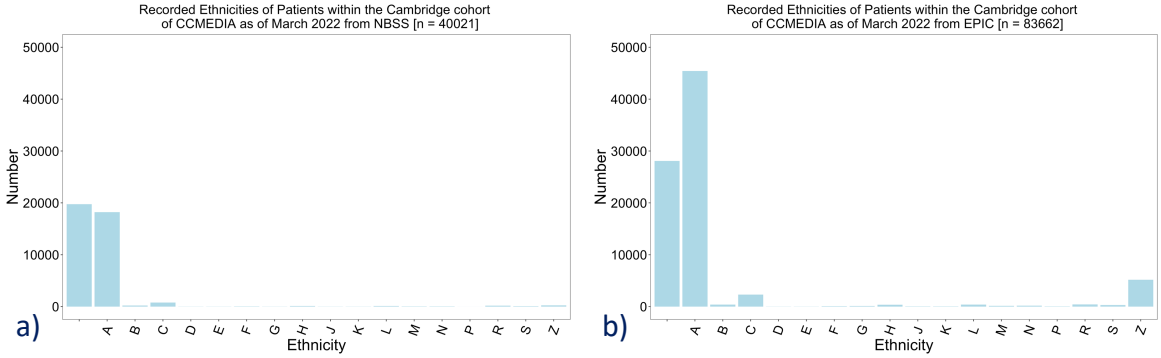


Figure 4-8 – Ethnicity data distribution at Cambridge using National Breast Screening System (NBSS) and Electronic Health Record (EHR) EPIC data. *a) NBSS, b) EPIC EHR. Ethnicity codes are provided in Table 4-6. NBSS: National breast screening system.*

4.4.5 Database content - Mammographic breast density

Density is not routinely reported by readers in the NHSBSP. However, the Breast Imaging-Reporting and Data System (BI-RADS) 5th edition density score was obtained for all cases in the CC-MEDIA database. Raw DICOM data was processed by Volpara (research version - VolparaResearch32_L30Enabled_v2, Wellington, New Zealand) to generate the Volumetric Breast Density (VBD) of each case. The VBD was then converted in Volpara Density Grade, which is consistent with BI-RADS 5th edition. Processed DICOM data was processed by one of the AI algorithm systems (DL-3) used in this research to generate the BI-RADS 5th edition density score for each case. Figure 4-9 shows the distribution of 1 years' worth (2017) of data from Cambridge where both raw and processed data was available.

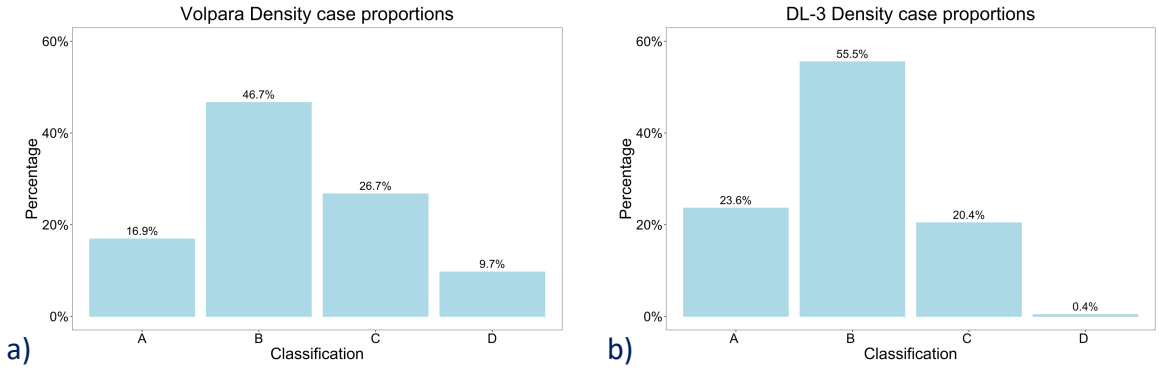


Figure 4-9 – Breast imaging-reporting and data system (BI-RADS) 5th edition mammographic density distribution for cases in one year (2017) of data at Cambridge with both raw and processed four views mammograms available [n = 18246]. *a) Volpara raw density distribution, b) DL-3 processed density distribution. Cases with breast implants were removed from this dataset.*

Demonstrating a similar distribution in the Volpara population mammographic density distribution as per previous publications[75,265]. Whereas the density distribution from DL-3 using processed data shifted the population distribution to the left providing overall lower density assessments for cases.

4.4.6 Database content - Histopathological information

The clinical metadata collected from each site included the invasive status (ICD-10 code), histological grade (assigned using Nottingham grading system), and histological size for cancer cases. Where there were gaps in data, the missing data was collected by hand from histopathology reports on the EHR systems at each site. Histopathological information was taken from the surgical pathology report where available. If the surgical histopathology was unavailable the core biopsy histopathology was used. Information regarding the use of neoadjuvant chemotherapy and hormone therapy was not available alongside this information and so the histopathological size and grade could differ at the time of diagnosis for some cases. Furthermore cancers diagnosed during the Covid-19 pandemic were treated with an increase use of hormone therapy whilst the availability of operations was limited, this would also have an impact on the histopathological size and grade of cancers.

4.5 Technical setup of an AI algorithm testing environment

An AI algorithm testing environment was setup at the University of Cambridge (developed by Richard Black, Medical Physicist at Cambridge University Hospitals NHS Foundation Trust). Two computers were available in this environment with the following technical setup:

- System 1 – OpenSUSE Leap 15.3 operating system, 12 central processing units (CPU), 32 GB random access memory (RAM).
- System 2 - OpenSUSE Leap 15.3 operating system, 56 CPU, 1024 RAM, 3 NVIDIA Quadro RTX 8000 graphics cards.

On both systems the following software was installed to allow for company installation as well as data processing; Teamviewer, Docker, dcmtk 3.6.5, libvirtd v7.1.0, and qemu-kvm v5.2.0 (virtualisation).

4.6 Uses of the database

To date the database has been used for the following research applications. Those applications with an asterisk (*) next to them are the applications detailed in the remaining chapters of this thesis, the remaining applications are part of ongoing work by other researchers.

- *Benchmark existing AI algorithms for interval and next round cancer detection – Chapters 5, 6 and 7
- *Benchmark existing AI algorithms for stand-alone cancer detection – Chapter 6
- *Benchmark existing AI algorithms for screening triage – Chapter 7
- *To assess the relationship between AI algorithm accuracy and mammographic breast density – Chapters 5, 6 and 7

- To evaluate the accuracy of AI algorithm prompt location for the detection of cancer
- To evaluate breast density tools for both raw and processed data
- To evaluate breast cancer screening risk stratification tools
- To evaluate the impact of prior image availability on AI algorithm performance

4.7 Discussion

4.7.1 Overall discussion

Developing a large multi-site mammographic imaging database is a complex task, involving numerous governance, approvals and technical setup requirements. The involvement of patients and the public in the setup highlighted the importance of clear communication regarding access and processing of data as well as the acceptability of using data without consent and with commercial collaborators for this type of research. In addition, the formation of the DAC means the data is treated with a high level of governance oversight from staff with expertise at both sites to ensure the security and correct use of the data in research. The systematic collection of a large representative cohort for breast cancer screening provides an extensive resource for AI algorithm benchmarking as well as for feedback to AI companies regarding their performance to allow for the further development of algorithms. The availability of SDCs as well as ICs and next round cancers (NRCs) over the ten-year study time period allows for the robust assessment of algorithms for the detection of cancers as well as the potential for the earlier detection of cancer. Using this database AI algorithms can be tested for numerous applications including stand-alone detection and normal case triage in a UK screening setting. Another advantage of the database is the inclusion of raw data at one site, allowing for the calculation of mammographic breast density which is not routinely reported within the NHSBSP. This database is of similar size to recently developed databases in the UK, USA and Sweden, and overcomes the limitation of early mammographic databases which were small in size and only contained screen film mammography.

4.7.2 Limitations

However, this database is limited to East Anglia and thus not representative of the entire UK population in terms of demographics. Furthermore, there is limited availability of ethnicity information at both sites to provided sufficient data for subgroup analysis to evaluate AI algorithms performance in order to detect bias. In addition, this database does not have any image level / pixel level annotations at present and so it is not possible to evaluate the precision of AI algorithm prompt locations which are provided alongside continuous case score outputs. Lastly, the overlap with OMI-DB is required to be taken into account when selecting cases for algorithm testing, by removing

cases identified as being extracted into OMI-DB, as these cases may have been used for model training.

4.8 Conclusion

The CC-MEDIA database is a large 127,000 case mammographic medical imaging database that is representative of the NHSBSP in case distribution. The clinical metadata available provides a robust method to identify the ground truth for different cases cohorts when testing various applications of AI algorithms in breast cancer screening. The governance of the database by the DAC ensures the security of the data and that robust protocols are followed when sharing data. Collecting data from a ten-year period provides sequential screening information which is vital for testing numerous applications of AI algorithms for breast cancer screening. However, this data is limited to one region of the UK only and thus is not completely representative of diverse UK population in terms of ethnicity and socio-economic factors.

Chapter 5 – Performance of artificial intelligence algorithms for interval cancer detection

5.1 Aims

In this chapter the performance of three commercial AI algorithms is investigated for the detection of interval cancers, using an enriched dataset from two UK screening sites. This study evaluated the potential benefit from AI algorithms for the earlier detection of breast cancer. Interval cancer literature and screen programme standards were used to pre define thresholds for the AI algorithms operating points and all algorithms were tested independent of the commercial vendor. The results from this study will help the planning of both retrospective and prospective studies for the use of AI algorithms as stand-alone readers.

Contents of this chapter have been presented at the Radiological Society of North America conference 2021 [abstract ID - #2021-SP-12762-RSNA] and accepted for presentation at the European Congress of Radiology 2022 [abstract number - #12040].

5.2 Introduction

Breast cancer screening programmes aim to detect breast cancer at an earlier stage when the cancer is asymptomatic, which has been shown to improve both morbidity and mortality outcomes[266]. Interval cancers (ICs) occur in the time period between screening rounds. In the UK, operating a triennial programme, the acceptable IC rate is set at 3.7/1000 women screened[101,103]. Overall the survival outcomes of ICs are worse than screen detected cancers[102]. It is estimated ~77% ICs could not be seen at screening (normal / benign), ~16% have minimal signs (uncertain) and ~7% were visible (suspicious)[103]. Duty of candour is defined as a healthcare professionals responsibility to be "honest with patients and people in their care when something that goes wrong with their treatment or care causes, or has the potential to cause, harm or distress", thus all ICs classified as false negative (suspicious) in the UK programme at the IC audit are required to be disclosed to patients[103,267]. There are numerous reasons ICs are not be detected at screening. These include not present at time of screening and developed in the interval, low sensitivity of mammography (especially in dense breasts due to masking), cancer radiological appearance (this can be either a stable appearance or the signs can be minimal on mammography), and perception or interpretation error (either not seen or seen and dismissed)[268,269].

Artificial intelligence (AI) algorithms for detection and diagnosis tasks (CADe+x) have demonstrated good performance for screen detected cancers (SDCs)[133]. However, as highlighted in the 2021 UK National Screening Committee (NSC) report the use of AI systems for IC detection is scarce,

especially for UK data[136]. Lång *et al*, tested one AI algorithm (Transpara v1.5.0) using a dataset of 429 ICs from five years of Swedish screening data, and found the AI algorithm could detect 11.2% of potentially visible cancers at the previous screen at a 4.0% recall rate[270]. In addition, 28.4% of minimal sign or false negative cancers were correctly located by the AI prompts. Of the ICs detected at a risk score 10 (the highest category score) 23.0% patients died or they developed stage IV disease and thus were clinically significant cancers[270]. Larsen *et al* tested an updated version of the AI algorithm (v1.7.0) used in Lång *et al*, and applied it to a large Norwegian dataset of more than 47,000 women, containing 205 ICs[271]. Larsen *et al* found 44.9% of ICs were detected at a risk score of 10 (a 10.0% recall rate), and 30.7% at a 5.8% recall rate. Hinton *et al* applied a ResNet50 architecture algorithm to a dataset of 182 ICs diagnosed within 12 months of screening, with an age and race matched screen detected cancer dataset of 173 cancers, from nine years of US screening. They found an accurate classification of 74.0% for ICs and 77.0% for SDCs[272]. Dembrower *et al*, tested an AI algorithm (Lunit v5.5.0) using a dataset of seven years of Swedish screening data with 7364 women which included 200 ICs. They found 12.0-27.0% of ICs had the highest 1.0-5.0% of scores, and the AI score was shown to be a better predictor than automated breast density (LIBRA) for the detection of ICs, OR 2.01 [95% CI 1.98-2.18] and 1.59 [95% CI 1.50-1.68] respectively[134]. Other studies have also included ICs within their datasets but have either not reported the separate performance for ICs or the dataset was small in size[138,273].

This study aimed to provide evidence for the use of AI algorithms for IC detection with UK screening data. In addition, this study aimed to evaluate three commercial AI algorithms using the same large unseen dataset to carry out independent performance benchmarking.

5.3 Methods

5.3.1 Sample size

The required sample size for this study was calculated using the method described in Arkin *et al,* to determine the minimum number of cases required to estimate the true performance of an algorithm for benchmarking[274]. As described in the literature it is estimated 23.0% of ICs were visible at the previous screening (false negatives – suspicious / uncertain) and therefore a reference proportion of 20.0% and 30.0% was used[103]. Applying these reference proportions and a 95.0% confidence interval, between 246 - 323 cancers were required for this study.

5.3.2 Data

Patient data was obtained from the existing CC-MEDIA database described in Chapter 4, where data was collected from two National Health Service Breast Screening Programme (NHSBSP) sites

(Cambridge and Norwich) under existing ethical approval (HRA REC 20/LO/0104, HRA CAG 20/CAG/0009, PHE RAC BSPRAC_090).

Women age greater than or equal to 47 years old who attended screening at either site were included. IC cases were identified using the existing cancer registry (CREGX) query on the National Breast Screening System (NBSS) from January 2011 to December 2020 at Cambridge, and January 2011 to May 2021 at Norwich. A python (Python Software Foundation, http://www/python.org, version 3.8)[260] script was used to query a database of all women screened at each site, to randomly select three age and screening year matched controls to every IC case. The two-view screening Full Field Digital Mammography (FFDM) images for each case were used. Cases were excluded where they did not include the full four views, and as per each companies' manufacturer protocol images containing an implant, pacemaker or other device were excluded. Cases were also excluded following a discussion with Public Health England (PHE) if the IC was not a primary breast cancer (e.g. mesothelioma, melanoma, colorectal cancer metastasis). IC radiological classifications were taken from the original screen reader IC audit. Where histopathological data was missing, this was hand searched for using Electronic Health Records (EHR) at each site, for further detail please see Chapter 4 Section 4.4.5. The case selection process is shown in the Standards for Reporting of Diagnostic Accuracy Studies (STARD) diagram in Figure 5-1[275].

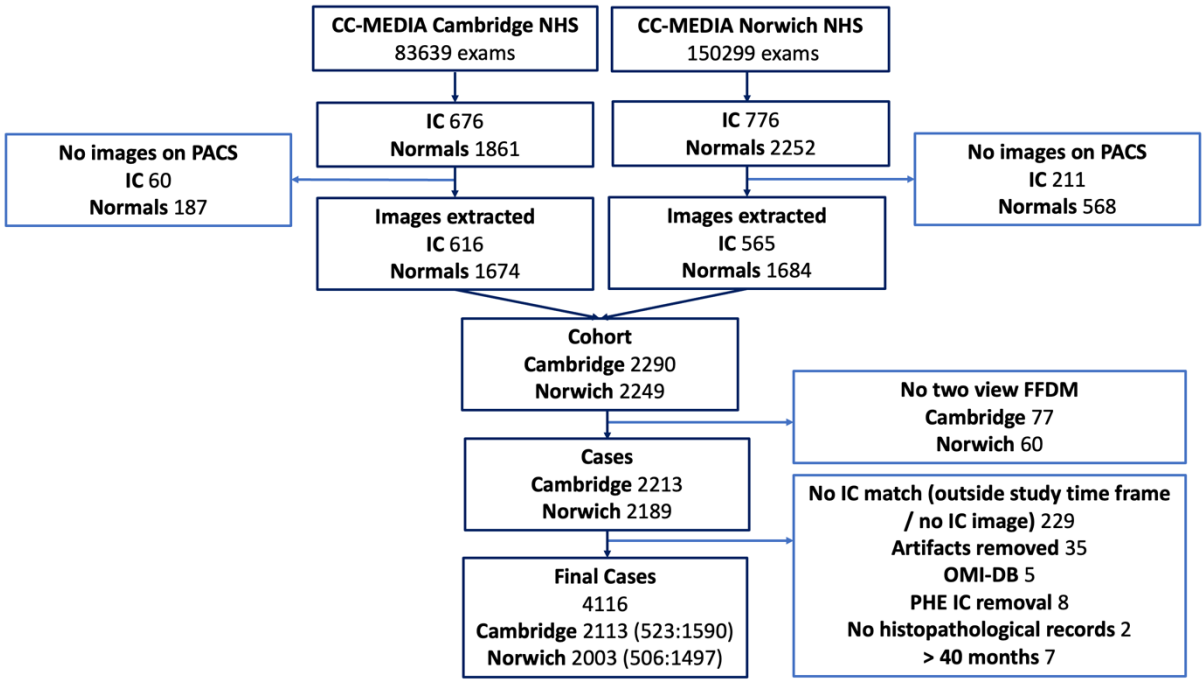


Figure 5-1 – Standards for Reporting of Diagnostic Accuracy Studies (STARD) flow diagram of cases included and excluded in this study. *FFDM: Full Field Digital Mammogram, IC: Interval cancer, NHS: National Health Service, OMI-DB: The Optimam Mammography Image Database, PHE: Public Health England, PACS: Picture Archiving and Communication System.*

5.3.3 Ground truth

The ground truth for an IC case was a confirmed histopathological diagnosis, within 40 months of screening, as per the NHSBSP definition[101]. ICs were classified by radiologists as part of the routine IC audit using the NHSBSP definitions, which were updated in August 2017[101]:

- Satisfactory - Normal – "normal or benign mammographic features" and "readers found no reason to recall".
- Satisfactory with learning points - Uncertain – "seen with hindsight, difficult to perceive, not obviously malignant" and "not all readers would recall. Case may provide learning".
- Unsatisfactory – Suspicious – "appearance is obviously malignant" with "all readers reviewing the images agree that they would recall. Woman should have been recalled".

A case was classified as 'normal' if there was a routine recall from screening, more than 912 days after their initial screen, and no breast cancer was detected in this time period. Figure 5-2 provides an overview of the cases included and examples of different IC cases included.

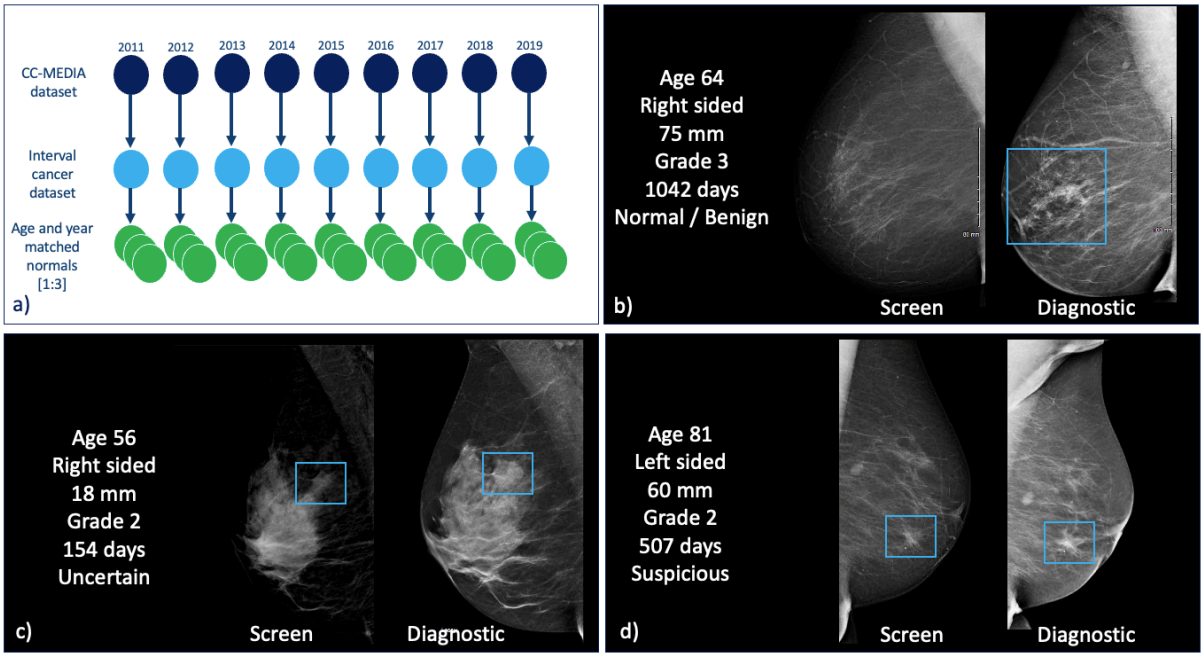


Figure 5-2 – Example of cases included in the study. *a) Interval cancer cases selected from CC-MEDIA cohort were matched at a ratio of 1:3 with normal cases based on year of screen and age at screen. Only cases from 2011-2019 were included due to the follow-up time period required of 912 days and so only cases up until 2019 could be included as screening data was available until end of 2020 at Cambridge and mid 2021 at Norwich, b) an example of a normal / benign classified interval cancer case, c) an example of an uncertain interval cancer case, and c) an example of a suspicious interval cancer case.*

5.3.4 AI tools

Three commercial AI algorithms were independently tested. Each tool was installed within the University of Cambridge research environment, and companies did not have access to their tools

during testing or the results from the study. Details of each algorithms training, required input and output as well as operating system are outlined in Table 5-1.

Tool	DL-1	DL-2	DL-3
Training Screening Programme Readers - Frequency (%UK)	Double – triennial Double – biennial Single – annual (10.0%)	Double – biennial (0.0%)	Double – triennial Single - biennial Single – annual (4.3%)
OMI-DB	Yes	No	Yes
Training Cases n	>200000	>200000	>150000
Training Cases Age Range	40-74	50-70	18-90
Training Cancers	SDC / IC / NRC	SDC	NA
Training Cases Vendors (%)	Hologic (80.0%) GE (10.0%) Siemens (10.0%)	Hologic (41.0%) GE (5.0%) Siemens (36.0%) Philips (< 1.0%) Fuji (7.0%) Agfa (6.0%) Kodak (4.0%)	Hologic (32.3%) GE (65.6%) Siemens (1.7%) Philips (0.3%)
Data	Processed FFDM	Processed FFDM	Processed FFDM
OS	Lunix	Lunix	Lunix
Output	Case level Continuous score (0-10)*	Case level Continuous score (0-10)*	Case level Continuous score (0-10)*

Table 5-1 – Artificial intelligence (AI) algorithm characteristics. *FFDM: Full field digital mammogram, GE: General Electric, IC: Interval cancer, NA: Not available, NRC: Next round cancer, OS: Operating System, OMI-DB: The Optimam Mammography Image Database, SDC: Screen detected cancer. *Output scores were adjusted to the same 0-10 scale.*

5.3.5 Thresholds

Three different methods for identifying thresholds were used in this study and were all based on using the AI algorithms at either a 96.0% specificity (NHSBSP consensus specificity) or 30.0% sensitivity (estimated visible IC rate), for use as stand-alone system for IC detection, Figure 5-3.b. The first threshold is the 'pre-specified specificity / sensitivity' (threshold 1) where the tools are operated at the pre-defined operating points of 96.0% specificity or 30.0% sensitivity within the study data for each site. The second threshold is the 'identified year operating points' (threshold 2) for each algorithm, which were found using 10,206 cases (229 cancers (150 SDCs and 79 ICs)) of Cambridge 2017 data from the main CC-MEDIA database. Both the 'pre-specified specificity / sensitivity' (threshold 1) and 'identified year operating points' (threshold 2) thresholds were then applied to the Cambridge and Norwich data in this study. The last threshold was the 'identified Cambridge operating points' (threshold 3) for each algorithm, where the operating point was identified on the study Cambridge data and then applied to the study Norwich data.

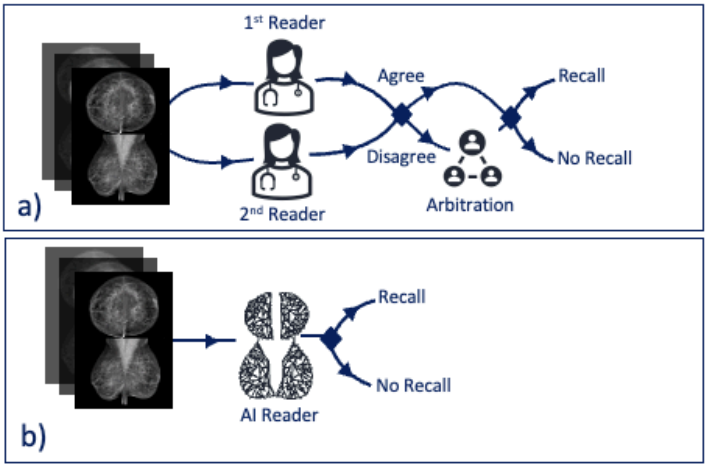


Figure 5-3 – Proposed workflow image for testing the artificial intelligence (AI) systems as stand-alone readers for interval cancer (IC) detection. *a) Routine UK double reading workflow, b) stand-alone artificial intelligence algorithm reading at 96.0% specificity and 30.0% sensitivity thresholds workflow.*

5.3.6 Statistical analysis

All statistical analysis took place in R (R Foundation for Statistical Computing, Vienna, Austria, version 4.0.4)[225], using packages: ggplot2, dplyr, tidyr, lme4, pROC, precrec, lubridate, epiR, data.table and VennDiagram[276–284]. The overall predictive performance of each AI algorithm was evaluated by calculating the area under the receiver operating characteristic curve (AUROC), proportion of true positive (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity and specificity.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

To investigate the variability between sites and mammography machine vendors the results from each site are reported separately. Data is presented as integer number and percentage (*n* (%)), or median and interquartile range [IQR 25th – 75th centile range] as appropriate.

A multivariable model was created for a combination of all three individual algorithms using a generalised linear mixed effects model and Cambridge data. This model was then applied to Norwich data to check for overfitting. DeLong's test was used to assess for a statistically significant difference between the AUROC curve of individual AI algorithms using 2000 bootstrapping examples.

Subgroup analysis for each algorithm based on IC detection at different categories of; age, radiological classifications, time interval to diagnosis (months), mammographic machine vendor, invasive status of cancer, invasive tumour size, invasive tumour grade and mammographic breast density was performed using both Cambridge and Norwich data. The true integer values and

sensitivity were reported as well as Chi squared χ^2 test was used to investigate if there was a statistically significance between categories[285]. In all analyses, 95.0% confidence intervals are used and p-values < 0.05 were considered statistically significant.

5.3.7 Reporting

Each AI algorithm was assigned a Deep Learning (DL) Identifier (ID) for the purposes of this study. The de-identified results of all algorithms were reported back to the companies prior to publication. The individual companies' results were re-identified and presented back to each company for their own performance; the companies could not alter any reporting or methods used.

5.4 Results

5.4.1 Data

In total 8,452 images from Cambridge and 8,012 images from Norwich were included in the study dataset. 2,113 cases from Cambridge contained 523 IC cases (24.8%), and 2,003 cases from Norwich contained 506 IC cases (25.3%). Study case cohort characteristics are provided in Table 5-2.

		Cambridge Normal Cases n (%)	Cambridge Interval Cancers n (%)	Norwich Normal Cases n (%)	Norwich Interval Cancers n (%)
Total Cases n		2113		2003	
Age at Screening	Median	57.0 [51.0-64.0]		60.0 [54.0-67.0]	
	47-49	187 (8.8%)	62 (2.9%)	145 (7.2%)	52 (2.6%)
	50-54	430 (20.4%)	142 (6.7%)	258 (12.9%)	89 (4.4%)
	55-59	261 (12.4%)	89 (4.2%)	308 (15.4%)	105 (5.2%)
	60-64	318 (15.1%)	107 (5.1%)	222 (11.1%)	77 (3.8%)
	65-69	298 (14.1%)	91 (4.3%)	390 (19.5%)	129 (6.4%)
	70+	96 (4.5%)	32 (1.5%)	174 (8.7%)	54 (2.7%)
Year of Screen	2011	8 (0.4%)	7 (0.3%)	0 (0.0%)	0 (0.0%)
	2012	218 (10.3%)	70 (3.3%)	10 (0.5%)	13 (0.7%)
	2013	231 (10.9%)	72 (3.4%)	166 (8.3%)	66 (3.3%)
	2014	313 (14.8%)	105 (5.0%)	353 (17.6%)	115 (5.7%)
	2015	264 (12.5%)	87 (4.1%)	293 (14.6%)	97 (4.8%)
	2016	198 (9.4%)	64 (3.0%)	283 (14.1%)	91 (4.5%)
	2017	238 (11.3%)	76 (3.6%)	237 (11.8%)	77 (3.8%)
	2018	120 (5.7%)	42 (2.0%)	153 (7.6%)	46 (2.3%)
	2019	0 (0.0%)	0 (0.0%)	2 (0.1%)	1 (0.05%)
FFDM Vendor	GE	26 (1.2%)	16 (0.8%)	1490 (74.4%)	502 (25.1%)
	Philips	1501 (71.0%)	473 (22.4%)	0 (0.0%)	0 (0.0%)
	Hologic	26 (1.2%)	15 (0.7%)	7 (0.3%)	4 (0.2%)
	Sectra	37 (1.8%)	19 (0.9%)	0 (0.0%)	0 (0.0%)
Density BI-RADS$^{\alpha}$	a	342 (16.2%)	46 (2.2%)	224 (11.2%)	20 (1.0%)
	b	875 (41.4%)	231 (10.9%)	894 (44.6%)	238 (11.9%)
	c	368 (17.4%)	236 (11.2%)	361 (18.0%)	232 (11.6%)
	d	5 (0.2%)	10 (0.5%)	18 (0.9%)	16 (0.8%)

Table 5-2 – Summary of testing dataset characteristics. *Integer values with percentages in brackets (%) and median with Interquartile range in square brackets [IQR] are provided. BI-RADS: Breast imaging-reporting and data system, FFDM: Full Field Digital Mammography, GE: General Electric. $^{\alpha}$DL-3 5th edition BI-RADS density scores on processed Full Field Digital Mammograms.*

The FFDM images were from 48.0% Philips's, 49.4% GE, 1.3% Hologic, and 1.4% Sectra mammography machines. The median age of the entire cohort was 59.0 [IQR 53.0–65.3] years old and the median time interval between screening and follow-up normal recall as 1071.0 [IQR 1041.0–1105.0] days. IC cases had a median time interval from screening to diagnosis of 690.0 [IQR 465.0–911.0] days at Cambridge and 670.5 [IQR 434.2–880.8] days at Norwich. The majority of cases (78.4%) were classified as normal / benign, with 16.6% assigned uncertain and 3.3% suspicious classification at the routine IC audit. IC characteristics are provided in Table 5-3 and Table 5-4.

		Cambridge Interval Cancers n (%)	Norwich Interval Cancers n (%)
Total Cases n		523	506
Interval (months)	0-12	85 (16.3%)	83 (16.4%)
	12-24	205 (39.2%)	201 (39.7%)
	24-36	233 (44.5%)	222 (43.9%)
	36-40	0 (0.0%)	0 (0.0%)
Radiological Audit Classification	Normal / Benign	429 (82.0%)	382 (75.5%)
	Uncertain	80 (15.2%)	92 (18.2%)
	Suspicious	8 (1.5%)	27 (5.3%)
	Unclassifiable	5 (1.0%)	5 (1.0%)
	Missing	1 (0.2%)	0 (0.0%)
Density BI-RADS$^{\beta}$	a	20 (3.8%)	-
	b	110 (21.0%)	-
	c	114 (21.8%)	-
	d	76 (14.5%)	-
	Missing	203 (38.8%)	-

Table 5-3 – Interval cancer (IC) characteristics by case. *Integer values with percentages in brackets (%) are provided. BI-RADS: Breast imaging-reporting and data system. $^{\beta}$ Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms Cambridge data.*

One AI algorithm (DL-3) provided a density score based on processed data for the entire cohort which was used in study analysis, Table 5-2. Volpara mammographic breast density (research version – VolparaResearch32_L30Enabled_v2, Wellington, New Zealand) was only available for Cambridge cases, where the raw mammographic data was available (67.3% of Cambridge cases), Table 5-3.

		Cambridge Interval Cancers n (%)	Norwich Interval Cancers n (%)
Total Lesions n		535	519
Invasive Status	Invasive	474 (88.6%)	490 (94.4%)
	Non-invasive	55 (10.3%)	29 (5.6%)
	Missing	6 (1.1%)	0 (0.0%)
Invasive Tumour Size$^{\delta}$ (mm)	< 15	114 (24.1%)	146 (29.8%)
	>= 15	278 (58.6%)	288 (58.8%)
	Missing	82 (17.3%)	56 (11.4%)
Invasive Tumour Grade$^{\delta}$	1	49 (10.3%)	65 (13.3%)
	2	223 (47.0%)	231 (47.1%)
	3	183 (38.6%)	180 (36.7%)
	Missing	19 (4.0%)	14 (2.9%)

Table 5-4 – Interval cancer (IC) characteristics by lesions. *Integer values with percentages in brackets (%) are provided. $^{\delta}$Invasive lesions only.*

5.4.2 Algorithm results

The area under the receiver operating characteristic curve (AUROC) was 0.710 [95% CI 0.691–0.730], 0.713 [95% CI 0.695–0.732], 0.732 [95% CI 0.715–0.750] for DL-1, DL-2, and DL-3 respectively when

testing on the entire cohort. When tested on Cambridge data the AUROC was 0.719 [95% CI 0.692–0.746], 0.723 [95% CI 0.698–0.748], 0.726 [95% CI 0.701–0.752], and on Norwich data was 0.713 [95% CI 0.686–0.740], 0.704 [95% CI 0.677–0.730], 0.760 [95% CI 0.736–0.784] for DL-1, DL-2, and DL-3 respectively.

ROC curve plots for comparison between sites is shown in Figure 5-4.a. and between AI algorithms in Figure 5-4.b. All algorithms perform similarly on Cambridge and Norwich data. However, the AUROC of DL-3 is statistically significantly greater than DL-1 and DL-2 when tested on all and Norwich data ($p < 0.05$).

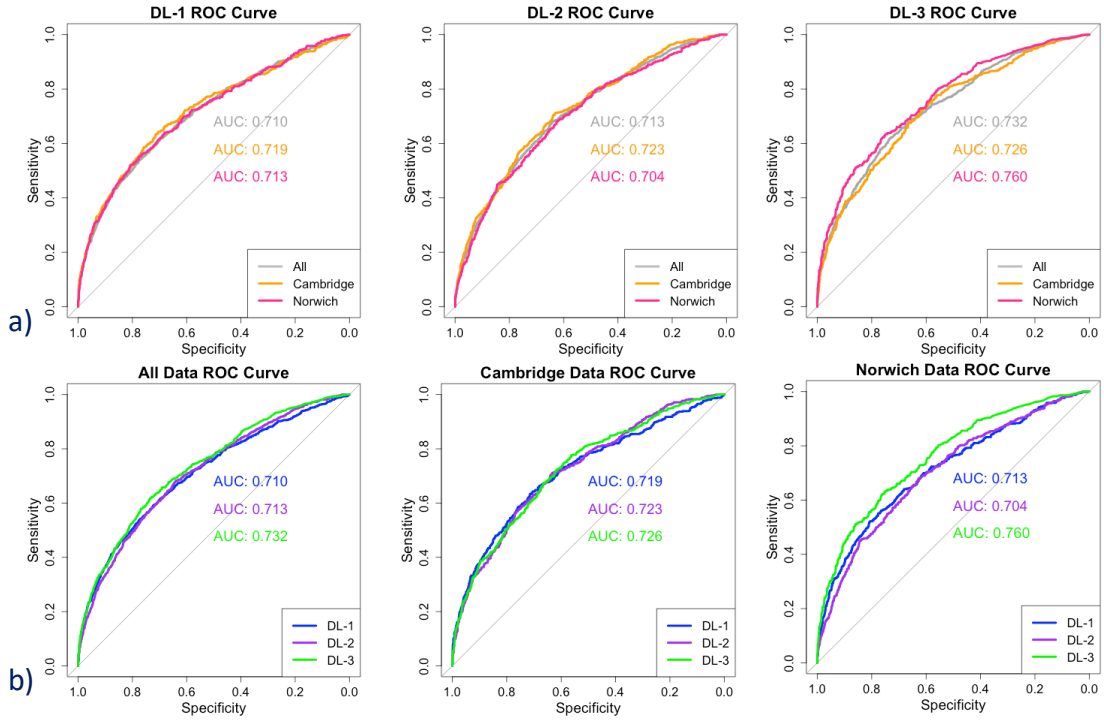


Figure 5-4 – Receiver operating characteristic (ROC) curves for all three artificial intelligence (AI) algorithms at each site. *a) For each artificial intelligence algorithm with the overall results in grey, Cambridge in orange and Norwich in pink, b) for each site with the results for DL-1 are in blue, DL-2 in purple, and DL-3 in green.*

Testing using the 'pre-specified specificity / sensitivity' (threshold 1) thresholds on Cambridge data at 96.0%, specificity, found a sensitivity of 23.7% , 21.6%, 23.1% and at 30.0% sensitivity specificity was 93.8% , 93.1%, 93.0% for DL-1, DL-2, and DL-3 respectively, results are shown in Table 5-5.

Threshold	a) Sensitivity	a) Specificity	b) Sensitivity	b) Specificity
96.0% specificity (DL-1)	23.7% [19.0-28.7]	96.0%	21.6% [17.0-26.4]	96.7% [95.3-97.9]
96.0% specificity (DL-2)	21.6% [17.2-26.4]	96.0%	21.8% [17.0-26.4]	96.0% [94.1-97.3]
96.0% specificity (DL-3)	23.1% [17.8-27.3]	96.0%	20.8% [16.3-26.0]	96.5% [95.1-97.7]
30.0% sensitivity (DL-1)	30.0%	93.8% [91.8-95.6]	2.9% [1.9-9.6]	99.9% [99.7-100]
30.0% sensitivity (DL-2)	30.0%	93.1% [90.5-94.8]	2.1% [1.5-5.2]	100% [99.9-100]
30.0% sensitivity (DL-3)	30.0%	93.0% [90.9-95.0]	0.4% [0.2-5.6]	100% [99.9-100]

Table 5-5 – Cambridge data testing of three artificial intelligence (AI) algorithms. *a) At the 'pre-specified specificity / sensitivity' (threshold 1) for 96.0% specificity, and 30.0% sensitivity, b) at the 'identified year operating points' (threshold 2) from Cambridge external year cohort testing. 95.0% confidence intervals are in square brackets [95.0% CI].*

Applying the 'identified year operating points' (threshold 2) on Cambridge 2017 data at 96.0% specificity, found a specificity of 96.7%, 96.0% and 96.5%, and sensitivity of 21.6% , 21.8%, 20.8% respectively for DL-1, DL-2, and DL-3. At 30.0% sensitivity DL-1, DL-2, and DL-3 specificity was 99.9% , 100.0%, 100.0% and sensitivity was 2.9%, 2.1%, 0.4% respectively. Figure 5-5 shows the distribution of IC cases and normal cases from Cambridge data by the assigned continuous score for each AI algorithm with the four different operating points used in this study.

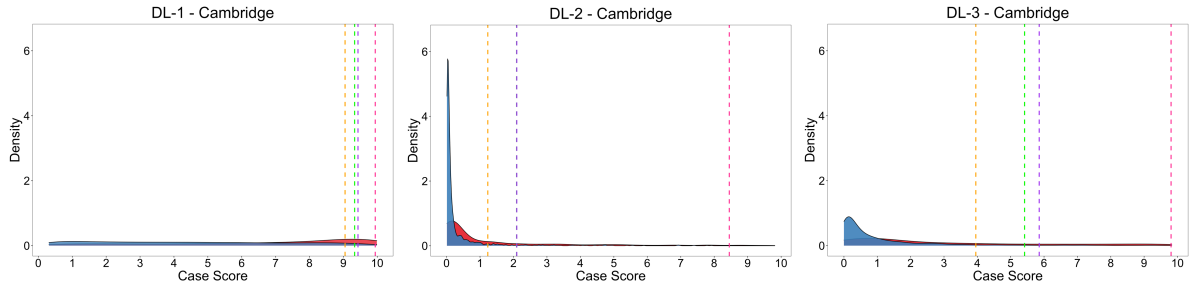


Figure 5-5 – Cambridge data testing density plots for each artificial intelligence (AI) algorithm. *Interval cancer case distribution is shown in red and normal case distribution is in blue. The green line represents the 'pre-specified specificity / sensitivity' (threshold 1) 96.0% specificity operating point for each algorithm and the orange line the 30.0% sensitivity operating point on Cambridge study data. The purple line represents the 'identified year operating points' (threshold 2) 96.0% specificity operating point for each algorithm and the pink line the 30.0% sensitivity operating point.*

Applying the 'pre-specified specificity / sensitivity' (threshold 1) thresholds on Norwich data at 96.0%, specificity, the sensitivity was 23.3%, 16.4%, 27.9%. At 30.0% sensitivity DL-1, DL-2, and DL-3 specificity was 94.1% , 91.2%, 95.4% respectively, the results are shown in Table 5-6.

Threshold	a) Sensitivity	a) Specificity	b) Sensitivity	b) Specificity	c) Sensitivity	c) Specificity
96.0% Specificity (DL-1)	23.3% [18.4-29.1]	96.0%	36.8% [31.8-42.3]	90.0% [87.2-92.5]	39.3% [34.2-44.7]	88.6% [86.0-91.3]
96.0% Specificity (DL-2)	16.4% [13.0-21.0]	96.0%	16.2% [12.9-20.1]	96.3% [94.5-97.9]	16.2% [12.9-20.1]	96.3% [94.5-97.9]
96.0% Specificity (DL-3)	27.9% [22.7-32.8]	96.0%	13.8% [9.9-18.2]	99.0% [98.2-99.7]	14.8% [11.1-20.0]	98.8% [98.1-99.6]
30.0% Sensitivity (DL-1)	30.0%	94.1% [91.0-95.7]	6.5% [2.8-11.7]	99.5% [99.1-99.9]	47.4% [42.7-52.6]	83.4% [79.8-87.2]
30.0% Sensitivity (DL-2)	30.0%	91.2% [88.6-93.4]	2.6% [0.0-5.9]	99.9% [99.7-100]	24.1% [18.8-28.7]	93.7% [91.4-95.3]
30.0% Sensitivity (DL-3)	30.0%	95.4% [92.9-97.0]	1.0% [0.2-2.0]	100% [100-100]	20.6% [12.6-25.3]	98.0% [96.7-98.7]

Table 5-6 – Norwich data testing of three artificial intelligence (AI) algorithms. *a) At the 'pre-specified specificity / sensitivity' (threshold 1) for 96.0% specificity, and 30.0% sensitivity, b) at the 'identified year operating points' (threshold 2) from Cambridge external year cohort testing, c) at the 'identified Cambridge operating points' (threshold 3) from Cambridge data in this study. 95.0% confidence intervals are in square brackets [95.0% CI].*

Testing using the 'identified year operating points' (threshold 2) on Norwich data at 96.0% specificity, the specificity of each AI algorithm was 90.0%, 96.3% and 99.0%, and the sensitivity was 36.8%, 16.2%, 13.8% for DL-1, DL-2, and DL-3 respectively. At 30.0% sensitivity DL-1, DL-2, and DL-3 specificity was 99.5%, 99.9%, 100% and sensitivity was 6.5%, 2.6%, 1.0% respectively.

Applying the 'identified Cambridge operating points' (threshold 3) on Norwich data at 96.0% specificity, the specificity was 88.6%, 96.3% and 98.8%, and the sensitivity was 39.3%, 16.2%, 14.8% respectively for DL-1, DL-2, and DL-3. At 30.0% sensitivity DL-1, DL-2, and DL-3 specificity was 83.4%, 93.7%, 98.0% and sensitivity was 47.4%, 24.1%, 20.6% respectively.

Figure 5-6 shows the distribution of IC cases and normal cases from Norwich data by the assigned score for each AI algorithm with the six different operating points used in this study.

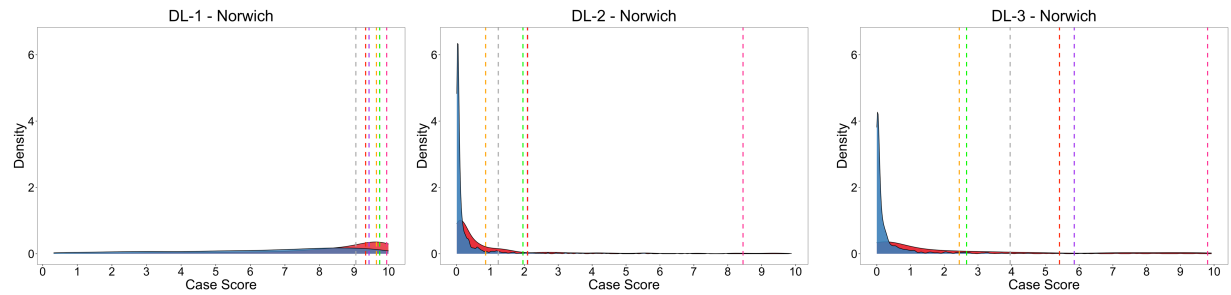


Figure 5-6 – Norwich data testing density plots for each artificial intelligence (AI) algorithm. *Interval cancer cases distribution is shown in red and normal case distribution is in blue. The green line represents the 'pre-specified specificity / sensitivity' (threshold 1) 96.0% specificity operating point for each algorithm and the orange line the 30.0% sensitivity operating point on Norwich study data. The purple line represents the 'identified year operating points' (threshold 2) 96.0% specificity operating point for each algorithm and the pink line the 30.0% sensitivity operating point. The red line represents the identified Cambridge operating points' (threshold 3) 96.0% specificity operating point for each algorithm and the grey line the 30.0% sensitivity operating point.*

5.4.3 Combined algorithm results

Combining the performance of all three DL algorithms (DL-1, DL-2, DL-3) using Cambridge data resulted in an AUROC of 0.738 [95% CI 0.713–0.764], which was not statistically significant different to the individual AI algorithms performance (p = 0.302–0.508). And at the threshold of 96.0% specificity, the sensitivity of the combined model was 25.4% [95% CI 21.4–30.0]. The contribution to the combined model was similar from both DL-1, DL-2 and DL-3. The ROC plot for each model on Cambridge data is shown in Figure 5-7.

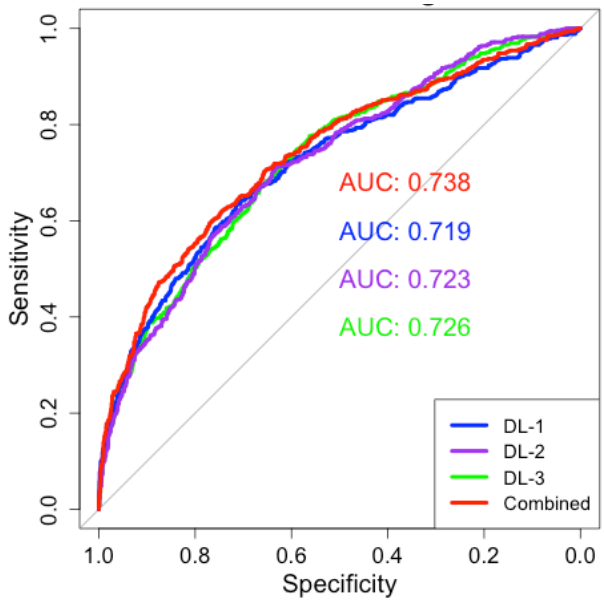


Figure 5-7 – Combined model receiver operating characteristic (ROC) curve on Cambridge data compared to individual artificial intelligence (AI) algorithms (DL-1, DL-2, DL-3) performance. *Results for DL-1 are in blue, DL-2 in purple, DL-3 in green, and the Combined model in red, with area under the receiver operating characteristic curve values provided for each algorithm.*

Applying the combined model to Norwich data also resulted in an AUROC of 0.738, which was statistically significantly different to DL-1, DL-2 and DL-3 ($p < 0.05$). At the 96.0% specificity operating point the combined model sensitivity on Norwich data was 25.7% [95% CI 21.0–31.6]. Applying the 96.0% operating point from Cambridge testing of the combined model, the sensitivity was 12.8% [95% CI 10.1–15.6] and specificity was 99.1% [95% CI 98.5–99.5]. The ROC plots for each model on Norwich data are shown in Figure 5-8.

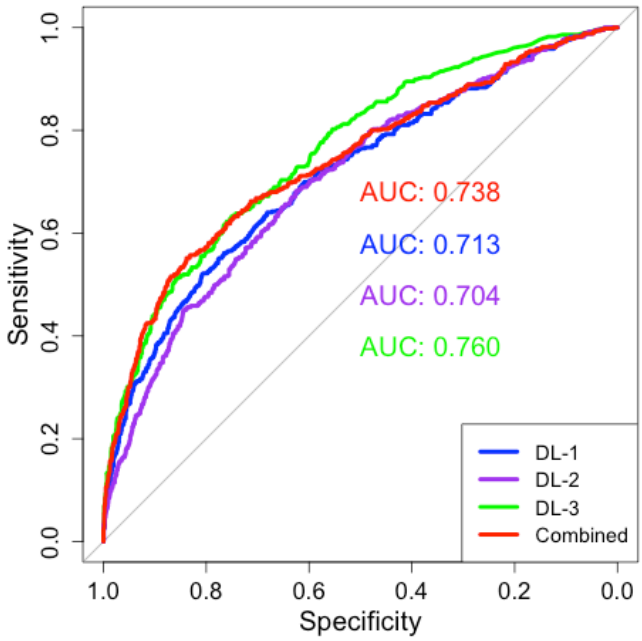


Figure 5-8 – Combined model receiver operating characteristic (ROC) curve on Norwich data compared to individual artificial intelligence (AI) algorithms (DL-1, DL-2, DL-3) performance. *Results for DL-1 are in blue, DL-2 in purple, DL-3 in green, and the Combined model in red, with area under the receiver operating characteristic curve values provided for each algorithm.*

5.4.4 Sub-group analysis

Sub group analysis on the entire cohort to evaluate each AI algorithms performance across key IC characteristic parameters at the 96.0% specificity 'identified year operating points' (threshold 2) is detailed in Table 5-7 and Table 5-8. Threshold 2 was used in this subgroup analysis as this threshold was found on a separate dataset reducing the bias of the threshold as well as the same threshold was used for Cambridge and Norwich data for each algorithm. When interpreting these results please refer to Table 5-5 and 5-6 which details the sensitivity and specificity at this threshold. When re-applying this threshold to Norwich there was a decrease in specificity with an increase in sensitivity for DL-1 and the opposite for DL-3, with the performance DL-2 remaining stable. Therefore the number of cancers detected by DL-1 at this threshold is greater than that for DL-2 and DL-3, however with the trade-off of decreased specificity. Overall detection was greater for ICs occurring in the first year, and for cancers that were classified as suspicious at the IC audit for all

three AI tools. On the other hand, detection was lower for grade 3 and less than 15 mm in size invasive tumours, however due to missing data this analysis is not definitive.

Interval cancer parameter		Total	DL-1		DL-2		DL-3	
Total Cases n		1029	299		196		179	
Age at Screening	47-49	114	25 (21.9%)	0.328	19 (16.7%)	0.754	16 (14.0%)	0.826
	50-54	231	60 (26.0%)		37 (16.0%)		42 (18.2%)	
	55-59	194	62 (32.0%)		39 (20.1%)		36 (18.6%)	
	60-64	184	49 (26.6%)		37 (20.1%)		28 (15.2%)	
	65-70+	306	103 (34.5%)		64 (22.7%)		57 (17.7%)	
FFDM Vendor	GE	518	194 (37.5%)	< 0.01	85 (16.4%)	< 0.01	72 (13.9%)	< 0.01
	Philips	473	96 (20.3%)		101 (21.4%)		98 (20.7%)	
	Hologic	19	3 (15.8%)		4 (21.1%)		3 (15.8%)	
	Sectra	19	6 (31.6%)		6 (31.6%)		6 (31.6%)	
Interval (months)	0-12	168	56 (33.3%)	0.577	38 (22.6%)	0.563	39 (23.2%)	0.199
	13-24	406	118 (29.0%)		76 (18.7%)		65 (16.0%)	
	35-36	455	125 (27.5%)		82 (18.0%)		75 (16.5%)	
Radiological Audit Classification	Normal / Benign	811	180 (22.3%)	< 0.01	119 (14.7%)	< 0.01	103 (12.8%)	< 0.01
	Uncertain	172	90 (52.3%)		58 (33.9%)		58 (33.9%)	
	Suspicious	35	26 (74.3%)		16 (45.7%)		14 (41.2%)	
	Unclassifiable	10	3 (30.0%)		3 (30.0%)		4 (40.0%)	
Density BI-RADS$^{\beta}$	a	20	1 (5.0%)	0.187	3 (15.0%)	0.752	4 (20.0%)	0.213
	b	110	20 (18.2%)		20 (18.2%)		14 (12.7%)	
	c	114	32 (28.1%)		28 (24.6%)		25 (21.9%)	
	d	76	15 (19.7%)		16 (21.1%)		21 (27.6%)	

Density BI-RADS$^{\alpha}$	a	66	13 (19.7%)	0.093	10 (15.2%)	0.212	8 (12.1%)	0.357
	b	469	155 (33.0%)		86 (18.3%)		75 (16.0%)	
	c	468	128 (27.4%)		99 (21.2%)		93 (19.9%)	
	d	26	3 (11.5%)		1 (3.8%)		3 (11.5%)	
Invasive Status	Invasive	964	283 (29.4%)	0.578	186 (19.3%)	0.963	170 (17.6%)	0.966
	Non-invasive	84	28 (33.3%)		16 (19.0%)		15 (17.9%)	

Table 5-7 – Subgroup analysis of cases using all interval cancer (IC) data from both Cambridge and Norwich sites. *The total number of interval cancer cases detected at the 96.0% specificity 'identified year operating points' (threshold 2) for each artificial intelligence algorithm is reported. Sensitivity is reported in round brackets. BI-RADS: Breast imaging-reporting and data system, FFDM: Full Field Digital Mammography. $^{\beta}$Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammogram Cambridge data, $^{\alpha}$DL-3 5th edition BI-RADS scores from processed full field digital mammogram data at both sites. p values were determined by using Chi squared χ^2 test to compare against the detected proportion of interval cancer cases / lesions by each artificial intelligence algorithm for each interval cancer characteristic category. p-values < 0.05 were considered statistically significant.*

At this threshold there was no statistically significant difference in the cancers detected by each AI tool for; patient age, interval to diagnosis, BI-RADS mammographic breast density, invasive status, and grade ($p > 0.05$). There was however a statistically significant difference between radiological classification groups and mammographic machine vendor for all of the three AI algorithms ($p < 0.05$). In addition, there was a statistically significant difference for DL-3 invasive tumour size ($p < 0.05$).

Interval cancer parameter		Total	DL-1		DL-2		DL-3	
Total Invasive Lesions n		964	283		186		170	
Invasive Tumour Grade$^{\delta}$	1	114	37 (32.5%)	0.171	23 (20.2%)	0.350	23 (20.2%)	0.085
	2	454	146 (32.2%)		97 (21.4%)		92 (20.3%)	
	3	363	89 (24.5%)		60 (16.5%)		49 (13.5%)	
Invasive Tumour Size$^{\delta}$ (mm)	< 15	260	71 (27.3%)	0.337	39 (15.0%)	0.055	34 (13.1%)	0.034
	>= 15	566	180 (31.8%)		124 (21.9%)		115 (20.3%)	

Table 5-8 – Subgroup analysis of lesions using all interval cancer (IC) data from both Cambridge and Norwich sites. *The total number of interval cancer lesions detected at the 96.0% specificity 'identified year operating points' (threshold 2) for each artificial intelligence algorithm is reported. Sensitivity is reported in round brackets. $^{\delta}$Report by invasive lesions for size and grade. p values were determined by using Chi squared χ^2 test to compare against the detected proportion of interval cancer lesions by each artificial intelligence algorithm for each interval cancer characteristic category. p-values < 0.05 were considered statistically significant.*

The AI algorithms did overlap in the ICs detected. However, the AI algorithms did not identify identical IC cases as shown in Figure 5-9, for threshold 1 and 2 at 96.0% specificity.

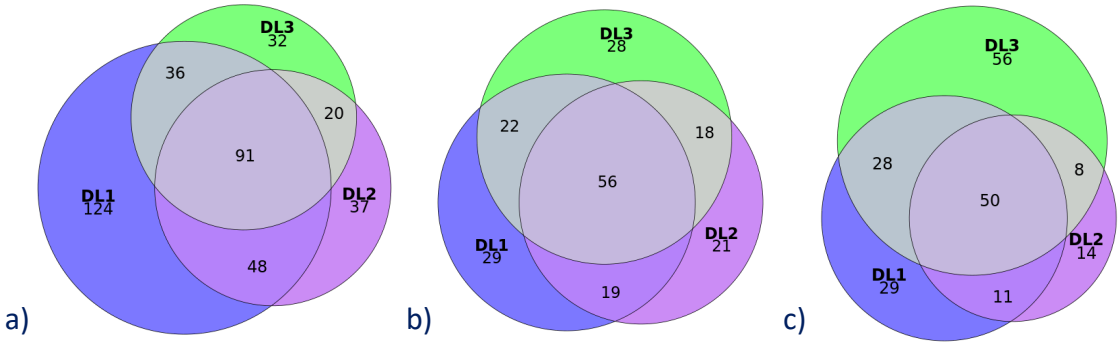


Figure 5-9 – Proportional Euler diagram of each artificial intelligence (AI) algorithms interval cancer (IC) detection. *a) At threshold 2 (96.0% specificity), using all interval cancer data from both Cambridge and Norwich sites, b) at threshold 1 (96.0% specificity), using all interval cancer data from Cambridge, c) at threshold 1 (96.0% specificity), using all interval cancer data from Norwich.*

5.4.5 Failure analysis

A case classified as a suspicious IC that was not detected by all methods, human readers and AI algorithms, at the 96.0% specificity threshold 2 is shown in Figure 5-10. This was a case of a 59-year-old patient, diagnosed with a left sided grade 2, 140 mm invasive cancer, 987 days after screening.

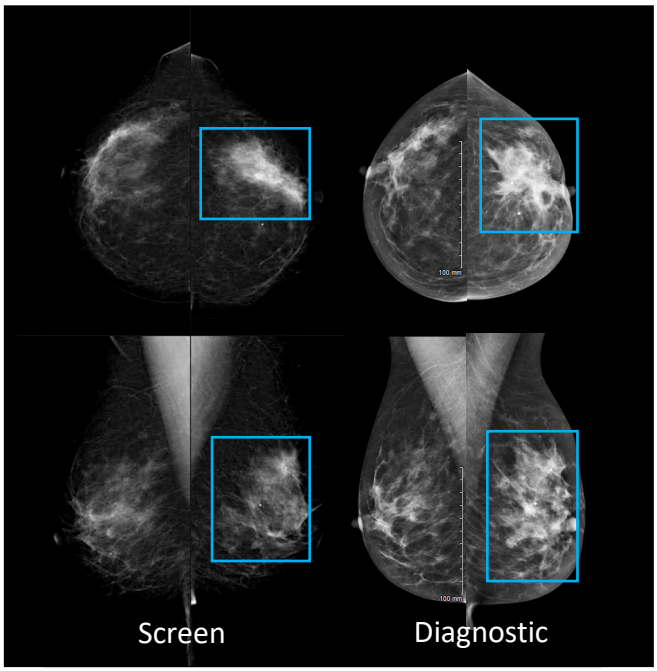


Figure 5-10 – False negative case, which was not detected by all three commercial artificial intelligence (AI) algorithms. *The screen and diagnostic images were annotated by a breast radiologist to show the true location of the cancer.*

A case classified as a normal / benign IC that was detected by all AI algorithms, at the 96.0% specificity threshold 2, is shown in Figure 5-11. This was a case of a 52-year-old patient, diagnosed with a left sided grade 3, 10 mm invasive cancer, 789 days after screening.

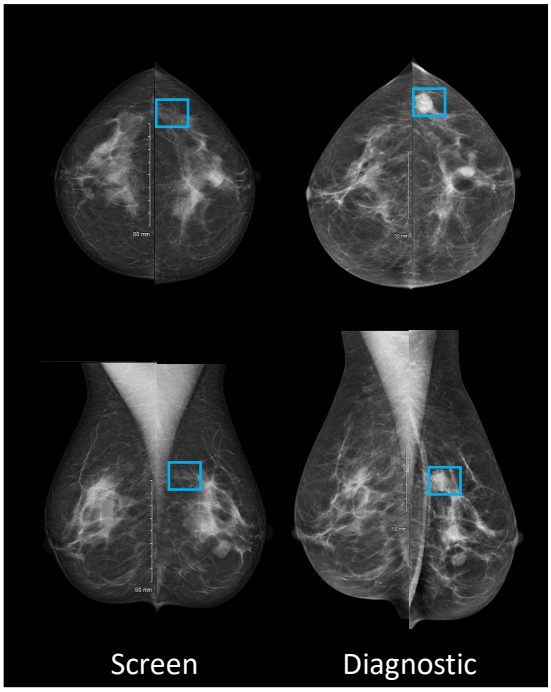


Figure 5-11 – True positive case, which was detected by all three commercial artificial intelligence (AI) algorithms. *The screen and diagnostic images were annotated by a breast radiologist to show the true location of the cancer.*

5.5 Discussion

The three commercial AI algorithms performed similarly and maintained acceptable performance at the 'pre-specified specificity / sensitivity' (threshold 1) for stand-alone IC detection. Thus, AI algorithms could play a role in the earlier detection of cancers. When using the algorithms at the same specificity as the screening programme double reader performance (96.0%), 21.6%-23.7% of ICs at Cambridge and 16.4%-27.9% of ICs at Norwich were detected. This is similar to the expected reported percentage of visible cancers that could have been detected, ~20.0-30.0%, at the previous screen[103]. Although this result was found using the 'pre-specified specificity / sensitivity' (threshold 1), which is not the threshold used in routine practice as this cut off is drawn from a population without SDCs and also a dataset enriched with IC cases. When transferring operating points identified at one site (Cambridge) using a one-year cohort (2017) with 2.2% cancers (SDC and IC) to both sites, performance was maintained for the Cambridge site, whilst there was a shift in performance shown for two out of the three algorithms (DL-1 and DL-3) when applied at the Norwich site. A significant shift was seen for all AI algorithms at the 30.0% sensitivity thresholds at both sites, with a very high specificity (99.5%-100.0%) whilst very low sensitivity was achieved (0.4%-

6.5%). This is expected due to the change in cancer proportions between the two datasets. Consistency / reliability of transferring operating points between sites should be monitored and is a key metric in performance. In addition, the dataset used to identify the threshold should be clearly documented in order to allow for monitoring where there is variation between sites e.g. mammography machine manufacturer. Based on this analysis, DL-2 demonstrated good generalizability and reliability to other sites with stable performance in the 96.0% specificity threshold 2 at both Cambridge and Norwich.

Sensitivity and specificity should be stated when using AUROC to report model performance as demonstrated in this study where model (e.g. DL-3) achieved the highest AUROC on Cambridge site data. However, as we were evaluating model performance at one extreme of the ROC curve (96.0% specificity), in order to avoid an increase in recall rates and thus costs of assessment clinics, the sensitivity when reported for the model with the highest AUROC (DL-3) is lower than another model (DL-1) at this threshold. Thus, AUROC should not be the only metric reported and should not be the deciding factor of an AI algorithms performance when under taking evaluation in a breast screening programme task.

As there is an overlap of the Cambridge database with The Optimam Mammography Image Database (OMI-DB) database (2012-2016), detailed in Chapter 4, these cases were identified and removed from this study to ensure that the same cases were not used in training and testing[286]. Two out of the three companies used OMI-DB in their training of the algorithm which may explain the good performance in Cambridge Philips data, despite Philips's data being used for a small percentage of training. To account for this the algorithms were tested on the completely independent Norwich dataset, that has never been used for the training of any AI algorithm. DL-3 performance improved when tested on Norwich data compared to Cambridge data, this is likely due to the significant proportion of GE images used in the DL-3 algorithm training.

Combining the three AI algorithms into one model did not significantly increase performance compared to a single algorithm when tested on Cambridge data, however there was a statistically significant difference between the Combine model and all three algorithms when tested on Norwich data. Thus there was no overfitting displayed and further work is needed to determine if there is an advantage of using different AI systems together for screen reading tasks.

The AI algorithms did not preferentially detect specific IC characteristics, other than for the radiological classification of cases (uncertain and suspicious) and mammographic machine vendor. Importantly, there was no difference found between invasive size and grade categories of ICs detected by all three AI systems, except for one system and invasive size. The algorithms did detect

different ICs to each other at threshold 1 and 2, therefore it may be possible that these systems could be used in tandem with all three systems operating independently to increase IC detection. This is the first study to compare three separate algorithms for the use in IC detection on UK data, as well as using the largest set of IC cases reported, addressing the gap in evidence identified in the NSC report 2021. The results found in this study are similar to the results in Lång *et al*, where 11.2% of visible ICs were detected and correctly located using Transpara v1.5.0 at 4.0% recall rate, Larsen *et al* where 30.7% ICs were detected at a 5.8% recall rate using Transpara v1.7.0, and Dembrower et al where 27.0% of ICs were detected using Lunit v5.5.0 at a 5.0% recall rate[134,270,271]. McKinney *et al* reported slightly lower values of 2.7-9.4% of ICs were detected when using an in house algorithm from Google[138] .

There is a potential role of these systems to be used to guide supplemental imaging in breast screening, as shown in Wanders *et al*, where 50.9% of women who developed an IC were identified at 90.0% specificity by combining Transpara v1.6.0 with LIBRA density in an enriched cohort[287]. Also Dembrower *et al* showed the rule in triage identified 12.0-27.0% ICs in the highest 1.0-5.0% cases suggesting a hybrid tailored screening approach could be made feasible by using AI algorithms[134].

There are limitations to this study such as the overall small study cohort without the representative class-imbalance of routine screening. In addition, not using the annotation provided by AI tools to confirm correct AI tool location identification of a cancer, which is critical for ICs with no radiological signs to guide further assessment unlike in Lång *et al* where they did confirm the location for each IC cases[270]. Thus, it is not possible to conclude if the cancers identified based on the threshold score only would be detected at assessment without additional correct location prompting by the AI system. In addition, this was a retrospective study and so it is not known how a human reader would behave with a prompt on a cancer that two readers have previously dismissed. Further prospective studies are required to confirm these results. Furthermore, there was missing cancer information at both sites, this was due to both data not being recorded and well as patients not undergoing further investigations or operations when their IC was diagnosed. Lastly, it should also be considered that the invasive size used in the analysis maybe subject to change due to the effect of neo-adjuvant chemotherapy, which is not commonly available through the automated extraction of data from NBSS.

5.6 Conclusion

The three AI algorithms were able to detect ICs at the preceding screening mammogram, detecting between 16.0-27.0% ICs across two UK screening sites at a 96.0% specificity threshold. However, when translating identified operating points from a year cohort from one study site to the other

there was a significant variation in performance for two out of the three algorithms and thus stability must be monitored across sites when translating operating points. It is unknown how readers would react to such cases being flagged where no location information is provided, and what is the best deployment route for algorithms to maintain such performance for IC detection in the real-time screening workflow. Thus, future prospective studies using the identified operating points across UK screening sites are required as well as sufficient follow-up to monitor the impact of AI algorithms on IC rates.

Chapter 6 – Performance of stand-alone deep learning algorithms in a UK screening cohort for detection and diagnosis

6.1 Aims

In this chapter three commercial artificial intelligence (AI) algorithms for stand-alone screen reading are investigated using a representative cohort from two UK screening centres. Performance was compared against that of a human reader, for three stand-alone reading approaches and non-inferiority was demonstrated for the AI algorithms at various benchmarks. The inclusion of interval and next round cancers allowed for the robust assessment of AI algorithms performance for the earlier detection of breast cancer. The results from this chapter provided data to plan future prospective studies.

Contents of this chapter have been accepted for presentation at the European Congress of Radiology 2022 [abstract number - #12040] and submitted to the European Society of Breast Imaging conference 2022 [abstract ID - #A-165].

6.2 Introduction

Traditional computer aided detection (CAD) algorithms have been used as clinical decision support systems, predominantly in the USA screening programme. However CAD systems have been shown to increase the recall rate with little improvement in reader performance, especially for experienced readers[125,288]. With the increasing improvement in performance of deep learning (DL) methods it has been proposed that these AI tools could be deployed as computer aided detection and diagnosis (CADe+x) stand-alone systems either entirely independently or alongside existing readers[133]. Many stand-alone systems have been tested and shown to be non-inferior to the first reader / single reader performance and even superior in some cases when used as a stand-alone system[133,289,290]. However no algorithm has been shown to be superior to the standard double reading performance whilst maintaining acceptable recall rates, suggesting that DL will not replace human reading entirely in these programmes at present[133,289]. Few algorithms have been compared on the same independent dataset for benchmarking against acceptable performance thresholds[137,149]. The UK National Screening Committee (NSC) report in 2021 concluded that further evidence, retrospective and prospective, using UK data, is required before these systems are implemented into the National Health Service Breast Screening Programme (NHSBSP)[136].

This study aimed to evaluate the performance of three different AI algorithms for stand-alone detection and diagnosis (CADe+x) of breast cancer, using a representative UK screening cohort from two sites, and comparing against UK reader performance thresholds. This will provide evidence for

three AI algorithm deployment approaches as well as identifying AI algorithm thresholds for prospective studies.

6.3 Methods

6.3.1 Sample size

The sample size for this study was calculated to determine the minimum number of cases required to reliably detect a meaningful difference between the AI algorithm performance and reader performance. The method was derived from Arkin *et al* for 'comparing a variable and a fixed proportion'[274]. The fixed proportion was determined by the screen readers sensitivity in the study reported cohorts (2017), in order to determine non-inferiority of the AI algorithm in comparison to the average UK reader performance.

The key metrics involved in the calculation are:

- a - Reference proportion – which is the average three-year single first reader and double reader sensitivity at the two screening sites, 62.9% and 67.4% respectively
- Effect size – which is the size of difference required to be shown between the groups. This was set at 10%
- b - (Reference proportion - Effect size)
- Power - (1-P(Type 2 error)) – which was set to 95% (β = 0.05)
- Significance level - P(Type 1 error) – which was set at 0.025

$$Sample\ size\ n = \left[\frac{Z_{\alpha_2}\sqrt{\pi_a(1-\pi_a)} + Z_{\beta}\sqrt{\pi_b(1-\pi_b)}}{(\pi_a - \pi_b)} \right]^2$$

The first calculation, to demonstrate that the AI algorithm is as good as the average single first reader (independent) performance over three years (screen detected cancers (SDCs) plus interval cancers (ICs)) when used as a stand-alone system, found that 313 cancers were required for this study. The second calculation, to show the AI algorithm plus a single first reader and arbitration is as good as the average double reading three yearly performance (SDCs plus ICs), found that 300 cancers were required for this study. Therefore a sufficient sample size between 300 and 313 cancers was required for this study.

6.3.2 Data

Patient data was obtained from the CC-MEDIA database described in Chapter 4, where data was collected from two NHSBSP sites (Cambridge and Norwich) under existing ethical approval (HRA REC 20/LO/0104, HRA CAG 20/CAG/0009, PHE RAC BSPRAC_090). All study data was de-identified prior to use in this research. Processed Full Field Digital Mammogram (FFDM) image data (right and left

craniocaudal (CC) and mediolateral oblique views (MLO)) and corresponding clinical metadata was retrospectively collected for all women who attended routine three yearly screening between January 1 2017 and December 31 2017 in order to obtain a sufficient sample size. Cases were excluded if they had an incomplete mammogram (less than two views of each breast or images not available on Picture Archiving and Communication System (PACS)), no ground truth was available, if the case was part of high-risk screening or the screen was documented as a technical recall. Cancer cases were also removed where they did not meet the specified definition, such as secondary melanoma metastasis recorded as an IC and confirmed following discussions with Public Health England (PHE). IC cases were removed if the interval from screening was recorded as longer than 40 months. As per the AI algorithm manufacturers documentation breast implants, pacemakers (including loop recorders) were excluded as well as any cases where only raw data was available or a pixel error occurred. Examples of artefacts excluded from the study cohort are shown in Figure 6-1.

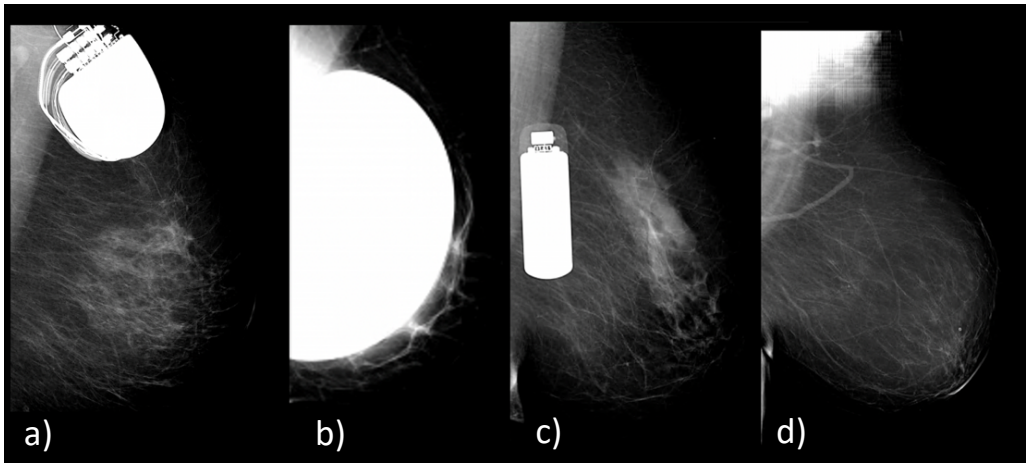


Figure 6-1 – Mediolateral oblique (MLO) views of mammogram artefacts removed from the study. *a) Pacemaker, b) breast implant, c) loop recorder device, d) pixel error.*

The study case selection process is shown in a Standards for Reporting of Diagnostic Accuracy Studies (STARD) diagram in Figure 6-2[275].

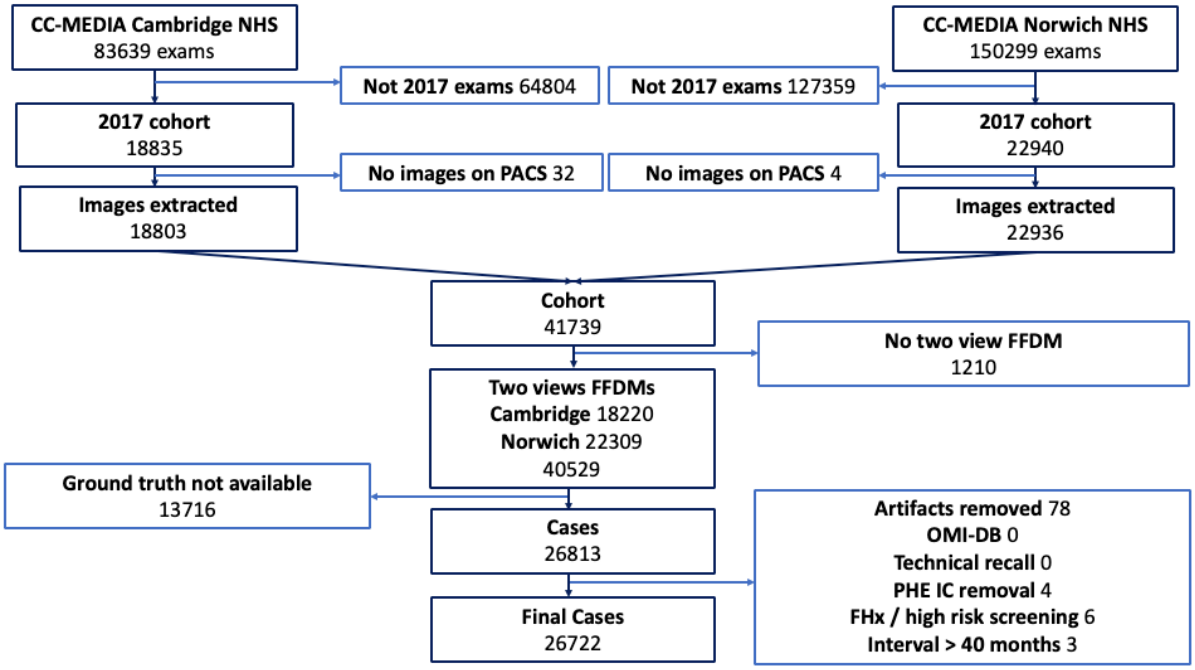


Figure 6-2 – Standards for Reporting of Diagnostic Accuracy Studies (STARD) flow diagram of cases included and excluded in this study. *FFDM: Full Field Digital Mammogram, FHx: Family history, IC: Interval cancer, NHS: National Health Service, OMI-DB: The Optimam Mammography Image Database, PHE: Public Health England, PACS: Picture Archiving and Communication System.*

One exam was included per patient. All exams had not previously been seen by any AI algorithm. All images were stored in JPEG Lossless DICOM format and no additional pre-processing other than that performed by the mammography vendor and that performed by the AI algorithm occurred. Corresponding clinical metadata was available for each case, including each readers decision at screening. Trainee readers were removed from this analysis and replaced with the first and second trained reader decision. The invasive status (ICD-10 code), histological grade (assigned using Nottingham grading system), and histological size, was obtained using an automated National Breast Screening System (NBSS) query, for further detail please see Chapter 4 Section 4.4.5.

6.3.3 Ground truth

The NHSBSP is a triennial screening programme, thus women are screened every 34-36 months using FFDM. The ground truth for normal cases was defined as a final reader action of routine recall (RR) more than > 912 days (30 months) after their previous screen, to account for early recall of women to three-year screening and a confirmed 'no cancer diagnoses' within three years. The study follow-up time period overlaps with the pause in screening during the Covid-19 pandemic, therefore cases were excluded if sufficient follow-up information was not available. We used this definition of a 'normal' case to provide a robust ground truth for these cases.

Cancer cases were identified using the existing NBSS queries. All cancers received a confirmed histopathological diagnosis and were classed as either a: SDC, next round cancer (NRC), future round cancer (FRC), IC, or next round interval cancer (NRIC), such that;

- SDCs were recalled and diagnosed at the screening episode included in the study, within 90 days (3 months).
- NRCs were recalled at the next screening episode, after the screening episode included in this study.
- FRCs were recalled at the second screening episode, after the screening episode included in this study.
- ICs occurred in the interval following a negative screen, within 1216 days (40 months) of the screening episode, and received a confirmed histopathological diagnosis.
- NRICs occurred less than 1216 days (40 months) after the next round screening episode, and received a confirmed histopathological diagnosis.

6.3.4 AI tools

Three commercial AI algorithms were installed at the University of Cambridge. Two AI algorithms were hosted in a local environment using a virtual machine connection and one AI algorithm was run using hardware supplied by the AI company. The AI companies did not have access to their algorithms following the successful setup installation and at no time had access to the study data. Details regarding the training data used by each AI algorithm as well as the technical setup and algorithm output is outlined in Chapter 5 Table 5-1.

Density was calculated using Volpara (research version - VolparaResearch32_L30Enabled_v2, Wellington, New Zealand) and DL-3. The Breast Imaging-Reporting and Data System (BI-RADS) 5th edition density score from both Volpara and DL-3 is reported in this study.

6.3.5 Thresholds

SDCs and ICs, occurring within the three-year screening interval, were classified as cancer cases in both the study and when identifying any study thresholds.

Three thresholds were used in this study. The first was set at the single first reader three yearly specificity for the entire study cohort (96.6%) (threshold 1). The second threshold was identified using one year of Cambridge study data (2018) to identify the operating point for each AI algorithm at the first reader specificity performance (96.6%) (threshold 2). The 2018 Cambridge cohort used to identify this threshold consisted of 12,455 cases of which 239 were cancer cases (183 SDCs (1.5%), and 56 ICs (0.5%)). The third threshold (threshold 3) of 99.0% specificity, was also identified using the Cambridge 2018 data cohort.

Each AI algorithms performance was then assessed using these three thresholds. Adapted screening reading workflows, are outlined in Figure 6-3. Figure 6-3.b, shows how the AI algorithm performance alone was compared to the single first independent reader using threshold 1 and threshold 2. Figure 6-3.c, shows the combined AI and human reader approach, where the AI algorithm was set at threshold 2 and combined with the single human first reader. If there was discordance the final action decision was used (either second reader or arbitration) and the overall performance was compared to double reading performance as shown in Figure 6-3.a.

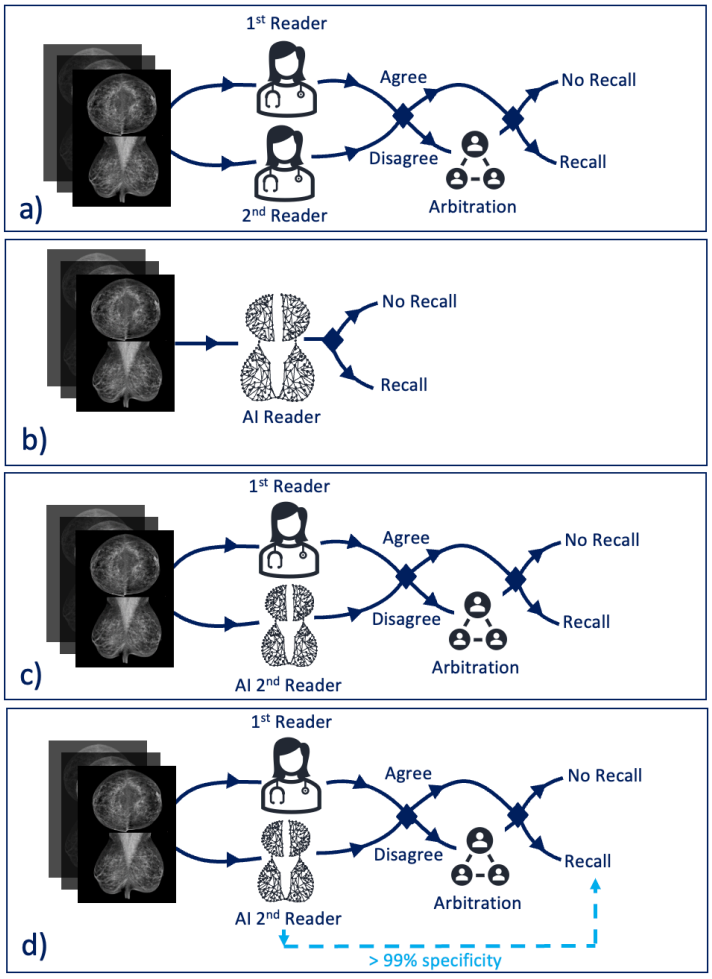


Figure 6-3 – Proposed workflow deployment of a stand-alone computer aided detection and diagnosis (CADe+x) artificial intelligence (AI) algorithm. *a) Routine UK double reading workflow, b) stand-alone artificial intelligence algorithm reader, c) single human and artificial intelligence algorithm reader, with arbitration where there is discordance, d) auto recall of cases, not recalled by single human and artificial intelligence algorithm workflow, for cases that score above the artificial intelligence algorithm threshold of 99.0% specificity.*

Figure 6-3.d, demonstrates the use of the auto recall threshold where all cases above the 99.0% specificity threshold of the AI algorithm were automatically recalled. Any cases below this threshold but above the 96.6% of the AI algorithm, and those recalled by the first reader, were recalled. Where there was discordance between the AI algorithm and the first reader (not including cases above the

99.0% specificity threshold) cases were referred to arbitration where the final action decision was taken. The results from this workflow were also compared to double reading performance as shown in Figure 6-3.a.

6.3.6 Statistical analysis

All statistical analysis took place in R version 4.0.4 (R Foundation for Statistical Computing, Vienna, Austria)[225], using the packages detailed in Chapter 5 Section 5.3.6.

The overall predictive performance of each AI algorithm was evaluated using area under the receiver operating characteristic (AUROC) curve, the partial AUROC (pAUROC) at 96.0-100% specificity, and area under the precision recall curve (AUPRC). Due to the imbalanced nature of the data (3% cancers to 97% normal cases) precision and sensitivity were the primary outcome measures for this study.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Performance of each AI algorithm was compared to readers performance, using one sample one tailed z-test to determine if the algorithm was non-inferior. The percentage of cancers detected from each category (SDC, IC, NRC, FRC and NRIC) was calculated for the AI algorithm at each threshold. Perturbation analysis took place to test the robustness of each AI algorithm against changes in performance thresholds.

A multivariable model was created through the combination of all three individual algorithms using a generalised linear mixed effects model on Cambridge site data. This combined model was then tested using Norwich site data to check for overfitting. DeLong's test was used to assess for a statistically significant difference between the AUROC curve of individual AI algorithms using 2000 bootstrapping examples.

Finally, sub group analysis to evaluate each AI algorithms performance for SDC and IC detection in the following categories took place; age at screening, breast density, invasive status, invasive grade and size of cancers as well as mammographic machine vendor. Further sub group analysis took place for ICs using the interval between screening and diagnosis, as well as the radiological audit classifications assigned to each case. A Chi squared χ^2 test was used to investigate if there was a statistically significance between categories[285]. In all analyses, p-values < 0.05 were considered statistically significant and 95% confidence intervals were calculated, using bootstrapping with 2000 samples or through an approximation method from Simel *et al* using the epiR package[291].

6.3.7 Reporting

Each AI algorithm was assigned a DL-ID for the purposes of this study. For additional details please refer to section 5.3.7 in Chapter 5. This study is reported in accordance with The Checklist for Artificial Intelligence in Medical Imaging (CLAIM) criteria[167].

6.4 Results

6.4.1 Data

In total 26,722 cases were included in this study, 11,924 cases (44.6%) were from Cambridge and 14,798 cases (55.4%) were from Norwich. Patient characteristics of the study cohort are shown in Table 6-1. The median age for the entire cohort was 59.0 [IQR 54.0–63.0].

	Cambridge n (%)		Norwich n (%)
Total Cases n	11924		14798
FFDM Vendor			
GE	121 (1.0%)		14798 (100%)
Philips	11803 (99.0%)		0 (0.0%)
Age at Screening			
Median [IQR]	57.0 [54.0-63.0]		59.0 [55.0-64.0]
47-49	13 (0.1%)		70 (0.5%)
50-54	4002 (33.6%)		2958 (20.0%)
55-59	2802 (23.5%)		5290 (35.8%)
60-64	2928 (24.6%)		2915 (19.7%)
65-69	1826 (15.3%)		2787 (18.8%)
70+	353 (3.0%)		778 (5.3%)
Density BI-RADS	Volpara$^{\beta}$	DL-3$^{\alpha}$	DL-3$^{\alpha}$
a	1968 (16.5%)	2755 (23.1%)	2353 (15.9%)
b	5474 (45.9%)	6568 (55.1%)	8660 (58.5%)
c	3247 (27.2%)	2548 (21.4%)	3614 (24.4%)
d	1217 (10.2%)	53 (0.4%)	171 (1.2%)
Missing	18 (0.2%)	0 (0.0%)	0 (0.0%)
Cancers			
SDC Rate per 1000 screens	152 8.1/1000		180 7.9/1000
IC Rate per 1000 screens	84 4.5/1000		90 3.9/1000
NRC Rate per 1000 screens	99 7.5/1000		155 9.6/1000
FRC	0		1*
NRIC	13*		15*

Table 6-1 – Summary of testing dataset characteristics. *Integer values with percentages in brackets (%) and median with Interquartile range in square brackets [IQR] are provided. BI-RADS: Breast imaging-reporting and data system, FRC: Future round cancer, FFDM: Full Field Digital Mammography, GE: General Electric, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, SDC: Screen detected cancer. *Rate was calculated by the total number of women screened that year, there was incomplete follow-up time period information from which to calculate an accurate rate for these groups. $^{\alpha}$DL-3 5th edition BI-RADS density scores on processed full field digital mammograms. $^{\beta}$Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data.*

The majority of Cambridge cases were classed as b / c BI-RADS density, which was consistent with the expected distribution across the reported screening population[75].

In total 506 three-year cancer cases (SDCs and ICs) were included in this study, 236 from Cambridge and 270 from Norwich. A total of 254 NRC cases were also included. The characteristics of the SDC and NRC cases in the study cohort are shown in Table 6-2.

		Cambridge SDC n (%)		Cambridge NRC n (%)		Norwich SDC n (%)	Norwich NRC n (%)
Total Cases n		152		99		180	155
Total Lesions n		159		104		184	161
Round length*$^{\lambda}$ [IQR]		35.6 [35.1-36.1]		41.7 [36.7-45.2]		35.2 [35.1-36.1]	39.0 [35.5-39.8]
Age at Screening $^{\lambda}$	Median [IQR]	62.0 [56.0-67.0]		59.0 [54.0-65.0]		64.0 [59.0-68.0]	60.0 [56.0-65.0]
	47-49	0 (0.0%)		1 (1.0%)		0 (0.0%)	0 (0.0%)
	50-54	37 (24.3%)		29 (29.3%)		15 (8.3%)	20 (12.9%)
	55-59	25 (16.5%)		22 (22.2%)		36 (20.0%)	50 (32.3%)
	60-64	29 (19.1%)		18 (18.2%)		40 (22.2%)	34 (21.9%)
	65-69	44 (29.0%)		22 (22.2%)		56 (31.1%)	35 (22.6%)
	70+	17 (11.2%)		7 (7.1%)		33 (18.3%)	16 (10.3%)
Invasive Status	Invasive	134 (84.3%)		86 (82.7%)		152 (82.6%)	136 (84.5%)
	Non-invasive	24 (15.1%)		18 (17.3%)		30 (16.3%)	25 (15.5%)
	Missing	1 (0.6%)		0 (0.0%)		2 (1.1%)	0 (0.0%)
Invasive Tumour Size$^{\delta}$	< 15 mm	73 (54.5%)		40 (46.5%)		87 (57.2%)	71 (52.2%)
	>= 15 mm	59 (44.0%)		34 (39.5%)		61 (40.1%)	55 (40.4%)
	Missing	2 (1.5%)		12 (14.0%)		4 (2.6%)	10 (7.4%)
Invasive Tumour Grade$^{\delta}$	1	23 (17.2%)		10 (11.6%)		44 (29.0%)	37 (27.2%)
	2	76 (56.7%)		60 (69.8%)		81 (53.3%)	65 (47.8%)
	3	30 (22.4%)		11 (12.8%)		26 (17.1%)	27 (19.9%)
	Missing	5 (3.7%)		5 (5.8%)		1 (0.7%)	7 (5.1%)
		Volpara$^{\beta}$	DL-3$^{\alpha}$	Volpara$^{\beta}$	DL-3$^{\alpha}$	DL-3$^{\alpha}$	DL-3$^{\alpha}$
Density BI-RADS$^{\lambda}$	a	18 (11.8%)	30 (19.7%)	13 (13.1%)	22 (22.2%)	15 (8.3%)	19 (12.3%)
	b	79 (52.0%)	92 (60.5%)	52 (52.5%)	50 (50.5%)	116 (64.4%)	89 (57.4%)
	c	43 (28.3%)	29 (19.1%)	24 (24.2%)	27 (27.3%)	49 (27.2%)	47 (30.3%)
	d	12 (7.9%)	1 (0.7%)	10 (10.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 6-2 – Cancer characteristics by lesions and cases. *With integer values and percentages in brackets (%). $^{\delta}$Invasive lesions only. BI-RADS: Breast imaging-reporting and data system, NRC: Next round, SDC: Screen detected cancer. $^{\lambda}$Cases only. $^{\alpha}$DL-3 5th edition BI-RADS density scores on processed full field digital mammograms.$^{\beta}$ Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data. *Round length is shown in months from the previous screen, cases without a previous screen or screened more than six years previously are removed from this analysis. The round length was increased for next round cancers due to the pause in screening during the Covid-19 pandemic which is described in Chapter 4.*

In total 174 ICs, 84 from Cambridge and 90 from Norwich were included in the study cohort. The characteristics of the IC cases in the study cohort are shown in Table 6-3. The median time to diagnosis was 825.5 [IQR 531.0–1002.5] days for all ICs at Cambridge and 725.5 [IQR 486.8–964.0] days at Norwich.

		Cambridge IC n (%)		Norwich IC n (%)
Total Cases n		84		90
Total Lesions n		86		100
Age at Screening[λ]	Median [IQR]	58.0 [54.0-65.3]		62.0 [55.0-68.0]
	47-49	0 (0.0%)		0 (0.0%)
	50-54	27 (32.1%)		17 (18.9%)
	55-59	19 (22.6%)		23 (25.6%)
	60-64	14 (16.7%)		13 (14.4%)
	65-69	14 (16.7%)		21 (23.3%)
	70+	10 (11.9%)		16 (17.8%)
Invasive Status	Invasive	74 (86.0%)		93 (93.0%)
	Non-invasive	9 (10.5%)		6 (6.0%)
	Missing	3 (3.5%)		1 (1.0%)
Invasive Tumour Size[δ]	< 15 mm	16 (21.6%)		36 (38.7%)
	>= 15 mm	48 (64.9%)		48 (51.6%)
	Missing	10 (13.5%)		9 (9.7%)
Invasive Tumour Grade[δ]	1	10 (13.5%)		18 (19.4%)
	2	33 (44.6%)		40 (43.0%)
	3	31 (41.9%)		32 (34.4%)
	Missing	0 (0.0%)		3 (3.2%)
		Volpara[β]	DL-3[α]	DL-3[α]
Density BI-RADS[λ]	a	8 (9.5%)	11 (13.1%)	2 (2.2%)
	b	34 (40.5%)	43 (51.2%)	51 (56.7%)
	c	29 (34.5%)	29 (34.5%)	36 (40.0%)
	d	12 (14.3%)	1 (1.2%)	1 (1.1%)
	Missing	1 (1.2%)	0 (0.0%)	0 (0.0%)
Interval (months)[λ]	0-12	13 (15.5%)		18 (20.0%)
	12-24	23 (27.4%)		28 (31.1%)
	24-36	48 (57.1%)		44 (48.9%)
	36-40	0 (0.0%)		0 (0.0%)
Radiological Audit Classification[λ]	Normal/ Benign	69 (82.1%)		60 (66.7%)
	Uncertain	11 (13.1%)		28 (31.1%)
	Suspicious	0 (0.0%)		0 (0.0%)
	Unclassifiable	0 (0.0%)		1 (1.1%)
	Missing	4 (4.8%)		1 (1.1%)

Table 6-3 – Interval cancer (IC) characteristics by lesions and cases. *With integer values and percentages in brackets (%). Invasive Tumour Size in millimetres (mm). BI-RADS: Breast imaging-reporting and data system, IC: Interval cancer. [δ]Invasive lesions only. [λ]Cases only. [α]DL-3 5th edition BI-RADS density scores on processed full field digital mammograms.[β] Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data.*

The majority of IC cases were classified as normal / benign in keeping with the reported UK distribution[103]. No cases were classified as suspicious, which was likely to be due to the change in national reporting of interval cancers in 2017[101].

6.4.2 Algorithm results

The overall AUROC for DL-1, DL-2, DL-3 was 0.868 [95% CI 0.849–0.887], 0.885 [95% CI 0.869–0.902] and 0.894 [95% CI 0.878–0.910] respectively. ROC curves for each AI algorithm are shown in Figure 6-4. All algorithms maintained a similar AUROC performance for both sites ($p > 0.05$). The AUROC of DL-3 was statistically significantly different to DL-1 on all and Norwich data, and DL-2 on Norwich data ($p < 0.05$). DL-1 and DL-2 were also statistically significantly different to each other on all and Norwich data ($p < 0.05$). The comparator ROC curves for each site are shown in Figure 6-5.

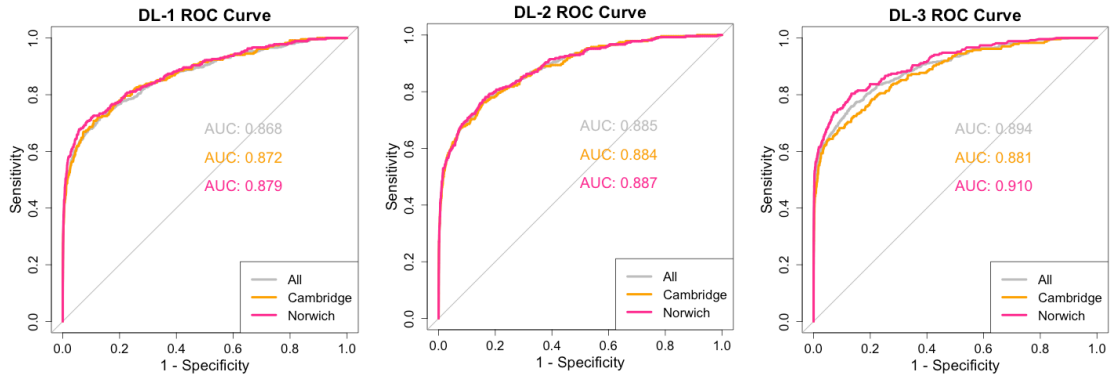


Figure 6-4 – Receiver operating characteristic (ROC) curves per artificial intelligence (AI) algorithm. *The overall results are in grey, Cambridge in orange and Norwich in pink. Area under the receiver operating characteristic curve values are provided for each site.*

The pAUROC, from 96.0% to 100% specificity, for DL-1, DL-2, DL-3 was 0.744 [95% CI 0.723–0.764], 0.739 [95% CI 0.720–0.760] and 0.774 [95% CI 0.754–0.794] respectively on all data.

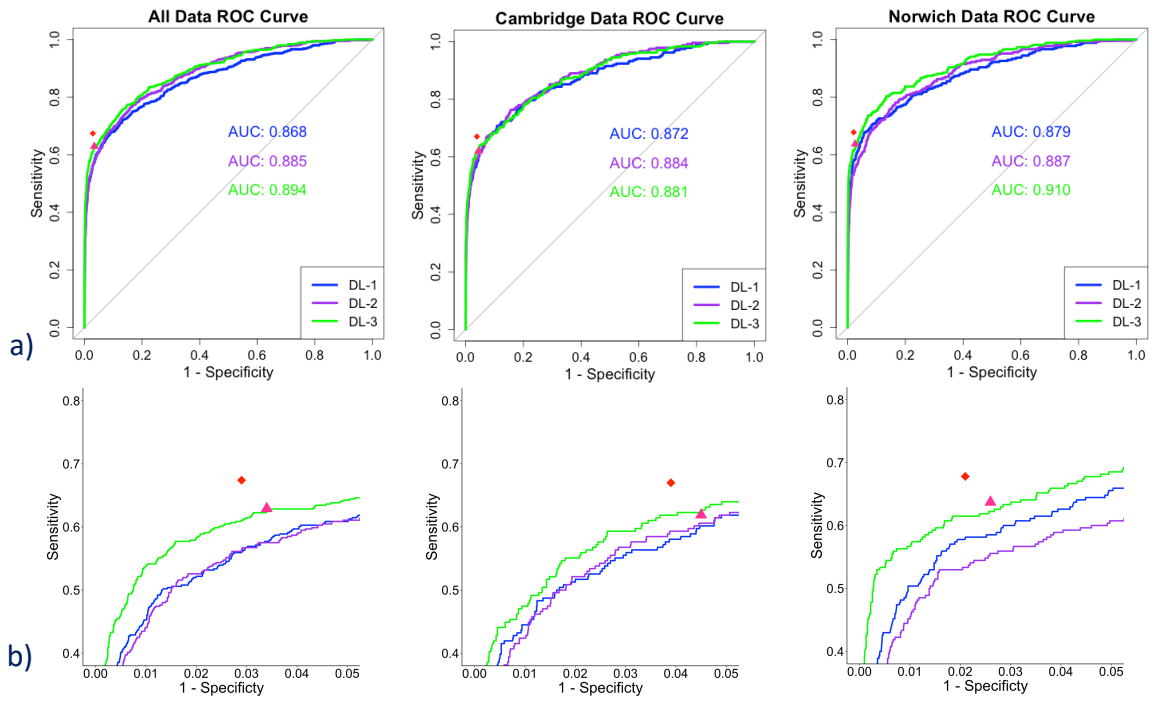


Figure 6-5 – Receiver operating characteristic (ROC) curves per site. *a) receiver operating characteristic curves per site, with the area under the receiver operating characteristic curve values provided for each algorithm, b) partial receiver operating characteristic curves to show the performance of each artificial intelligence algorithm between 95.0% and 100% specificity at each site. The results for DL-1 are in blue, DL-2 in purple, and DL-3 in green. A pink triangle represents the first reader performance, and a red diamond represents the overall double reader performance at each site.*

The pAUROC when tested on Cambridge data was lower for all algorithms; 0.739 [95% CI 0.709–0.769], 0.737 [95% CI 0.709–0.767] and 0.759 [95% CI 0.730–0.787] for DL-1, DL-2 and DL-3 respectively. On Norwich data all algorithms achieved a higher pAUROC compared to all and Cambridge data, with DL-1 achieving a pAUROC of 0.761 [95% CI 0.732–0.789], DL-2 0.741 [95% CI 0.714–0.769], and DL-3 0.791 [95% CI 0.762–0.818]. The pAUROC of DL-3 was statistically significantly different ($p < 0.05$) when compared to DL-1 and DL-2 on all, Cambridge and Norwich data. The pAUROC of DL-1 was also statistically significantly different ($p < 0.05$) compared to DL-2 on Norwich data.

The overall AUPRC for DL-1, DL-2, DL-3 was 0.440, 0.407, 0.513 respectively, Figure 6-6. The drop in DL-2 and DL-3 precision, shown in the precision recall curves (PRC), was due to either missing a true positive case or including more false positives at a high recall threshold. Although both curves recover, the curve for DL-2 remains consistently lower than DL-3.

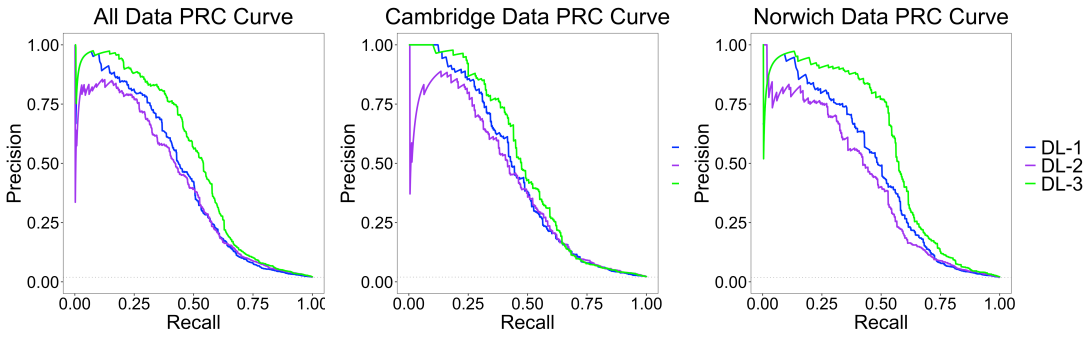


Figure 6-6 – Precision recall curves (PRC). *For DL-1 in blue, DL-2 in purple and DL-3 in green.*

When the AI algorithm threshold is set at the first screen reader specificity (96.6%) (threshold 1), DL-1, DL-2 and DL-3 were non-inferior relative to the single first reader, as shown in Table 6-4. DL-3 was also non-inferior to the double reader sensitivity. The AI algorithms detected more NRC (Δ +4.5%~+9.9%) and IC (Δ +5.2%~+8.0%) compared to the first reader, when these systems were used as stand-alone CADe+x readers. However, the number of SDCs found by all AI algorithms was less than the first reader (Δ -4.9%~-10.9%).

At the identified threshold 2, DL-2 maintained performance and was non-inferior to the single first human reader. However, the sensitivity of DL-1 improved with the trade-off of reduced specificity and the opposite was found for DL-3, whilst both algorithms sensitivity remained non-inferior to the first reader performance, Table 6-5. DL-1 sensitivity was also non-inferior to the double reader performance.

All three AI algorithms were able to detect a greater proportion of ICs and NRCs at both threshold 1 (96.6% specificity) and threshold 2 (Cambridge 2018 first reader 96.6% specificity performance) for the earlier detection of cancer compared to the human reader workflows offsetting the reduced rate of SDCs.

	Double reader	First reader	DL-1	DL-2	DL-3
AUROC	-	-	0.868	0.885	0.894
pAUROC	-	-	0.744	0.739	0.774
AUPRC	-	-	0.440	0.407	0.513
Sensitivity	67.4% [63.1-71.5]	62.9% [58.5-67.1]	57.7% [53.4-62.1] p = 0.016	57.5% [53.2-61.9] p = 0.02	62.5% [58.3-66.8] p < 0.01
-	-	-	Non-inferior	Non-inferior	Non-inferior
Specificity	97.1%	96.6%	96.6%	96.6%	96.6%
Precision	31.3%	26.0%	24.7%	24.6%	26.2%
Recall Rate	4.1%	4.6%	4.4%	4.4%	4.5%
Cancers					
SDC n (%)	332 (100%)	302 (91.0%)	266 (80.1%)	266 (80.1%)	286 (86.1%)
IC n (%)	9 (5.2%)	16 (9.2%)	26 (14.9%)	25 (14.4%)	30 (17.2%)
NRC n (%)	10 (3.9%)	13 (5.1%)	31 (12.2%)	32 (12.6%)	31 (12.2%)
FRC n (%)	0 (0.0%)	0 (0.0%)	1 (100%)	1 (100%)	1 (100%)
NRIC n (%)	2 (7.1%)	3 (10.7%)	2 (7.1%)	2 (7.1%)	1 (3.6%)

Table 6-4 – Stand-alone artificial intelligence (AI) algorithm application compared to the single first reader – threshold 1. *All algorithm thresholds set at the first reader, 96.6% specificity (threshold 1). AUROC: Area under the receiver operating characteristic curve, AUPRC: Area under the precision recall curve, FRC: Future round cancer, IC: Interval cancer, NRC: Next round, NRIC: Next round interval cancer, pAUROC: Partial area under the receiver operating characteristic curve, SDC: Screen detected cancer. 95.0% confidence intervals are shown in square brackets [95.0% CI]. p values are calculated using a one-sided z-test.*

	DL-1	DL-2	DL-3
Sensitivity	64.8% [61.3-68.2] p < 0.01	56.7% [53.0-60.5] p = 0.045	58.9% [55.3-62.5] p < 0.01
-	Non-inferior	Non-inferior	Non-inferior
Specificity	92.8% [92.5-93.1] p < 0.01	96.8% [96.7-97.0] p < 0.01	97.9% [97.8-98.0] p < 0.01
Precision	14.8%	25.6%	35.2%
Recall Rate	8.3%	4.2%	3.2%
SDC n (%)	287 (86.5%)	264 (79.5%)	275 (82.8%)
IC n (%)	41 (23.6%)	23 (13.2%)	23 (13.2%)
NRC n (%)	59 (23.2%)	32 (12.6%)	18 (7.1%)
FRC n (%)	1 (100%)	1 (100%)	0 (0.0%)
NRIC n (%)	6 (21.4%)	2 (7.1%)	1 (3.6%)

Table 6-5 – Stand-alone artificial intelligence (AI) algorithm application compared to the single first reader – threshold 2. *All algorithm thresholds set using the operating point identified using Cambridge 2018 data, threshold 2. FRC: Future round cancer, IC: Interval cancer, NRC: Next round, NRIC: Next round interval cancer, SDC: Screen detected cancer. 95.0% confidence intervals are shown in square brackets [95.0% CI]. p values are calculated using a one-sided z-test.*

The distribution of scores for each category of cases along with the cut off points of each threshold are shown in Figure 6-7.

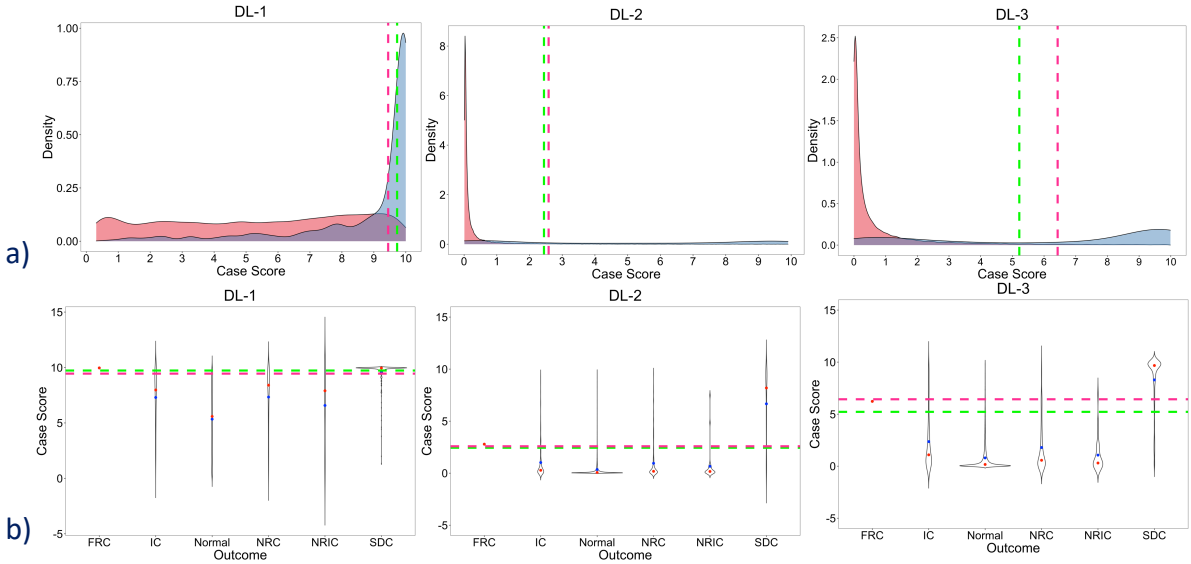


Figure 6-7 – Individual artificial intelligence (AI) algorithm score distributions normalised from 0-10. *a) Density plots, where normal cases are in red and cancer cases (screen detected and interval cancers) are in blue, and b) violin plots where the blue dot in the violin plot is the mean score and the red is the median score. The green line indicates the 96.6% specificity threshold (threshold 1) and the pink line is the operating point identified from the Cambridge 2018 data (threshold 2). FRC: Future round cancer, IC: Interval cancer, NRC: Next round, NRIC: Next round interval cancer, SDC: Screen detected cancer.*

	Double reader	First reader + DL-1	First reader + DL-2	First reader + DL-3
Sensitivity	67.4% [63.1-71.5]	67.0% [62.7-71.1] p < 0.01	65.6% [61.3-69.7] p < 0.01	65.4% [61.1-69.6] p < 0.01
-	-	Non-inferior	Non-inferior	Non-inferior
Specificity	97.1% [96.7-97.3]	97.4% [97.2-97.6] p < 0.01	97.6% [97.4-97.7] p < 0.01	97.6% [97.4-97.8] p < 0.01
Precision	31.3%	33.4%	34.2%	34.4%
Arbitration	2.7%	9.5%	6.3%	5.2%
Recall Rate	4.1%	3.8%	3.6%	3.6%
SDC n (%)	332 (100%)	326 (98.2%)	323 (97.3%)	321 (96.7%)
IC n (%)	9 (5.2%)	13 (7.5%)	9 (5.2%)	10 (5.8%)
NRC n (%)	10 (3.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
FRC n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
NRIC n (%)	2 (7.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 6-6 – Artificial intelligence (AI) algorithm (at threshold 2) combined with the single first reader (+/- arbitration where discordance) compared to double reading performance. *FRC: Future round cancer, IC: Interval cancer, NRC: Next round, NRIC: Next round interval cancer, SDC: Screen detected cancer. 95.0% confidence intervals are shown in square brackets [95.0% CI]. p values are calculated using a one-sided z-test.*

Combining the AI algorithm with the human reader decision, using the identified threshold 2 for each AI algorithm, resulted in non-inferior sensitivity and specificity performance. The overall recall

rate was lower, however the was arbitration rate was higher (Δ +2.5%~+6.8%), Table 6-6. A reduction in SDCs (Δ -1.8%~-3.3%) and reduction in NRCs (Δ -3.9%) with only a modest increase in IC detection (Δ +0.0%~+2.3%) was noted for all algorithms compared to double reader performance, Table 6-6.

Perturbation analysis demonstrated that all the algorithms performed similarly and are robust to changes in specificity. When adjusting the AI algorithms specificity to ~90.5% and in combination with the first reader, and arbitration for discordance, the performance for all AI algorithms was close to double reader sensitivity without increasing the overall recall rate, but with an increase in the arbitration rate, Table 6-7.

	AI-Specificity	Sensitivity	Specificity	Precision	Arbitration	Recall
DL-1 + readers	97.5%	65.4%	97.6%	34.5%	5.82%	3.60%
	96.5%	66.2%	97.6%	34.5%	6.58%	3.63%
	95.5%	66.4%	97.5%	34.1%	7.34%	3.69%
	94.5%	66.4%	97.5%	33.8%	8.14%	3.72%
	93.5%	66.8%	97.5%	33.7%	8.94%	3.76%
	92.5%	67.0%	97.4%	33.3%	9.75%	3.81%
	91.5%	67.0%	97.4%	33.1%	10.60%	3.83%
	90.5%	67.0%	97.4%	33.0%	11.47%	3.84%
DL-2 + readers	97.5%	65.4%	97.6%	34.4%	5.70%	3.60%
	96.5%	65.6%	97.5%	34.0%	6.48%	3.65%
	95.5%	65.8%	97.5%	33.7%	7.31%	3.70%
	94.5%	66.0%	97.5%	33.4%	8.13%	3.75%
	93.5%	66.4%	97.4%	33.1%	8.97%	3.80%
	92.5%	66.8%	97.4%	33.0%	9.79%	3.84%
	91.5%	66.8%	97.4%	32.8%	10.6%	3.86%
	90.5%	67.0%	97.3%	32.7%	11.5%	3.88%
DL-3 + readers	97.5%	66.0%	97.6%	34.4%	5.48%	3.64%
	96.5%	66.4%	97.5%	34.0%	6.23%	3.70%
	95.5%	66.4%	97.5%	33.6%	7.02%	3.75%
	94.5%	66.4%	97.4%	33.3%	7.84%	3.78%
	93.5%	66.4%	97.4%	32.8%	8.66%	3.84%
	92.5%	66.6%	97.3%	32.6%	9.55%	3.87%
	91.5%	66.6%	97.3%	32.3%	10.40%	3.90%
	90.5%	67.0%	97.3%	32.2%	11.27%	3.94%

Table 6-7 – Perturbation analysis when adjusting the specificity threshold for the artificial intelligence (AI) algorithm, then combining with the first reader and final action arbitration decision if there is discordance.

6.4.3 Scenario D 99.0% specificity auto recall threshold

Including the auto recalled cases above the identified 99.0% specificity threshold of each AI algorithm resulted in an overall increase in sensitivity and decrease in specificity, Table 6-8. On average sensitivity increased by +0.8~+3.4%% and specificity decreased by -0.8~-2.3%, compared to

the results in Table 6-6 where the 99.0% specificity AI threshold was not implemented. This is also reflected in the increased recall rate (Δ +0.8~+2.3%) and decreased arbitration rate (Δ -0.9~-2.4%). There was an overall increase in the NRCs (Δ +1.8%~+9.9%), and ICs (Δ +2.3%~+9.7%) detected due to the auto recall implementation, thus Scenario D facilitates the earlier detection of cancer at the expense of an increased recall rate.

	Double reader	First reader + DL-1	First reader + DL-2	First reader + DL-3
Sensitivity	67.4% [63.1-71.5]	70.4% [66.2-74.3]	66.4% [62.1-70.5]	67.4% [63.1-71.5]
Specificity	97.1% [96.7-97.3]	95.1% [94.9-95.4]	96.6% [96.3-96.8]	96.8% [96.6-970]
Precision	31.3%	21.9%	27.2%	28.9%
Arbitration	2.7%	7.1%	5.2%	4.3%
Recall Rate	4.1%	6.1%	4.6%	4.4%
Cases Flagged				
Total	-	947 (3.5%)	530 (2.0%)	533 (2.0%)
True Positive	-	274 (1.0%)	235 (0.9%)	272 (1.0%)
False Positive	-	673 (2.5%)	295 (1.1%)	261 (1.0%)
Cancers Flagged				
SDC n (%)	-	254 (76.5%)	229 (69.0%)	258 (77.7%)
IC n (%)	-	20 (11.5%)	6 (3.5%)	14 (8.1%)
NRC n (%)	-	24 (9.4%)	10 (3.9%)	8 (3.1%)
FRC n (%)	-	1 (100%)	0 (0.0%)	0 (0.0%)
NRIC n (%)	-	1 (3.6%)	1 (3.6%)	0 (0.0%)
Auto Recalled				
SDC n (%)	-	20 (6.0%)	15 (4.5%)	18 (5.4%)
IC n (%)	-	17 (9.8%)	4 (2.3%)	10 (5.8%)
NRC n (%)	-	21 (8.3%)	9 (3.5%)	7 (2.8%)
FRC n (%)	-	1 (100%)	0 (0.0%)	0 (0.0%)
NRIC n (%)	-	1 (3.6%)	1 (3.6%)	0 (0.0%)
Final Total Detected				
SDC n (%)	332 (100%)	326 (98.2%)	323 (97.3%)	321 (96.7%)
IC n (%)	9 (5.2%)	30 (17.2%)	13 (7.5%)	20 (11.5%)
NRC n (%)	4 (3.6%)	21 (8.3%)	9 (3.5%)	7 (2.8%)
FRC n (%)	6 (4.2%)	1 (100%)	0 (0.0%)	5 (3.5%)
NRIC n (%)	2 (7.1%)	1 (3.6%)	1 (3.6%)	0 (0.0%)

Table 6-8 – Artificial intelligence (AI) algorithm (at threshold 2) combined with the single first reader (+/- arbitration where discordance below 99.0% specificity for the algorithm and above 96.6% specificity) with cases auto recalled above the 99.0% specificity threshold (threshold 3) compared to double reading performance. *FRC: Future round cancer, IC: Interval cancer, NRC: Next round, NRIC: Next round interval cancer, SDC: Screen detected cancer.*

6.4.4 Combined algorithm results

The Combined model was created by combining DL-1, DL-2 and DL-3 AI algorithms. The Combined model achieved a performance of AUROC 0.886 [95% CI 0.860–0.912], with an improvement in AUROC of Δ +0.002~+0.014, and pAUROC of 0.761 [95% CI 0.733–0.793]. At the 96.6% specificity

threshold for the first reader specificity, the Combined model achieved a sensitivity of 61.4% [95% CI 54.7 – 67.4]. Figure 6-8 compares the ROC curves of the Combined model to, DL-1, DL-2 and DL-3 on the Cambridge data.

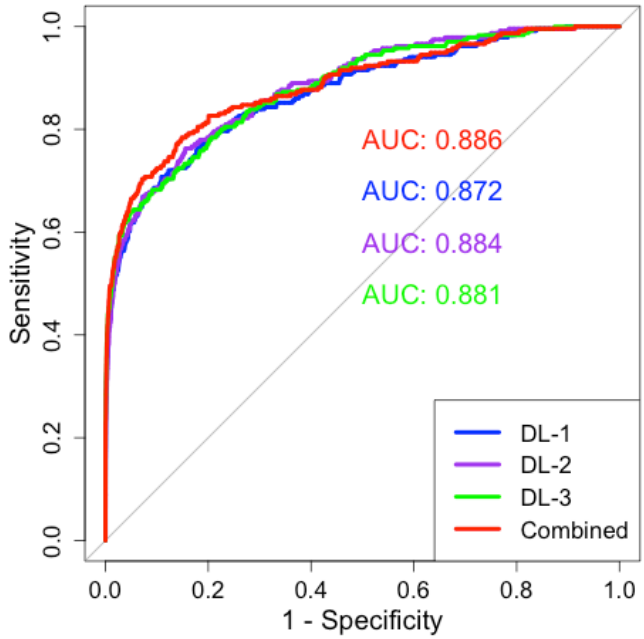


Figure 6-8 – Combined model receiver operating characteristic (ROC) curves on Cambridge data. *For DL-1 in blue, DL-2 in purple, DL-3 in green and the Combined algorithm performance in red, with area under the receiver operating characteristic curve values provided for each algorithm.*

The Combined model performance was not statistically significant from each individual AI algorithm performance ($p > 0.05$), as demonstrated in Table 6-9.

	DL-1	DL-2	DL-3
DeLongs test p value	0.4642	0.9093	0.806

Table 6-9 – DeLong's test comparison results for DL-1, DL-2, DL-3 compared to the Combined model performance on Cambridge data.

Taking the Combined model and then applying the model to Norwich data, found there was no overfitting of the model and that the model was generalisable to a different site using a different machine vendor (GE), achieving an AUROC of 0.902 [95% CI 0.880–0.925] and pAUROC of 0.783 [95% CI 0.756–0.810]. Figure 6-9 compares the ROC curves of the Combined model to, DL-1, DL-2 and DL-3 on Norwich data. Applying the 96.6% specificity threshold found on the Cambridge data using the Combined model, the Combined model on the Norwich data achieved a 99.8% [95% CI 99.8–99.9] specificity and 37.8% [95% CI 33.0–43.0] sensitivity.

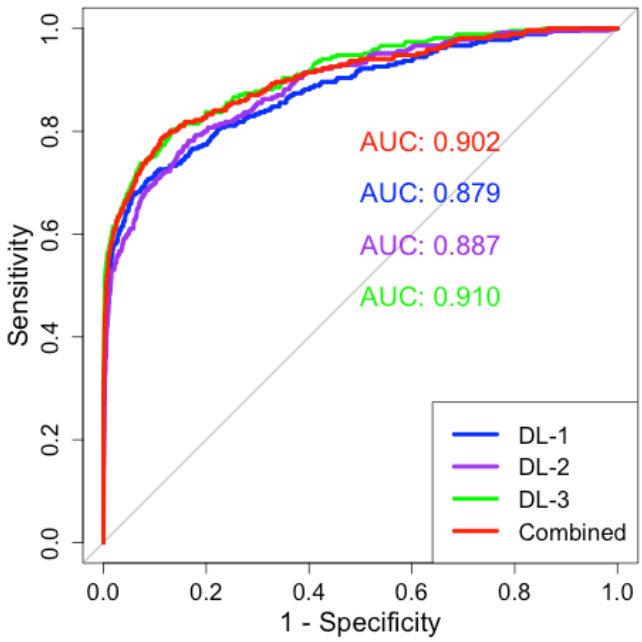


Figure 6-9 – Combined model receiver operating characteristic (ROC) curves on Norwich data. *For DL-1 in blue, DL-2 in purple, DL-3 in green and the Combined algorithm performance in red, with area under the receiver operating characteristic curve values provided for each algorithm.*

The Combined model performance was not statistically significant from DL-2 and DL-3 performance ($p > 0.05$). However, it was statistically significantly different from DL-1 as demonstrated in Table 6-10.

	DL-1	DL-2	DL-3
DeLongs test p value	< 0.01	0.2434	0.6288

Table 6-10 – DeLong's test comparison results for DL-1, DL-2, DL-3 compared to the Combined model performance on Norwich data.

6.4.5 Sub-group analysis

Performance of the AI algorithms was further assessed at the 96.6% specificity threshold (threshold 1) for sensitivity of the following subgroups; age at screening, mammographic machine vendor, invasive status, invasive size of cancer, invasive grade of cancer, and mammographic breast density categories for SDCs, Table 6-11, and ICs, Table 6-12, at all sites.

	n	First reader	DL-1	DL-2	DL-3
Total SDC Cases	332	302	266	266	286
Total Lesions$^{\delta}$	343	315	278	278	297
Total Invasive Lesions$^{\delta}$	286	266	238	240	253
Age at Screening					
< 60	113 (34.0%)	105 (92.9%)	88 (77.9%)	89 (78.8%)	96 (85.0%)
>= 60	219 (66.0%)	197 (90.0%)	178 (81.3%)	177 (80.8%)	190 (86.8%)
p value	-	0. 846319	0.806241	0.88204	0.902054
FFDM Vendor					
GE	184 (55.4%)	169 (91.8%)	159 (86.4%)	148 (80.4%)	160 (87.0%)
Philips	148 (44.6%)	133 (89.9%)	107 (72.3%)	118 (79.7%)	126 (85.1%)
p value	-	0.891552	0.284815	0.9576	0.896299
Invasive status$^{\delta}$					
Invasive	286 (83.3%)	266 (93.0%)	238 (83.2%)	240 (83.9%)	253 (88.5%)
Non-invasive	54 (15.7%)	46 (85.2%)	39 (72.2%)	37 (68.5%)	43 (79.6%)
Missing	3 (0.9%)	3 (100%)	1 (33.3%)	1 (33.3%)	1 (33.3%)
p value	-	0.916924	0.599303	0.493973	0.616753
Invasive Tumour Size$^{\delta}$					
< 15 mm	160 (55.9%)	144 (90.0%)	126 (78.8%)	133 (83.1%)	135 (84.4%)
>= 15 mm	120 (42.0%)	116 (96.7%)	106 (88.3%)	101 (84.2%)	112 (93.3%)
Missing	6 (2.1%)	6 (100%)	6 (100%)	6 (100%)	6 (100%)
p value	-	0.911418	0.772362	0.951474	0.82882
Invasive Tumour Grade$^{\delta}$					
1	67 (23.4%)	60 (89.6%)	55 (82.1%)	55 (82.1%)	57 (85.1%)
2	157 (54.9%)	147 (93.6%)	133 (84.7%)	134 (85.4%)	141 (89.8%)
3	56 (19.6%)	54 (96.4%)	45 (80.4%)	46 (82.1%)	50 (89.3%)
Missing	6 (2.1%)	5 (83.3%)	5 (83.3%)	5 (83.3%)	5 (83.3%)
p value	-	0.989649	0.996254	0.997324	0.994561
Density BI-RADS$^{\beta}$					
a	18 (5.4%)	17 (94.4%)	11 (61.1%)	13 (72.2%)	15 (83.3%)
b	79 (23.8%)	71 (89.9%)	59 (74.7%)	66 (83.5%)	71 (89.9%)
c	43 (13.0%)	39 (90.7%)	34 (79.1%)	36 (83.7%)	37 (86.1%)
d	12 (3.6%)	10 (83.3%)	6 (50.0%)	7 (58.3%)	7 (58.3%)
p value	-	0.996736	0.817878	0.889323	0.860564
Density BI-RADS$^{\alpha}$					
a	45 (13.6%)	42 (93.3%)	31 (68.9%)	35 (77.8%)	38 (84.4%)
b	208 (62.7%)	190 (91.3%)	174 (83.7%)	172 (82.7%)	185 (88.9%)
c	78 (23.5%)	69 (88.5%)	61 (78.2%)	59 (75.6%)	63 (80.8%)
d	1 (0.3%)	1 (100%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
p value	-	0.997149	-	-	-

Table 6-11 – Sub group analysis of DL-1, DL-2, DL-3 set at the first reader specificity threshold of 96.6% (threshold 1) for screen detected cancers (SDCs). *BI-RADS: Breast imaging-reporting and data system, FFDM: Full field digital mammography, SDC: Screen detected cancer.* $^{\delta}$Lesions reported. $^{\alpha}$*DL-3 5th edition BI-RADS density scores on processed full field digital mammograms.* $^{\beta}$*Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data. p values were determined by using Chi squared* χ^2 *test to compare against the detected proportion of cancers cases by the true distribution for each cancer characteristic category. p values < 0.05 were considered statistically significant.*

	n	Double reader	First reader	DL-1	DL-2	DL-3
Total IC Cases	174	9	16	26	25	30
Total Lesions$^{\delta}$	186	10	17	31	25	32
Total Invasive Lesions$^{\delta}$	167	7	16	30	19	30
Age at Screening						
< 60	86 (49.4%)	3 (3.5%)	4 (4.7%)	13 (15.1%)	12 (14.0%)	13 (15.1%)
>= 60	88 (50.6%)	6 (6.8%)	12 (13.6%)	13 (14.8%)	13 (14.8%)	17 (19.3%)
p value	-	0.346281	0.061133	0.956401	0.893963	0.537597
FFDM Vendor						
GE	91 (52.3%)	3 (3.3%)	7 (7.7%)	18 (19.8%)	9 (9.9%)	9 (9.9%)
Philips	83 (47.7%)	6 (7.2%)	9 (10.8%)	8 (9.6%)	16 (19.3%)	21 (25.3%)
p value	-	0.266994	0.512593	0.105847	0.127486	0.024046
Invasive status$^{\delta}$						
Invasive	167 (89.8%)	7 (4.2%)	16 (9.6%)	30 (18.0%)	19 (11.4%)	30 (18.0%)
Non-invasive	15 (8.1%)	2 (13.3%)	1 (6.7%)	1 (6.7%)	5 (33.3%)	2 (13.3%)
Missing	4 (2.2%)	1 (25.0%)	0 (0.0%)	0 (0.0%)	1 (25.0%)	0 (0.0%)
p value	-	0.118297	0.732246	0.327375	0.128396	0.700804
Invasive Tumour Size$^{\delta}$						
< 15 mm	52 (31.1%)	4 (7.7%)	5 (9.6%)	10 (19.2%)	4 (7.7%)	6 (11.5%)
>= 15 mm	96 (57.5%)	3 (3.1%)	11 (11.5%)	19 (19.8%)	14 (14.6%)	24 (25.0%)
Missing	19 (11.4%)	0 (0.0%)	0 (0.0%)	1 (5.3%)	1 (5.3%)	0 (0.0%)
p value	-	0.236243	0.756546	0.401431	0.394622	0.106786
Invasive Tumour Grade$^{\delta}$						
1	28 (16.8%)	1 (3.6%)	2 (7.1%)	4 (14.3%)	3 (10.7%)	4 (14.3%)
2	73 (43.7%)	5 (6.9%)	9 (12.3%)	21 (28.8%)	13 (17.8%)	18 (24.7%)
3	63 (37.7%)	1 (1.6%)	5 (7.9%)	5 (7.9%)	3 (4.8%)	8 (12.7%)
Missing	3 (1.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
p value	-	0.34277	0.663023	0.029639	0.105161	0.291008
Density BI-RADS$^{\beta}$						
a	8 (4.6%)	0 (0.0%)	1 (12.5%)	0 (0.0%)	1 (12.5%)	3 (37.5%)
b	34 (19.4%)	3 (8.8%)	3 (8.8%)	0 (0.0%)	4 (11.8%)	6 (17.6%)
c	29 (16.6%)	3(10.3%)	4 (13.8%)	5 (17.2%)	7 (24.1%)	7 (24.1%)
d	12 (6.9%)	0 (0.0%)	1 (8.3%)	3 (25.0%)	4 (33.3%)	5 (41.7%)
Missing	1 (0.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
p value	-	-	0.939331	-	0.518521	0.58906
Density BI-RADS$^{\alpha}$						
a	13 (7.4%)	0 (0.0%)	1 (7.7%)	0 (0.0%)	1 (7.7%)	4 (30.8%)
b	94 (53.7%)	5 (5.3%)	9 (9.6%)	17 (19.1%)	14 (14.9%)	14 (14.9%)
c	65 (37.7%)	4 (6.1%)	6 (9.1%)	9 (13.6%)	10 (15.2%)	12 (18.2%)
d	2 (1.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
p value	-	-	-	-	-	-

***Table 6-12 – Sub group analysis of DL-1, DL-2, DL-3 set at the first reader specificity threshold of 96.6% (threshold 1) for interval cancers (IC**). BI-RADS: Breast imaging-reporting and data system, FFDM: Full field*

digital mammography, IC: Interval cancer. [δ]Lesions reported. [α]DL-3 5th edition BI-RADS density scores on processed full field digital mammograms. [β] Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data. p values were determined by using Chi squared χ^2 test to compare against the detected proportion of cancers cases by the true distribution for each cancer characteristic category. p values < 0.05 were considered statistically significant.

The AI algorithms behaved similarly to true distribution of cancer cases in the all types of cancers detected ($p > 0.05$). A statistically significant difference between the distribution of ICs invasive grade for DL-1, and ICs mammographic machine vendor for DL-3 was found. Otherwise, no statistically significant difference was found between the distribution of each sub category and the types of cancers detected by human readers or AI algorithms. The AI algorithm performance is similar to human reader performance, and reduces in sensitivity as density increases. Performance of the AI algorithms was further assessed at the 96.6% specificity threshold (threshold 1) for sensitivity of the following IC subgroups; interval time in months and radiological audit classification, Table 6-13. The AI algorithms followed a similar distribution to the true distribution of IC over the interval time period (months) and detected more year three cancers than human readers. In addition, the algorithm like human readers picked up more uncertain cases, where there was potentially a visible sign "seen with hindsight", than normal / benign cases where there was no visible sign on case review.

	n	Double reader	First reader	DL-1	DL-2	DL-3
Total IC Cases	174	9	16	26	25	30
Interval (Months)						
0-12	31 (17.7%)	4 (12.9%)	6 (19.4%)	3 (9.7%)	4 (12.9%)	5 (16.1%)
12-24	51 (29.1%)	5 (9.8%)	10 (19.6%)	7 (13.7%)	5 (9.8%)	8 (15.7%)
24-36	92 (52.6%)	0 (0.0%)	0 (0.0%)	16 (17.4%)	16 (17.4%)	17 (18.5%)
36-40	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
p value	-	-	-	0.642967	0.545267	0.927792
Radiological Audit Classification						
Normal/Benign	129 (74.1%)	5 (3.9%)	9 (7.0%)	14 (10.9%)	16 (12.4%)	21 (16.3%)
Uncertain	39 (22.3%)	3 (7.7%)	7 (18.0%)	12 (30.8%)	8 (20.5%)	9 (23.1%)
Suspicious	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Unclassifiable	1 (0.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Missing	5 (2.9%)	1 (20.0%)	0 (0.0%)	0 (0.0%)	1 (20.0%)	0 (0.0%)
p value	-	-	-	-	-	-

Table 6-13 – Sub group analysis of DL-1, DL-2, DL-3 set at the first reader specificity threshold of 96.6% (threshold 1) for interval cancer (IC) specific categories. *IC: Interval cancer. p values were determined by using Chi squared χ^2 test to compare against the detected proportion of cancers cases by the true distribution for each cancer characteristic category. p values < 0.05 were considered statistically significant.*

Furthermore, the overlap of cases detected by each algorithm (DL-1, DL-2, DL-3) and the human first reader, at threshold 1, is shown in Figure 6-10. The Venn diagram demonstrates how the majority of SDC cases overlap for both the human reader and AI algorithms. Whereas, the IC and NRC cases detected differ between the AI algorithms as well as between the human first reader and AI algorithms.

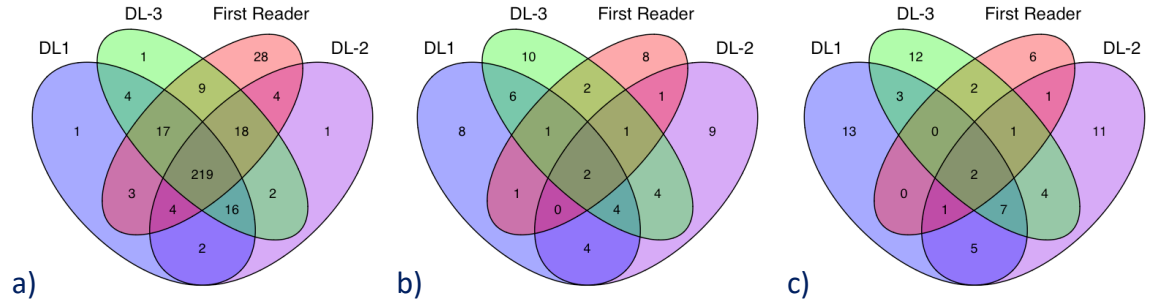


Figure 6-10 – Venn diagram – not proportional. *a) Screen detected cancer cases, b) interval cancer cases, c) next round cancer cases. For DL-1 in blue, DL-2 in purple, DL-3 in green and the first human reader in red.*

6.4.6 Failure analysis

Examples of cases missed by either the human readers, AI algorithms or both are shown below. A case classified as an uncertain IC that was not detected by all methods, human readers and AI algorithms is shown in Figure 6-11. This was a case of a 57-year-old patient, diagnosed with a left sided grade 2, 15 mm invasive cancer, 691 days after screening.

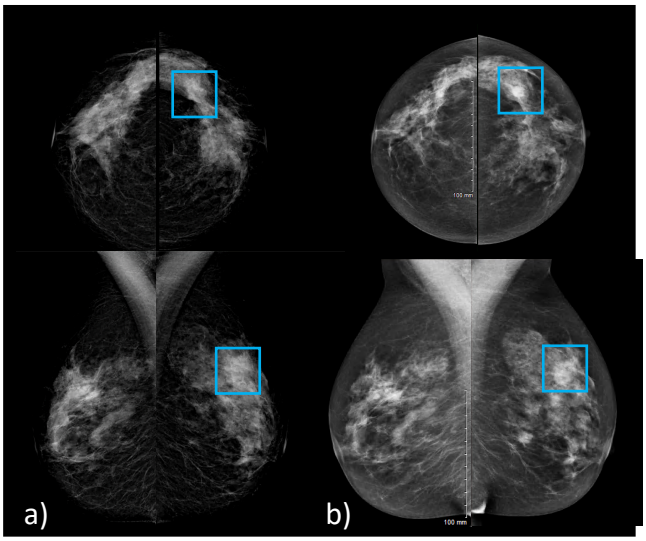


Figure 6-11 – Missing case analysis, case missed by both artificial intelligence (AI) and human readers. *a) Screening image b) diagnostic image, with a blue bounding box to show the location of the cancer. The screen and diagnostic images were annotated by a breast radiologist to show the true location of the cancer.*

A case classified as a normal / benign IC that was not detected by all human readers, but was recalled by all AI algorithms is shown in Figure 6-12. This was a case of a 57-year-old patient, diagnosed with a right sided grade 2, 21 mm invasive cancer, 806 days after screening.

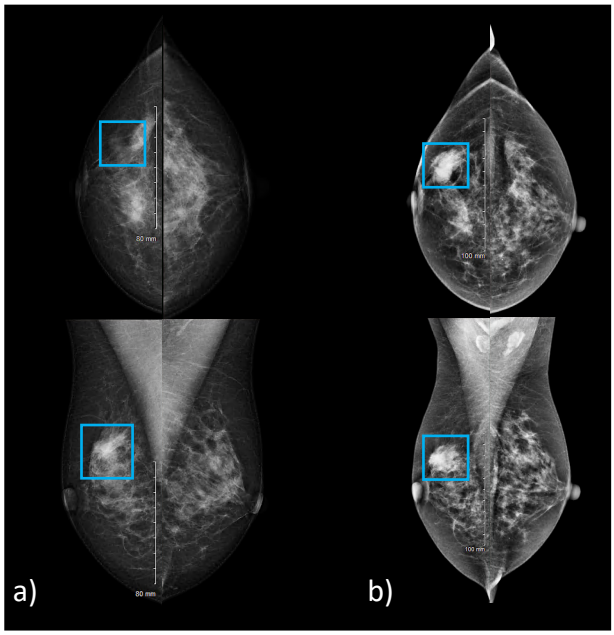


Figure 6-12 – Missing case analysis, case missed by all human readers and detected by all artificial intelligence (AI) algorithms. *a) Screening image b) diagnostic image, with a blue bounding box to show the location of the cancer. The screen and diagnostic images were annotated by a breast radiologist to show the true location of the cancer.*

A SDC case recalled at routine screening by all readers that was not detected by all AI algorithms, Figure 6-13. This was a case of a 51-year-old patient, diagnosed with a left sided grade 3, 3 mm invasive cancer, with a 107 mm non-invasive component.

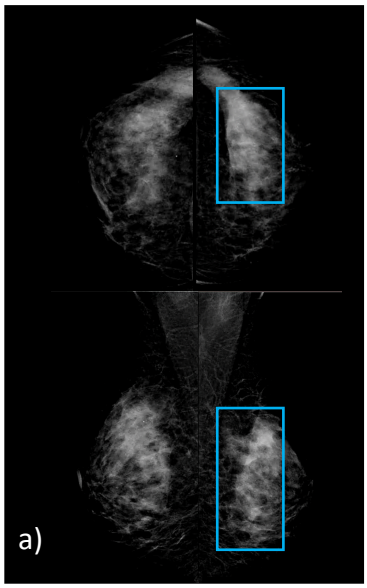


Figure 6-13 – Missing case analysis, case missed by all artificial intelligence (AI) algorithms. *a) Screening image, with a blue bounding box to show the location of the cancer. The screen images were annotated by a breast radiologist to show the true location of the cancer.*

6.5 Discussion

6.5.1 Overall performance

This study aimed to evaluate the performance of three commercial AI algorithms as stand-alone systems for the task of detection and diagnosis (CADe+x) in routine UK breast screening, using a large unenriched multi-vendor retrospective dataset from two UK NHSBSP sites. It provides an independent external validation which has not previously been performed, on UK data, for multiple algorithms simultaneously[133,138,235,236,292,293].

Overall all three AI algorithms achieved a good AUROC 0.868–0.910, pAUROC 0.737–0.791 and AUPRC 0.407–0.513 when using cancers diagnosed within 3 years of screening as cases, demonstrating that these algorithms are generalisable to the UK screening population across different sites and mammographic machine vendors. The AUROC and pAUC of DL-3 is statistically significantly different than DL-1 and DL-2 when tested on Norwich data, which is likely due to the predominant manufacturer used for training DL-3 (GE) is the same as the manufacture in the Norwich test set. Interestingly, the AUROC was not statistically significantly different between sites for the same AI algorithm, despite the algorithms either training on no or < 1% Philips data. Generalisability is further demonstrated as all the AI algorithms trained on less than 10% of UK data (triennial screening programme).

This study highlights the importance of reporting, AUROC alongside, pAUROC, AUPRC, sensitivity and precision, as the groups are unbalanced in screening with a large proportion of normal cases to cancer cases. Additional metrics of AUPRC, precision and sensitivity provide information regarding the cost trade off, such that there is a high cost for missing a cancer case, which is captured in these metrics, and is demonstrated in Figure 6-6 for DL-3 and DL-2 where the precision is significantly reduced at a high recall for either missing a cancer case or high rates of false positive recalls. The pAUROC allows for the evaluation of an AI algorithms performance at the extreme end of the curve, high specificity, where an algorithm operates for screening tasks to maintain recall rates and so provides a more accurate assessment of clinical performance compared to the overall AUROC.

Compared to the first reader, the sensitivity of all three algorithms were shown to be non-inferior at both threshold 1 and threshold 2 (first reader specificity 96.6%). The AI algorithms detected between 13.2-23.6% ICs and 4.5-28.8% NRCs which may offset the reduction in SDCs seen when using these systems as stand-alone readers. This is in keeping with previous studies where AI algorithms have been shown to be non-inferior and in certain cases superior to the first reader in double reading biennial and triennial screening programmes as well as in single reader annual programmes[137,138,149,290]. Rates of ICs detected were lower than in recent studies where 30.7% (63/205) of ICs were detected in the biennial screening programme of Norway, using a cohort of >

47,000 women, although this was at a higher recall rate of 5.8%[271]. In a study using a ten year UK and Hungarian screening cohort, 29.8% (111/373) of ICs were detected, although again the AI system was operating at a lower specificity (91.2%)[290].

When the algorithm is set at threshold 2 and combined with the first reader decision and final action decision where discordance, all of the three AI algorithms were non-inferior to double reading performance. However, there was an increase arbitration rate with a decrease in SDC rate and maintenance of IC detection. The decrease in recall rate and overall reduction in workload if this approach was implemented potentially provides a trade-off to the cancer detection and arbitration rate effects. Sharma *et al* reported a similar non-inferior sensitivity AI algorithm performance in a cohort of UK and Hungarian screening data, as well as similar increase in the arbitration rate[290].

Deployment of AI algorithms as the second reader is seen as a favourable initial deployment approach with a 'reader in the loop' for oversight of the algorithm's decisions, however the trade-off of reducing the workload of one reader, whilst significantly increasing arbitration, needs to be addressed as to what is an acceptable national level of increase in arbitration as well as who takes part in this arbitration and what information needs to be provided by the AI algorithm to arbitration readers. It is also important recall rates remain the same as existing screening standards so as not to increase the workload of assessment clinics, which are already a workload intensive and costly part of any screening programme.

6.5.2 Further analysis

The additional scenario of implementing an auto recall threshold (99.0% specificity), aims to overcome the bias caused by using the original arbitration decision of human readers, as a case can only be recalled if the overall human reader decision was to recall the case. At this threshold there was an overall increase in earlier detection of cancer (ICs and NRCs) at the expense of an increased recall rate. However, it is unknown if the ICs and NRCs recalled would be detected at an assessment clinic or with supplemental imaging.

Combining all three AI algorithms did not result in a statistically significant improvement in performance. Salim *et al* also found using a voting system of three different commercial AI algorithms did not improve performance compared to the best performing algorithm[149]. However in Schaffter *et al*, they implemented an ensemble method of the top performing eight algorithms as part of the DREAM challenge, and did show superior performance compared to the single best performing algorithm[137]. It was also suggested in the UK National Screening Committee report that using algorithms together could potentially improve overall performance[136,137]. Interestingly as shown in Figure 6-10.b the ICs and NRCs detected by each AI algorithm and human readers are

different and thus potential benefit for the early detection of cancers could be found by using these systems together.

Investigating the consistency of performance across different categories showed the algorithms detected cancers with a similar distribution to the true distribution across all sub groups. In addition AI algorithms demonstrated similar behaviour to human readers with a decrease in performance at the highest breast density category, which has previously been reported[138,149,290].

6.5.3 Limitations

There are limitations to this study. Firstly, comparing three yearly performance disadvantages the human reader as it provides the AI algorithms with the opportunity to detect cancers that were not detected by the human readers. Secondly, in practice human readers have access to both prior images and clinical information, which could disadvantage the AI algorithms. Recent developments have seen algorithms starting to use prior images within their decision-making process, and this information was not available in this study. In addition, as all the algorithms in this study are commercial, they are reported under a pseudonym (DL-ID). Whilst this limits the transparency of reporting certain parameters (e.g. model weights and layers) for reproducibility, it does provide an oversight as to the current performance of commercial AI algorithms for programme level decisions and thus evidence for the implementation of this technology as well as the planning of prospective studies. Part of this study uses simulation to estimate the performance of the AI as the second reader, as noted in the recently updated 2021 UK NSC report, simulation studies are unable to "measure the impact of AI on readers and their decisions". Ethnicity data was missing for a proportion of cases, when searching NBSS and Electronic Health Record (EHR) systems and so the assessment of AI algorithms for consistent non-biased performance based on case ethnicity was not possible. Histopathological size can be influenced by the use of neo-adjuvant chemotherapy, and this information was not commonly available alongside size information for analysis. Finally, two out of the three algorithms had access to UK The Optimam Mammography Image Database (OMI-DB) data, which includes a small proportion of data form Cambridge. Whilst, all time points for these cases were identified and removed from this study testing set, the Cambridge data is not wholly temporally independent from the training sets used by each AI algorithm and there is the potential for bias.

6.5.4 Future work

Further work should include evaluating the lesion level prompts provided by each algorithm to investigate the explainability as well as the possibility of use of algorithms as interactive clinical decision support systems. In addition, the development of the database over a ten-year period will

allow for the inclusion of prior mammogram information for AI algorithms which could result in an improvement in performance.

6.6 Conclusion

In conclusion all of the three commercial AI algorithms met the required benchmark of non-inferiority for the detection and diagnosis of breast cancer as a stand-alone single screen reader and in conjunction with a human reader in a double reading system. Thus, all of the three algorithms are suitable to proceed to prospective assessment. Further work is however required to confirm bias does not occur for certain patient groups, through the evaluation of AI algorithm performance for different ethnicities and in different socio-economic regions of the country as part of prospective studies.

Chapter 7 - Performance of stand-alone artificial intelligence algorithms in a UK screening cohort for high sensitivity and high specificity triage

7.1 Aims

In this chapter, the performance of three commercial artificial intelligence (AI) algorithms is investigated for high sensitivity and high specificity triage applications. A large representative screening cohort from two UK screening sites is used for this study in order to assess the tools performance at a high sensitivity for normal case rule out triage. In addition, each AI algorithm was evaluated for a high suspicious high specificity rule in triage application, for the detection of interval and next round cancers. A combined approach for both rule in rule out triage was then applied using the thresholds identified in the earlier studies. The results from this chapter provide data for planning prospective trials and adds to the UK evidence for investigating the use of AI algorithms for triage applications in breast cancer screening.

Contents of this chapter have been submitted to Radiological Society of North America conference 2022 [abstract ID - #2022-SP-2966-RSNA] and European Society of Breast Imaging conference 2022 [abstract ID - #A-165].

7.2 Introduction

Each year more than 2.5 million women are screened using mammography as part of the National Health Service (NHS) Breast Screening Programme (BSP), and an estimated 15,000 cancers are diagnosed, such that an estimated ~99.0% of women screened will not have a cancer at the time of screening[54]. Thus the vast majority of the screening workload is from 'normal' screens. Screening programmes like the NHSBSP employ a double reading system, where each case is read by two radiologists and if there is discordance between readers the case is arbitrated. Mammographic screen reading is therefore a repetitive task of high volume, which is prone to reader fatigue[294]. Many countries have also reported a scarcity of radiologists, especially in breast imaging[58]. Therefore solutions to improve the efficiency of screening are of interest for programmes. One solution would be to reduce the readers' workload and not have to read mammograms with a very low likelihood of a cancer by using an AI algorithm for automated 'normal' case triage. Mammograms below a threshold could be automatically assigned a 'normal' outcome and not read by a human reader or only read by one reader in a double reading programme[134,135,229]. In our systematic review and meta-analysis reported in Chapter 3 we found when applying this computer aided triage (CADt) approach

the number of exams could be reduced from 17.0-91.0% whilst missing 0.0-7.0% of cancer cases[133]. A recent study by Lång *et al*, found 19.1% of cases could be removed without missing a cancer. If 53.0% of cases were classified as normal, 10.3% of screen detected cancers (SDCs) would not be flagged of which 85.7% (6/7) were clearly visible, with a 27.8% reduction in false positive recalls[295]. What has not been fully quantified is the acceptable miss rate of these systems when used for this specific application, such as what sensitivity threshold should be used when setting the operating point of these systems[133]. Alternative triage reading approaches, have been suggested such that cases with the lowest scores are single read and the rest are double read. Using this approach Balta *et al* demonstrated a 32.6% reduction in workload for the second reader whilst estimated to miss no cancers[230]. Balta *et al* also found a reduction in recall rate (5.35% to 4.79% ($p < 0.01$)), a reduction in arbitration rate by 20.8% and an increase in positive predictive value (11.9% to 13.3% ($p < 0.01$))[230]. However, it is unknown if this would be replicated in the real-time clinical workflow and if reader performance would improve or at a minimum stay the same. Concerns raised are if there will be adverse effects if reading a smaller volume of exams and the impact on reader training to maintain the high standards of breast screening through exposure to different cases, both cancer and non-cancer.

An alternative method for stand-alone AI triaging is to triage highly suspicious cases with a high score for either automatic referral for assessment or supplemental imaging. This auto CADt rule in approach could improve the detection of interval (IC) and next round (NRC) cancers, thus potentially improving the survival outcomes of women through earlier detection. One approach suggested by Dembrower *et al* is for enhanced screening of those cases with the highest 1.0-5.0% scores using supplemental imaging (Digital Breast Tomosynthesis (DBT), Magnetic resonance imaging (MRI)), which estimated to increase the detection of ICs by 12.0-27.0% and NRCs by 14.0-35.0%[134]. Dembrower *et al* also incorporated a rule out triage approach which identified 60.0% of cases could be triaged out from human reading without missing a cancer[134]. Lauritzen *et al*, implemented both a normal rule out triage and a high suspicion auto recall to assessment triage. For the auto recall high suspicion triage only 0.08% of cases were recalled of which 8.8% were SDCs, 0.0% ICs and 0.14% NRCs. However, Lauritzen *et al* found an overall 63.0% workload reduction and 25.1% false positive reduction, whilst missing 12 (1.5%) SDCs when implementing the auto recall out threshold[296]. Overall, Lauritzen *et al* reported a non-inferior sensitivity (69.7% vs 70.8%) and specificity (98.6% vs 98.1%) when comparing the AI system workflow to the routine double reading workflow[296].

Such improvements in efficiency of reading could also benefit the patients by potentially providing faster results. The anxiety of waiting for screening results is often reported by patients attending screening. Furthermore, the overall cost effectiveness of the screening programme could be

improved by this strategic screening reading approach, through the reduction in the number of radiologists hours required for mammographic screen reading and instead utilising radiologists time for complex biopsies and time-consuming MRI reading.

This study addresses the gap in evidence highlighted in the UK National Screening Committee report, using a large external UK multi-vendor multi-site cohort for testing multiple AI algorithms in different triage approaches within an independent environment[136].

7.3 Methods

7.3.1 Data

All study data was obtained from the CC-MEDIA database described in Chapter 4, where data was collected from two NHSBSP sites (Cambridge and Norwich) under existing ethical approval (HRA REC 20/LO/0104, HRA CAG 20/CAG/0009, PHE RAC BSPRAC_090). All study data was de-identified prior to use in this research.

Cases were included if the following eligibility criteria was met; age more than 47 years old, complete two-view FFDM, took part in routine NHSBSP screening between January 1 2015 to December 31 2017 at Norwich and January 1 2017 to December 31 2018 at Cambridge. Cases were excluded if they were recorded as a technical recall, were part of high-risk screening, did not meet the specified case definition for ground truth, and any cases where there was an incomplete mammogram; less than four views, more than four views, images not available on Picture Archiving and Communication System (PACS) or only raw data was available. One time point per case was included, such that if a case appeared twice due to repeat screening within the study time frame the earliest time point was used for this study. Cancer cases were also removed where they did not meet the specified definition following discussion with Public Health England (PHE), and interval cancer cases were removed if the interval was recorded as longer than 40 months. Cases were not excluded if they had prior surgery, prior cancer or an artefact was included in the image (e.g. pacemakers). However, cases with implants were excluded. All cases were checked to ensure there was no overlap with any existing databases the tools had used for training and therefore the AI algorithms had not previously seen any data included in this study dataset.

The processed screening Full Field Digital Mammogram (FFDM) images were used by all AI algorithms for this study. The images were stored in Joint Photographic Experts Group (JPEG) Lossless Digital Imaging and Communications in Medicine (DICOM) format and no additional pre-processing other than that performed by the mammography vendor and that performed by the AI algorithm occurred. No prior images or clinical data was available for the algorithms to use.

Clinical metadata was collected for all cases included in this study from the National Breast Screening System (NBSS). The invasive status, histological grade, and histological size, was obtained using an automated NBSS query, for further detail please see Chapter 4 Section 4.4.5. The case selection process is shown in the Standards for Reporting of Diagnostic Accuracy Studies (STARD) diagram in Figure 7-1[275]. Cases were excluded if a sufficient ground truth follow-up was not available. In this study 34,889 cases were excluded as a second time point normal screen outcome was not available in the NBSS output. This is due to cases not returning to screening due to either non-attendance or they completed routine screening (aged 50-70) and did not self-refer. The study took place during the Covid-19 pandemic which also effected women being recalled to screening, as detailed in Chapter 4.

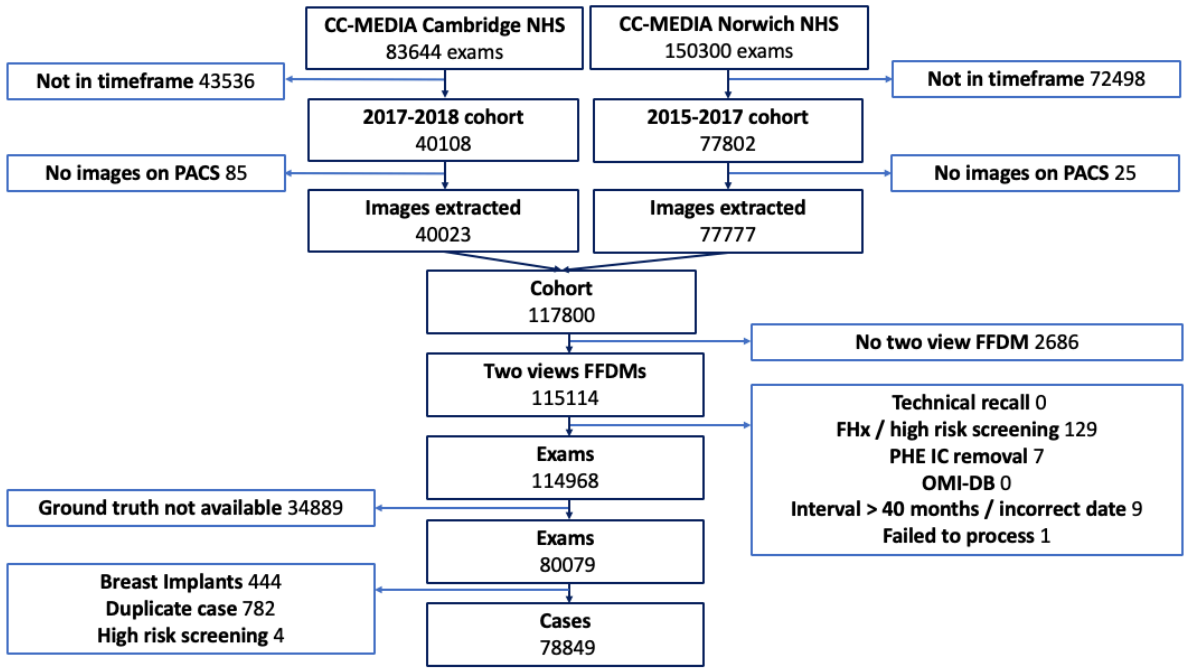


Figure 7-1 – Standards for Reporting of Diagnostic Accuracy Studies (STARD) flow diagram of cases included and excluded in this study. *FFDM: Full Field Digital Mammogram, FHx: Family history, IC: Interval cancer, NHS: National Health Service, OMI-DB: The Optimam Mammography Image Database, PHE: Public Health England, PACS: Picture Archiving and Communication System.*

7.3.2 Ground truth

The ground truth was determined for each case using definitions available from the NHSBSP as well as normal follow-up standards within the field. The study time period overlaps with the pause in screening during the Covid-19 pandemic, which lead to an increase in round length. For further details regarding the definition for the ground truth of each case please refer to Chapter 6 Section 6.3.3.

Figure 7-2 shows the sequence of different cancer outcomes. Previous round cancers and previous round interval cancers were only included if they had a further round outcome of either 'normal' or 'cancer' outcome.

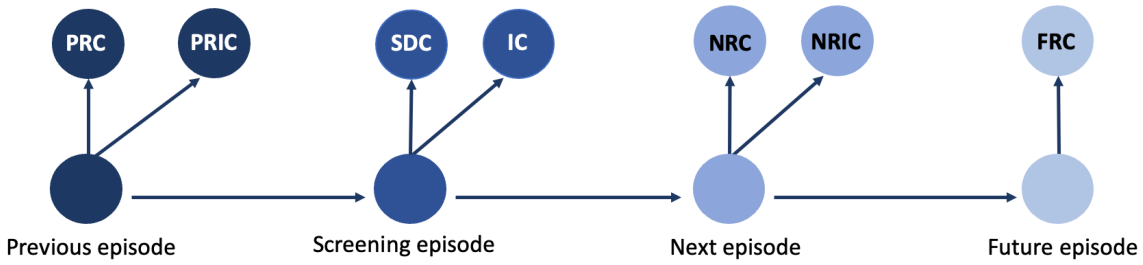


Figure 7-2 – Cancer outcomes for study cohort. *FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, PRC: Previous round cancer, PRIC: Previous round interval cancer, SDC: Screen detected cancer.*

The human readers were trained breast radiologists or breast radiographers who read as part of the NHSBSP, meeting the NHSBSP standards of reading 5000 mammograms a year and undertaking Personal Performance in Mammographic Screening (PERFOMS) testing each year[56]. Trainee readers were removed from this analysis.

7.3.3 AI tools

Three commercial AI algorithms, which use a deep learning (DL) convolutional neural network architecture, were installed at the University of Cambridge. Details regarding the training data used by each AI algorithm as well as the technical setup and algorithm output is outlined in Chapter 5 Table 5-1.

Breast Imaging-Reporting and Data System (BI-RADS) 5th edition density scores were provided from two systems; Volpara (research version – VolparaResearch32_L30Enabled_v2, Wellington, New Zealand) using raw full field digital mammography (FFDM) data, and DL-3 using FFDM processed data.

7.3.4 Thresholds

Four thresholds were used for the normal triage aspect of this study. The first threshold was set at 99.0% sensitivity for the AI algorithm (threshold 1) and the second threshold was set at 99.9% sensitivity (threshold 2), with SDCs classified as cases. The third threshold (threshold 3) was set at 85.0% sensitivity, and the fourth threshold was set at 70.0% specificity (threshold 4) for cancers occurring within 3 yearly screening (SDCs, ICs, NRCs). At these thresholds the AI algorithms performance was assessed for an adapted reader workflow, where an initial AI algorithm read takes place and if the case meets the threshold, then it is included in the alternative screening workflow, Figure 7-3.b or Figure 7-3.c. Scenario B, as shown in Figure 7-3.b, results in any case below the

threshold not being read by a human reader, whereas in Scenario C, Figure 7-3.c, the case is read by one reader (single first reader) only. Cases that do not meet this threshold proceed to routine double reading creating a simulated workflow.

For the high-suspicion rule in triage part of this study, the performance threshold was set at 94.0-99.0% specificity, for cancers occurring within 3 yearly screening (SDCs, ICs, NRCs). Two approaches are reported in this study, Scenario D (Figure 7-3.d) and Scenario E (Figure 7-3.e). In Scenario D any case above the AI algorithm threshold of 94.0-99.0% specificity is automatically recalled for supplemental imaging or assessment. Alternatively in Scenario E, any case above the threshold (94.0-99.0% specificity) not recalled by routine human reading would be referred for further supplemental imaging or assessment.

SDCs and ICs, occurring within the three-year screening interval, were classified as cancer cases for the calculation of overall sensitivity and specificity.

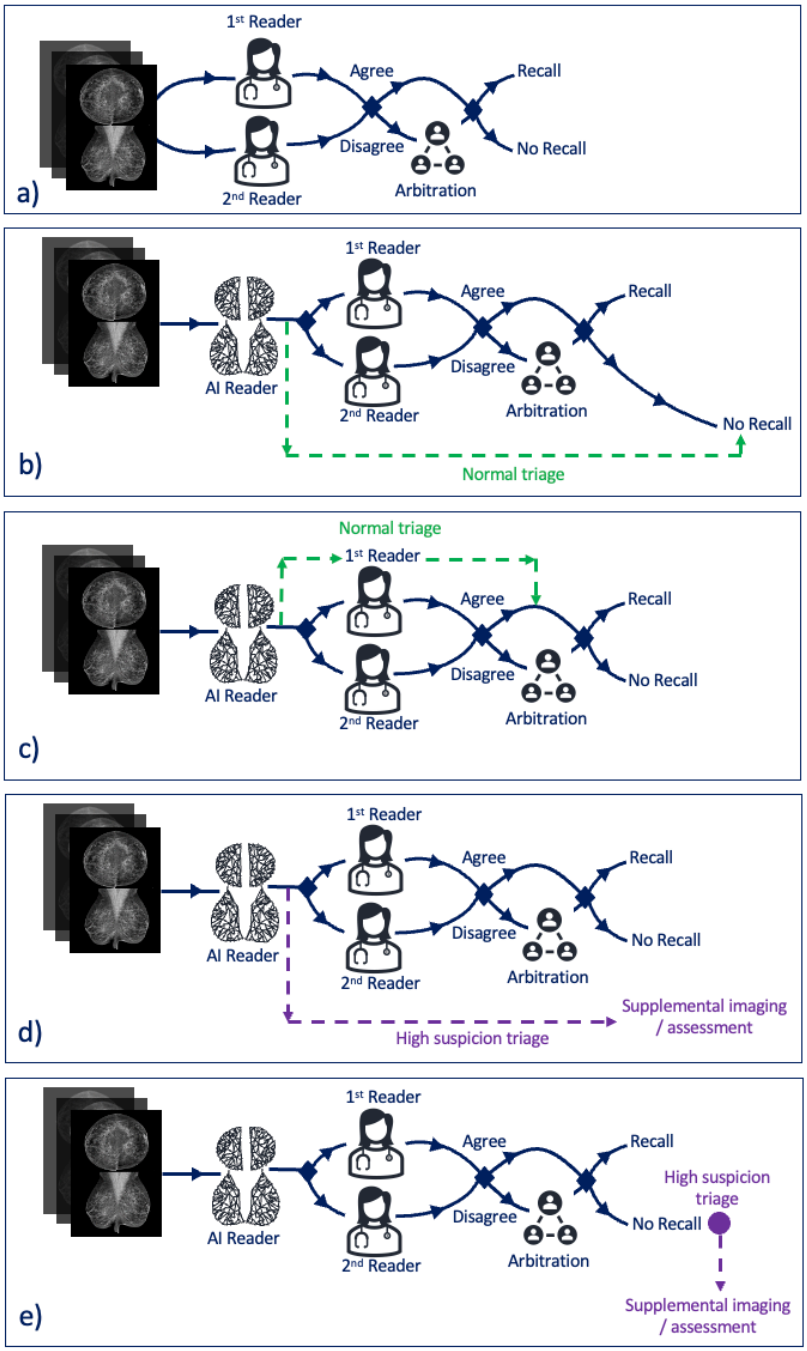


Figure 7-3 – Proposed workflow deployment approaches for stand-alone artificial intelligence (AI) systems as triage tools. *a) Routine UK double reading workflow, b) rule out normal triage of all cases below a set threshold, c) rule out normal triage of cases below a set threshold to single first reader reading, d) rule in high suspicion triage of all cases above a set threshold to supplemental imaging or assessment, e) rule in high suspicion triage of cases above a set threshold that were not recalled by routine double reading to supplemental imaging or assessment.*

7.3.5 Statistical analysis

All statistical analysis took place in R version 4.0.4 (R Foundation for Statistical Computing, Vienna, Austria)[225], using the packages detailed in Chapter 5 Section 5.3.6.

The overall predictive performance of each AI algorithm was evaluated using area under the receiver operating characteristic (AUROC) curve, and the partial AUROC (pAUROC) at 99.0-100% or 85.0-100% sensitivity and 94.0-99.0% specificity. The primary performance metrics for the normal triage study were; specificity, sensitivity, % cases triaged, % cancers missed due to normal triage and % false positive cases triaged. The primary performance metrics for the high-suspicious triage study were; sensitivity, specificity, % interval cancers detected, and % next round cancer detected. The effect on the overall recall rate and arbitration rate was also assessed for all triage approaches.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Performance of each AI algorithm was compared to readers performance, using one sample one tailed z-test to determine if the algorithm was non-inferior. Subgroup analysis of the SDCs missed by the AI algorithms as part of normal triaging at threshold 1 took place for the following categories; age at screening, mammographic machine vendor, invasive tumour size, invasive tumour grade, and mammographic breast density using both Cambridge and Norwich data. Further subgroup analysis in the same categories was calculated for the AI algorithms at the 94.0% specificity rule in triage threshold for ICs and NRCs. The true integer values and percentages were reported as well as Chi squared χ^2 test was used to investigate if there was a statistically significance between categories[285]. Data is presented as integer number and percentage (*n* (%)), or median and interquartile range (IQR) [25th – 75th centile range] as appropriate. DeLong's test was used to assess for a statistically significant difference between the AUROC curve of AI algorithms using 2000 bootstrapping examples. In all analyses, p-values < 0.05 were considered statistically significant and 95% confidence intervals were calculated, using bootstrapping with 2000 samples or through an approximation method from Simel *et al* using the epiR package[291].

7.3.6 Reporting

Each AI algorithm was assigned a DL Identifier (ID) for the purposes of this study. For additional details please refer to Section 5.3.7 in Chapter 5.

7.4 Results

7.4.1 Data

In total 78,849 cases were included. 24,563 (31.2%) cases were from Cambridge and 54,286 (68.8%) cases were from Norwich. The median age of the cohort was 59.0 years old [IQR 54.0–63.0]. The study cases characteristics are detailed in Table 7-1.

	Cambridge		Norwich
Total Cases n	24563		54286
Year of Screen			
2015	-		21017 (38.7%)
2016	-		19219 (35.4%)
2017	11956 (48.7%)		14050 (25.9%)
2018	12607 (51.3%)		-
FFDM Manufacturer			
GE	235 (1.0%)		54286 (100%)
Philips	24328 (99.0%)		-
Age at Screening			
Median [IQR]	58.0 [54.0-63.0]		59.0 [52.0-64.0]
47-49	76 (0.3%)		4030 (7.4%)
50-59	13871 (56.5%)		26752 (49.3%)
60-69	9607 (39.1%)		20785 (38.3%)
70+	1009 (4.1%)		2719 (5.0%)
Density BI-RADS	Volpara$^{\beta}$	DL-3$^{\alpha}$	DL-3$^{\alpha}$
a	3965 (16.1%)	5682 (23.1%)	8109 (14.9%)
b	11123 (45.3%)	13527 (55.1%)	31260 (57.6%)
c	6607 (26.9%)	5252 (21.4%)	14180 (26.1%)
d	2516 (10.2%)	102 (0.4%)	737 (1.4%)
Missing	48 (0.2%)	0 (0.0%)	0 (0.0%)
Cancers			
SDC Rate per 1000 screens	342 8.5/1000		545 7.0/1000
IC Rate per 1000 screens	167 4.2/1000		272 3.5/1000
NRC Rate per 1000 screens	184* 6.5/1000		504 8.1/1000
FRC	-		149*
Next round Interval cancers	18*		181*

Table 7-1 – Summary of testing dataset characteristics. *Integer values and percentages in brackets (%) and Interquartile range in square brackets [IQR]. BI-RADS: Breast imaging-reporting and data system, FRC: Future round cancer, FFDM: Full Field Digital Mammography, GE: General Electric, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, SDC: Screen detected cancer. $^{\alpha}$DL-3 5th edition BI-RADS density scores on processed full field digital mammograms. $^{\beta}$Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data. *Incomplete follow-up time period information from which to calculate an accurate rate.*

In total 69.0% of the study mammograms were from GE machines, whereas 31.0% were from Philips machines with predominantly Philips machines at Cambridge and GE machines at Norwich. Approximately 1/6th of the cohort of women aged 67-69 are were not included as they would not have self-referred to have a repeat screen, thus would not have met the threshold for the 'normal' ground truth used in this study. This has a potential knock-on effect for the overall density percentages reported, as women aged 67-69 have a higher proportion of BI-RADS category a and b cases. The cohort contains 887 (1.1%) SDCs, 439 (0.6%) ICs, and 688 (0.9%) NRCs. The characteristics of the SDC and NRC cases in the study cohort are shown in Table 7-2.

		Cambridge SDC n (%)		Cambridge NRC n (%)		Norwich SDC n (%)	Norwich NRC n (%)
Total Cases n		342		184		545	504
Total Lesions n		359		191		562	527
Round Length* (days)		1088 [1066-1105]		1221 [1080-1332]		1063 [1036-1085]	1078 [1064-1128]
Round Length* (months)		35.8 [35.0-36.3]		40.1 [35.5-43.8]		35.0 [34.1-35.7]	35.4 [35.0-37.1]
Age at Screening	Median [IQR]	62.0 [57.0-68.0]		60.0 [54.0-65.0]		62.0 [56.0-68.0]	61.0 [55.0-65.0]
	47-49	0 (0.0%)		1 (0.5%)		33 (6.1%)	25 (5.0%)
	50-59	123 (36.0%)		84 (45.7%)		171 (31.3%)	199 (39.5%)
	60-69	164 (47.9%)		86 (46.7%)		259 (47.5%)	237 (47.0%)
	70+	55 (16.1%)		13 (7.1%)		82 (15.0%)	43 (8.5%)
Invasive Status	Invasive	292 (81.3%)		160 (83.8%)		478 (85.1%)	432 (82.0%)
	Non-invasive	66 (18.4%)		30 (15.7%)		82 (14.6%)	92 (17.5%)
	Missing	1 (0.3%)		1 (0.5%)		2 (0.4%)	3 (0.6%)
Invasive Tumour Size[δ]	< 15	154 (52.7%)		72 (45.0%)		261 (54.6%)	244 (56.5%)
	>= 15	128 (43.8%)		60 (37.5%)		198 (41.4%)	154 (35.6%)
	Missing	10 (3.4%)		28 (17.5%)		19 (4.0%)	34 (7.9%)
Invasive Tumour Grade[δ]	1	52 (17.8%)		24 (15.0%)		132 (27.6%)	133 (30.8%)
	2	170 (58.2%)		98 (61.2%)		233 (48.7%)	187 (43.3%)
	3	53 (18.2%)		24 (15.0%)		104 (21.8%)	86 (19.9%)
	Missing	17 (5.8%)		14 (8.8%)		9 (1.9%)	26 (6.0%)
		Volpara[β]	DL-3[α]	Volpara[β]	DL-3[α]	DL-3[α]	DL-3[α]
Density BI-RADS[λ]	a	47 (13.7%)	72 (21.1%)	27 (14.7%)	41 (22.3%)	52 (9.5%)	59 (11.7%)
	b	169 (49.4%)	197 (57.6%)	95 (51.6%)	98 (53.3%)	334 (61.3%)	303 (60.1%)
	c	87 (25.4%)	72 (21.1%)	47 (25.5%)	45 (24.5%)	153 (28.1%)	141 (28.0%)
	d	35 (10.2%)	1 (0.29%)	15 (8.2%)	0 (0.0%)	6 (1.1%)	1 (0.2%)
	Missing	4 (1.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Table 7-2 – Screen detected (SDC) and next round cancer (NRC) characteristics by lesions and cases. *With integer values and percentages in brackets (%). Invasive Tumour Size in millimetres (mm). BI-RADS: Breast imaging-reporting and data system, NRC: Next round cancer, SDC: Screen detected cancer. [δ]Invasive lesions only. [λ]Cases only. [α]DL-3 5th edition BI-RADS density scores on processed full field digital mammograms. [β]Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data. *Cases not screened before or screened more than 6 years ago were not included in this calculation. Results for screen detected cancers between sites are comparable. However the results for next round cancers are not comparable between sites. This is because cases from 2017 onwards were effected by the pause in screening during the Covid-19 pandemic, which is described in Chapter 4. Cambridge data includes 2017-2018 cases (~84.8% cases effected) whereas Norwich includes 2015-2017 cases (~21.2% cases effected), thus this effect is seen in the Cambridge next round cancer results where the round length is increased. It is expected that the size and grade of cancers would increase as a consequence, however due to the increase use of hormone therapy in the pandemic this impact may not been seen in the histopathological size and grade.*

Of the 439 ICs, 167 (38.0%) were from Cambridge and 272 (62.0%) were from Norwich. The characteristics of the IC cases in the study cohort are shown in Table 7-3.

		Cambridge IC n (%)		Norwich IC n (%)
Total Cases n		167		272
Total Lesions n		170		275
Age at Screening	Median [IQR]	59.0 [54.0-66.0]		62.0 [55.0-68.0]
	47-49	1 (0.6%)		18 (6.6%)
	50-59	85 (50.9%)		99 (36.4%)
	60-69	58 (34.7%)		114 (41.9%)
	70+	23 (13.8%)		41 (15.1%)
Invasive Status	Invasive	145 (85.3%)		260 (94.5%)
	Non-invasive	14 (8.2%)		14 (5.1%)
	Missing	11 (6.5%)		1 (0.4%)
Invasive Tumour Size[δ]	< 15	39 (26.9%)		87 (33.5%)
	>= 15	86 (59.3%)		142 (54.6%)
	Missing	20 (13.8%)		31 (11.9%)
Invasive Tumour Grade[δ]	1	24 (16.6%)		31 (11.9%)
	2	62 (42.8%)		127 (48.8%)
	3	57 (39.3%)		95 (36.5%)
	Missing	2 (1.4%)		7 (2.7%)
		Volpara[β]	DL-3[α]	DL-3[α]
Density BI-RADS[λ]	a	12 (7.2%)	15 (9.0%)	10 (3.7%)
	b	65 (38.9%)	85 (50.9%)	131 (48.2%)
	c	58 (34.7%)	63 (37.7%)	122 (44.9%)
	d	31 (18.6%)	4 (2.4%)	9 (3.3%)
	Missing	1 (0.6%)	0 (0.0%)	0 (0.0%)
Interval (months)[λ]	0-12	25 (15.0%)		44 (16.1%)
	12-24	56 (33.5%)		100 (36.8%)
	24-36	86 (51.5%)		128 (47.1%)
	36-40	0 (0.0%)		0 (0.0%)
	Missing	0 (0.0%)		0 (0.0%)
Radiological Audit Classification[λ]	Normal/ Benign	138 (82.6%)		199 (73.2%)
	Uncertain	19 (11.4%)		59 (21.7%)
	Suspicious	0 (0.0%)		9 (3.3%)
	Unclassifiable	1 (0.6%)		5 (1.8%)
	Missing	9 (5.4%)		0 (0.0%)

Table 7-3 – Interval cancer (IC) characteristics by lesions and cases. *With integer values and percentages in brackets (%). BI-RADS: Breast imaging-reporting and data system, IC: Interval cancer. [δ]Invasive lesions only. [λ]Cases only. Invasive Tumour Size in millimetres (mm). [α]DL-3 5th edition BI-RADS density scores on processed full field digital mammograms.[β] Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data.*

The median time to diagnosis was 741.0 [IQR 502.5–963.0] days for all ICs at Cambridge and 703.5 [IQR 450.0–944.2] days at Norwich.

The double reader performance and single first reader performance, when combining outputs from both sites, is shown below in Table 7-4.

	Double reading	First reader
Sensitivity	68.9%	63.6%
Specificity	97.6%	97.0%
Precision	32.8%	26.4%
Arbitration	2.5%	-
Recall rate	3.5%	4.1%
n detected (%)		
SDC	887 (100%)	807 (91.0%)
IC	27 (6.2%)	36 (8.2%)
NRC	20 (2.9%)	29 (4.2%)
FRC	4 (2.7%)	7 (4.7%)
NRIC	7 (3.5%)	10 (5.0%)
FP	1840	2310

Table 7-4 – Double and single first reader performance at both Cambridge and Norwich. *FRC: Future round cancer, FP: False positive, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, SDC: Screen detected cancer.*

7.4.2 Rule-out triage – Threshold 1 and 2

The overall AUROC for DL-1, DL-2, DL-3 was 0.962 [95% CI 0.955–0.969], 0.966 [95% CI 0.961–0.972] and 0.975 [95% CI 0.970–0.980] respectively, when classifying cancers as SDCs only, Figure 7-4. The AUROC of DL-3 was statistically significantly different ($p < 0.05$) to that of DL-1 and DL-2 when tested on all and Norwich data. As well as DL-3 AUROC performance was statistically significantly different ($p < 0.05$) when comparing between Norwich and Cambridge site data.

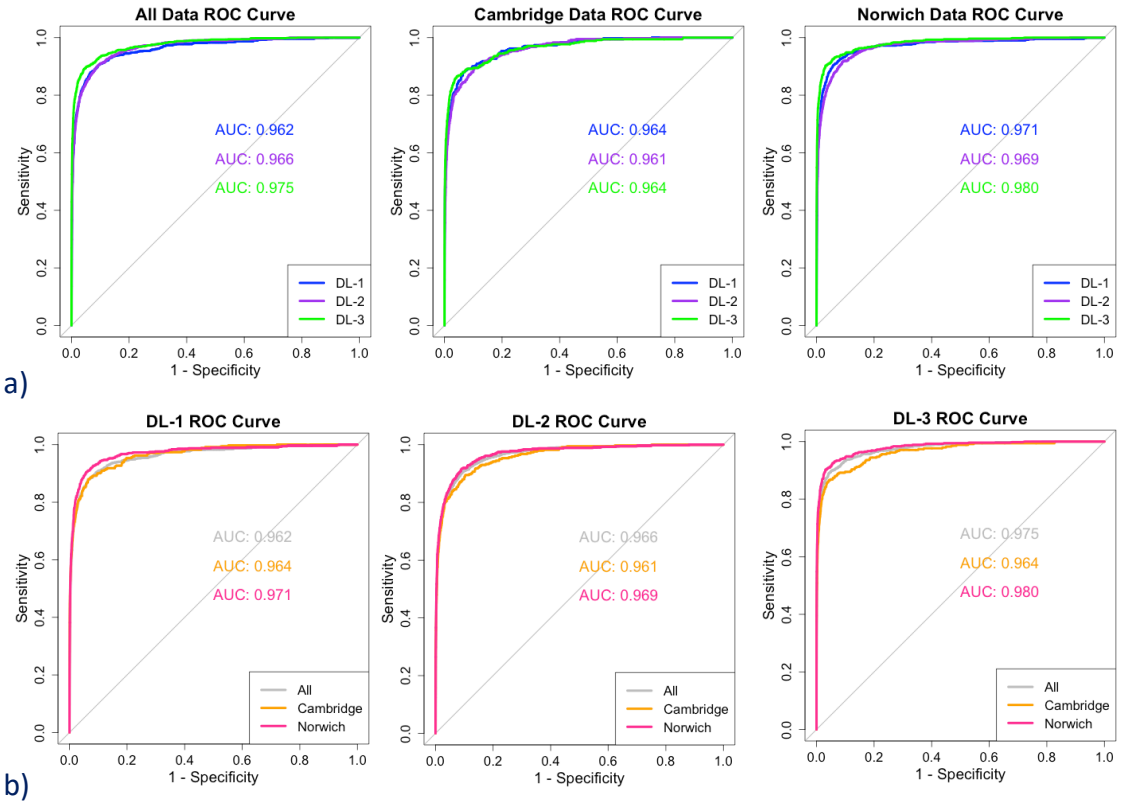


Figure 7-4 – Receiver operating characteristic (ROC) curves for screen detected cancers (SDCs) as cases. *a) results per site, b) results per algorithm. The overall results are in grey, Cambridge in orange and Norwich in pink. The results for DL-1 are in blue, DL-2 in purple, and DL-3 in green. Area under the receiver operating characteristic curve values are provided for each site and each algorithms performance.*

Firstly applying the Scenario B workflow at the 99.0% sensitivity threshold (threshold 1), where SDCs were classed as cases, resulted in 65.0%, 46.8% and 44.4% cases left to be read by a double reading workflow for DL-1, DL-2 and DL-3 respectively, Table 7-5. At this threshold all algorithms ruled out 9 (1.0%) SDCs, and between 100-222 (14.5-32.3%) NRCs and 74-114 (16.9-26.0%) ICs. DL-3 ruled out the highest number of false positive (FP) cases (n = 465), whereas DL-1 and DL-2 ruled out a similar volume (DL-1 n = 318 and DL-2 n = 369).

	Sensitivity Threshold	Specificity	Missed Cancers			FP out	% to read
			SDC	IC	NRC		
Cases	-	-	887	439	688	1840	-
DL-1 1)	99.0%	35.3% [30.0-57.0]	9 (1.0%)	74 (16.9%)	100 (14.5%)	318 (14.4%)	65.0%
DL-1 2)	99.9%	10.8% [10.6-28.0]	1 (0.1%)	14 (3.2%)	18 (2.6%)	68 (3.7%)	89.4%
DL-2 1)	99.0%	53.8% [35.9-66.4]	9 (1.0%)	107 (24.4%)	214 (31.1%)	369 (20.1%)	46.8%
DL-2 2)	99.9%	12.1% [11.9-29.2]	1 (0.1%)	12 (2.7%)	28 (4.1%)	25 (1.4%)	88.0%
DL-3 1)	99.0%	56.3% [38.1-66.7]	9 (1.0%)	114 (26.0%)	222 (32.3%)	465 (25.3%)	44.4%
DL-3 2)	99.9%	21.9% [21.4-38.1]	1 (0.1%)	21 (4.8%)	55 (8.0%)	131 (7.1%)	78.3%

Table 7-5 – Results at 1) 99.0% sensitivity threshold 1 and 2) 99.9% sensitivity threshold 2. *Missed cases are shown for screen detected cancers, next round cancers and interval cancers as well as the proportion of false positives ruled out. Where screen detected cancers were classed as cases only for the threshold identification. FP: False positive, IC: Interval cancer, NRC: Next round cancer, SDC: Screen detected cancer.*

The sensitivity (Δ -0.7%) and specificity (Δ +0.4%~+0.6%) is non-inferior. A lower arbitration rate (Δ -0.4%~-0.7%) and recall rate (Δ -0.4%~-0.6%) was also observed at threshold 1, Table 7-6.

Applying Scenario B workflow at the 99.9% sensitivity threshold (threshold 2) left in 89.4%, 88.0% and 78.3% cases required to be read by the double reading workflow for DL-1, DL-2 and DL-3 respectively. At this threshold, 1 (0.1%) SDC was missed, and between 18-55 (2.6-8.0%) NRCs and 12-21 (2.7-4.8%) ICs were missed from rule out triage, Table 7-5. The specificity and sensitivity were both non-inferior. The arbitration rate (Δ 0.0%~-0.2%) and recall rate (Δ 0.0%~-0.1%) were observed to not change at this threshold compared to double reading, Table 7-7.

	DL-1 + readers	DL-2 + readers	DL-3 + readers
Sensitivity threshold	99.0%	99.0%	99.0%
Sensitivity	68.2% [65.6-70.7] p < 0.01	68.2% [65.6-70.7] p < 0.01	68.2% [65.6-70.7] p < 0.01
Specificity	98.0% [97.9-98.1] p < 0.01	98.1% [98.0-98.2] p < 0.01	98.2% [98.1-98.3] p < 0.01
Precision	36.8%	37.6%	39.2%
Arbitration	2.1%	1.9%	1.8%
Recall rate	3.1%	3.1%	2.9%
n (%) Missed			
SDC	9 (1.0%)	9 (1.0%)	9 (1.0%)
IC	74 (16.9%)	107 (24.4%)	114 (26.0%)
NRC	100 (14.5%)	214 (31.1%)	222 (32.3%)
FRC	18 (12.1%)	58 (38.9%)	74 (49.7%)
NRIC	32 (16.1%)	82 (41.2%)	87 (43.7%)
Rule out			
Normal cases n (%)	27332 (34.7%)	41467 (52.6%)	43363 (55.0%)
Reader FP n (%)	318 (14.4%)	369 (20.1%)	465 (25.3%)

Table 7-6 – Results for DL-1, DL-2 and DL-3 at the 99.0% sensitivity (threshold 1) Scenario B. *FP: False positive, FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, SDC: Screen detected cancer. p values are calculated using a one-sided z-test.*

	DL-1 + readers	DL-2 + readers	DL-3 + readers
Sensitivity threshold	99.9%	99.9%	99.9%
Sensitivity	68.9% [66.3-71.3] p < 0.01	68.9% [66.3-71.3] p < 0.01	68.9% [66.3-71.3] p < 0.01
Specificity	97.7% [97.6-97.8] p < 0.01	97.6% [97.5-97.7] p < 0.01	97.8% [97.6-97.9] p < 0.01
Precision	33.6%	33.1%	34.4%
Arbitration	2.4%	2.5%	2.3%
Recall rate	3.4%	3.5%	3.4%
n (%) Missed			
SDC	1 (0.1%)	1 (0.1%)	1 (0.1%)
IC	14 (3.2%)	12 (2.7%)	21 (4.8%)
NRC	18 (2.6%)	28 (4.1%)	55 (8.0%)
FRC	2 (0.6%)	7 (4.7%)	28 (18.8%)
NRIC	7 (3.5%)	14 (7.0%)	21 (10.6%)
Rule out			
Normal cases n (%)	8342 (10.6%)	9371 (11.9%)	17017 (21.6%)
Reader FP n (%)	68 (3.7%)	25 (1.4%)	131 (7.1%)

Table 7-7 – Results for DL-1, DL-2 and DL-3 at the 99.9% sensitivity (threshold 2) Scenario B. *FP: False positive, FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, SDC: Screen detected cancer. p values are calculated using a one-sided z-test.*

Applying Scenario C at the 99.0% sensitivity threshold (threshold 1), resulted in 35.0-55.6% of cases classified as to be read by one reader only. Whilst maintaining SDC detection (99.9-100%) as well as an observed lower the arbitration rate (Δ -0.4%~-0.7%). However, the recall rate was observed to be higher (Δ +0.2%~+0.3%), Table 7-8. Specificity and sensitivity performance was non-inferior ($p < 0.01$).

	DL-1 + readers	DL-2 + readers	DL-3 + readers
Sensitivity	68.9% [66.3-71.3] p < 0.01	69.0% [66.4-71.5] p < 0.01	68.9% [66.4-71.4] p < 0.01
Specificity	97.4% [97.3-97.5] p < 0.01	97.4% [97.2-97.5] p < 0.01	97.4% [97.3-97.5] p < 0.01
Precision	31.5%	30.8%	30.9%
Arbitration	2.1%	1.9%	1.8%
Recall rate	3.7%	3.8%	3.7%
n (%) Detected			
SDC	886 (99.9%)	886 (99.9%)	887 (100%)
IC	27 (6.2%)	29 (6.6%)	27 (6.2%)
NRC	20 (2.9%)	21 (3.1%)	22 (3.2%)
FRC	4 (2.7%)	4 (2.7%)	5 (3.4%)
NRIC	8 (4.0%)	9 (4.5%)	7 (3.5%)
Rule out			
Single reading (%)	27565 (35.0%)	41937 (53.2%)	43869 (55.6%)
False positives n	1956	2019	2008

Table 7-8 – Results for DL-1, DL-2 and DL-3 at the 99.0% sensitivity (threshold 1) Scenario C. *FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, SDC: Screen detected cancer. p values are calculated using a one-sided z-test.*

7.4.3 Rule-out triage – Threshold 3 and 4

The AUROC for DL-1, DL-2, DL-3 was 0.813 [95% CI 0.802–0.824], 0.814 [95% CI 0.803–0.825], and 0.821 [95% CI 0.886–0.906] respectively, when classifying cancers as SDCs, ICs and NRCs, Figure 7-5. There was a statistically significant difference ($p < 0.05$) between the AUROC of DL-3 compared to DL-1 and DL-2 when tested on Norwich data. Applying Scenario B at the 85.0% sensitivity threshold (threshold 3), where SDCs, NRCs and ICs were classed as cases, the percentage of cases requiring double reading after applying the AI algorithm threshold was 49.8% for DL-1, 49.5% for DL-2, and 48.4% for DL-3. Applying Scenario B at the 70.0% specificity threshold (threshold 4), where SDCs, NRCs and ICs were classed as cases, the percentage of cases requiring double reading after applying the AI algorithm threshold was 31.2% for DL-1, 31.1% for DL-2, and 31.2% for DL-3. The results of this analysis for each AI algorithm are shown in Table 7-9.

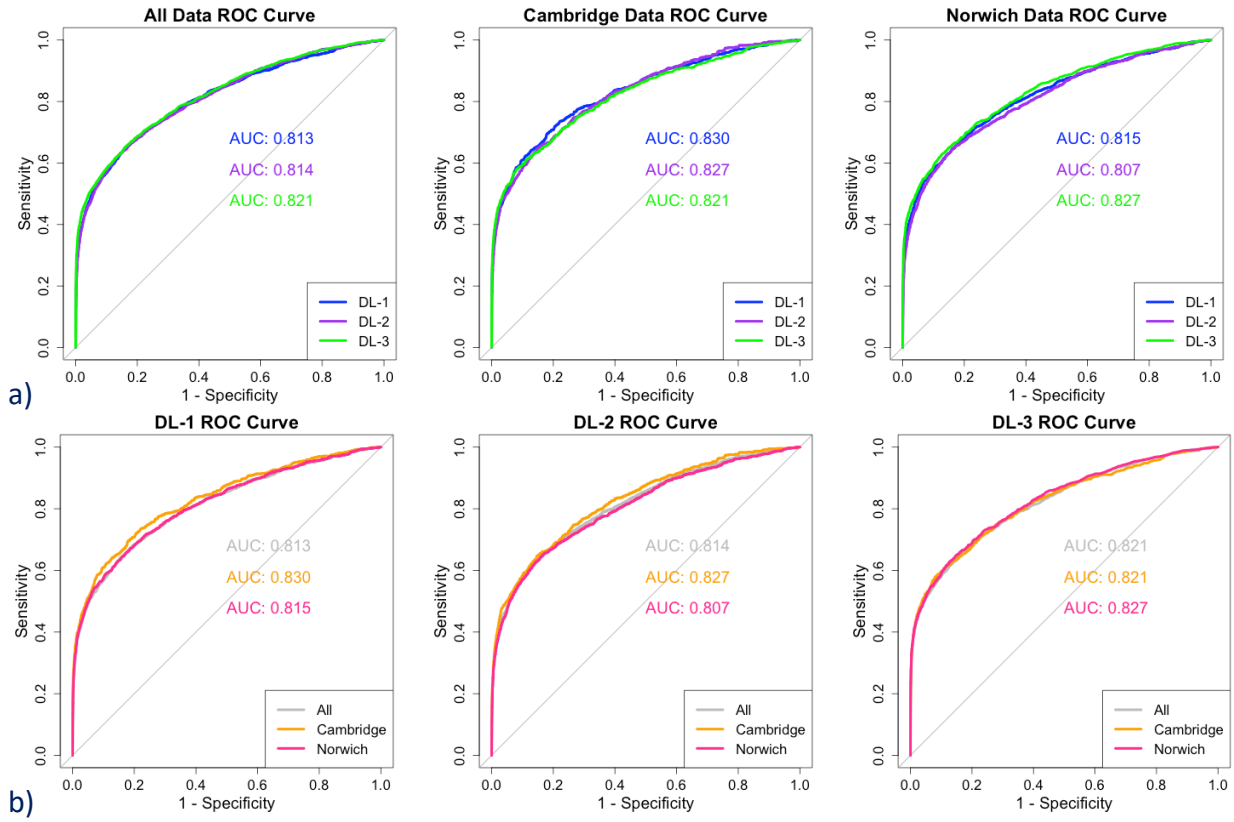


Figure 7-5 – Receiver operating characteristic (ROC) curves for screen detected cancers (SDCs), next round cancers (NRCs) and interval cancers (ICs) as cases. *a) for each site, b) for each algorithm. The overall results are in grey, Cambridge in orange and Norwich in pink. The results for DL-1 are in blue, DL-2 in purple, and DL-3 in green. Area under the receiver operating characteristic curve values are provided for each site and each algorithms performance.*

	Sensitivity Threshold	Specificity	Missed Cancers			FP out	% to read
			SDC	IC	NRC		
Cases	-	-	887	439	688	1840	-
DL-1 1)	85.0%	51.3% [47.5-56.9]	15 (1.7%)	118 (26.9%)	169 (24.6%)	507 (27.6%)	49.8%
DL-2 1)	85.0%	51.6% [48.4-54.7]	8 (0.9%)	95 (21.6%)	199 (28.9%)	338 (18.4%)	49.5%
DL-3 1)	85.0%	52.6% [50.0-55.6]	7 (0.8%)	97 (22.1%)	198 (28.8%)	420 (22.8%)	48.4%
	Specificity Threshold						
DL-1 2)	70.0%	75.3% [73.4-77.3]	34 (3.8%)	182 (41.5%)	281 (40.8%)	827 (45.0%)	31.2%
DL-2 2)	70.0%	74.8% [72.9-76.7]	20 (2.3%)	172 (39.2%)	316 (45.9%)	661 (35.9%)	31.1%
DL-3 2)	70.0%	75.8% [73.8-77.7]	19 (2.1%)	156 (35.5%)	313 (45.5%)	696 (37.8%)	31.2%

Table 7-9 – Results at 1) 85.0% sensitivity (threshold 3) and 2) results at 70.0% specificity (threshold 4). *Missed cases are shown for screen detected cancers, next round cancers and interval cancers as well as the proportion of false positives ruled out. Where screen detected cancers, next round cancers and interval cancers were classed as cases for the threshold identification. FP: False positive, IC: Interval cancer, NRC: Next round cancer, SDC: Screen detected cancer.*

	DL-1 + readers	DL-2 + readers	DL-3 + readers
Sensitivity	69.0% [66.4-71.5] p < 0.01	69.0% [66.4-71.5] p < 0.01	68.9% [66.4-71.4] p < 0.01
Specificity	97.4% [97.2-97.5] p < 0.01	97.4% [97.3-97.5] p < 0.01	97.4% [97.3-97.5] p < 0.01
Precision	30.8%	31.0%	31.1%
Arbitration	1.8%	2.0%	1.9%
Recall rate	3.8%	3.8%	3.7%
n (%) Detected			
SDC	886 (99.9%)	886 (99.9%)	887 (100%)
IC	29 (6.6%)	29 (6.6%)	27 (6.2%)
NRC	19 (2.8%)	21 (3.1%)	21 (3.1%)
FRC	4 (2.7%)	4 (2.7%)	5 (3.4%)
NRIC	9 (4.5%)	9 (4.5%)	7 (3.5%)
Rule out			
Single reading (%)	39625 (50.3%)	39923 (50.6%)	40719 (51.6%)
False positive n	2023	2005	1985

Table 7-10 – Results for DL-1, DL-2 and DL-3 at the 85.0% sensitivity (threshold 3) Scenario C. *FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, SDC: Screen detected cancer. p values are calculated using a one-sided z-test.*

	DL-1 + readers	DL-2 + readers	DL-3 + readers
Sensitivity	68.8% [66.2-71.3] p < 0.01	68.8% [66.2-71.3] p < 0.01	68.8% [66.2-71.3] p < 0.01
Specificity	97.2% [97.1-97.4] p < 0.01	97.2% [97.1-97.4] p < 0.01	97.3% [97.2-97.4] p < 0.01
Precision	29.9%	29.9%	30.3%
Arbitration	1.3%	1.5%	1.5%
Recall rate	3.9%	3.9%	3.8%
n (%) Detected			
SDC	883 (99.5%)	882 (99.4%)	885 (99.8%)
IC	29 (6.6%)	30 (6.8%)	27 (6.2%)
NRC	18 (2.6%)	20 (2.9%)	21 (3.1%)
FRC	4 (2.7%)	5 (3.6%)	6 (4.0%)
NRIC	9 (4.5%)	9 (4.5%)	7 (3.5%)
Rule out			
Single reading (%)	54281 (68.8%)	54292 (68.9%)	54266 (68.8%)
False positive n	2110	2102	2063

Table 7-11 – Results for DL-1, DL-2 and DL-3 at the 70.0% specificity (threshold 4) Scenario C. *FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, SDC: Screen detected cancer. p values are calculated using a one-sided z-test.*

Implementing the alternative Scenario C and threshold 3, specificity and sensitivity performance was found to be non-inferior ($p < 0.01$), Table 7-10. A lower arbitration rate (Δ -0.5%~-0.7%) was observed, whilst the recall rate (Δ +0.2%~+0.3%) was higher. Implementing the alternative Scenario C and threshold 4, specificity and sensitivity performance was found to be non-inferior ($p < 0.01$), Table 7-11. The arbitration rate (Δ -1.0%~-1.2%) was again lower, whilst the recall rate (Δ +0.3%~+0.4%) was higher.

The density and violin plots in Figure 7-6 show the distribution of cases and the assigned thresholds 1, 2, 3 and 4 for each AI algorithm.

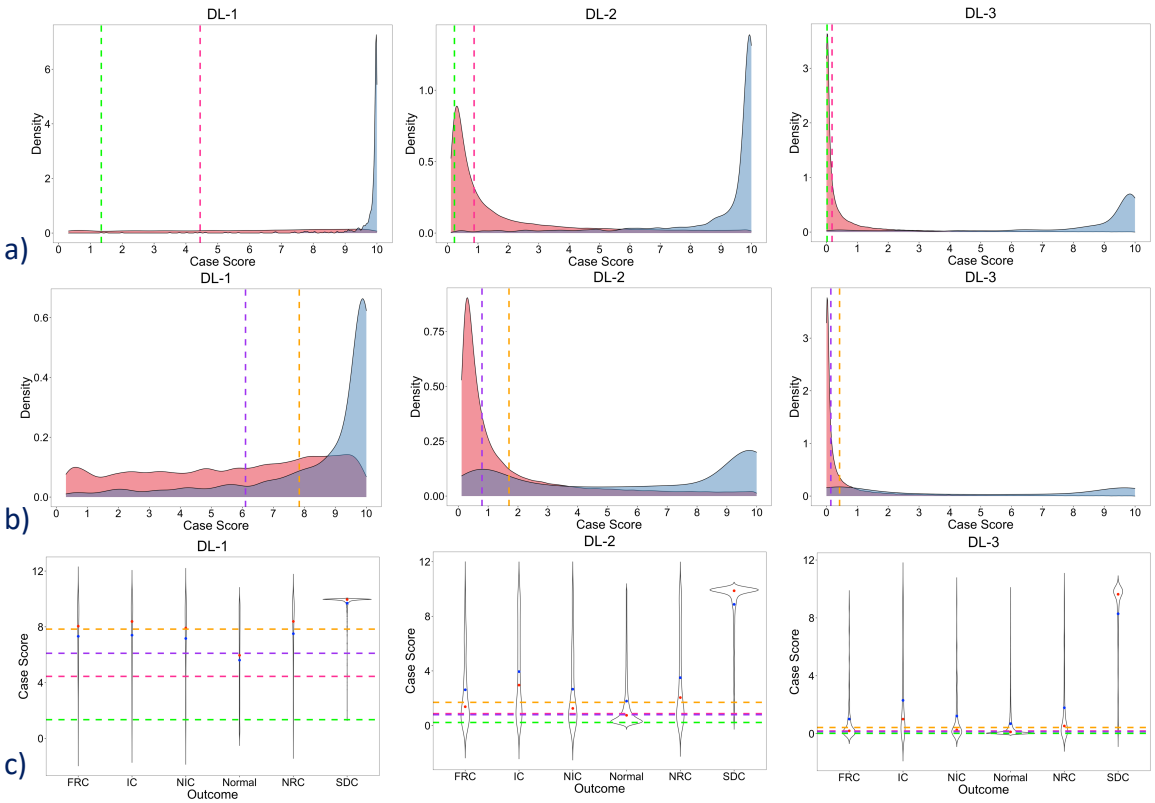


Figure 7-6 – Plots for rule out triage thresholds. *a) Density plot for screen detected cancers as cases, b) density plot for screen detected, next round and interval cancers as cases where the cancers are in blue and normal cases in red, c) violin plot for all cancer case types, where the blue dot in the violin plot is the mean score and the red is the median score. The pink line represents threshold 1 (99.0% sensitivity), the green line represents threshold 2 (99.9% sensitivity), the purple line represents threshold 3 (85.0% sensitivity) and the orange line represents threshold 4 (70.0% specificity). FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NIC: Next round interval cancer, SDC: Screen detected cancer.*

7.4.4 Rule-in triage

Scenario D applied 94.0-99.0% specificity cut-off to determine the percentage of ICs and NRCs with a high suspicion that should be referred for additional supplemental imaging / assessment, Table 7-12. At the lower 94.0% specificity threshold 101-115 (23.0-26.2%) ICs and 142-157 (20.6-22.8%) NRCs were ruled in for further assessment. At the 99.0% specificity threshold 26-46 (5.9-10.5%) ICs and 40-44 (5.8-6.4%) NRCs were recalled for further assessment.

SDC, IC, NRC as cases ruled in									
	DL-1			DL-2			DL-3		
Spec	%TRR	IC	NRC	%TRR	IC	NRC	%TRR	IC	NRC
99.0%	4.5%	40 (9.1%)	44 (6.4%)	4.5%	26 (5.9%)	40 (5.8%)	4.4%	46 (10.5%)	41 (6.0%)
98.0%	5.5%	56 (12.8%)	69 (10.0%)	5.4%	48 (10.9%)	82 (11.9%)	5.4%	62 (14.1%)	77 (11.2%)
97.0%	6.4%	68 (15.5%)	94 (13.7%)	6.4%	68 (15.5%)	96 (14.0%)	6.3%	74 (16.9%)	96 (14.0%)
96.0%	7.3%	82 (18.7%)	116 (16.9%)	7.3%	75 (17.1%)	115 (16.7%)	7.2%	88 (20.0%)	114 (16.6%)
95.0%	8.3%	94 (21.4%)	137 (19.9%)	8.2%	87 (19.8%)	129 (18.8%)	8.2%	101 (23.0%)	130 (18.9%)
94.0%	9.2%	101 (23.0%)	157 (22.8%)	9.2%	103 (23.5%)	149 (21.7%)	9.1%	115 (26.2%)	142 (20.6%)

Table 7-12 – Scenario D perturbations of specificity with screen detected cancers (SDCs), interval cancers (ICs) and next round cancers (NRCs) as cases. *IC: Interval cancer, NRC: Next round cancer, Spec: Specificity, %TRR: Percentage total recall rate.*

Table 7-13 shows the proportion of FRCs and NRICs that could have been detected at these thresholds (94.0-99.0% specificity) further increasing the proportion of cancers which could potentially be detected earlier. In addition, the number of additional false positive recalls, which would ultimately lead to an increase in the recall rate from this scenario is included in Table 7-13, and increases as the specificity threshold is reduced.

FP, FRC, NRIC ruled in									
	DL-1			DL-2			DL-3		
Spec	FP	FRC	NRIC	FP	FRC	NRIC	FP	FRC	NRIC
99%	753	9 (6.0%)	9 (4.5%)	757	3 (2.0%)	8 (4.0%)	760	4 (2.7%)	5 (2.5%)
98%	1508	12 (8.1%)	15 (7.5%)	1521	5 (3.4%)	11 (5.5%)	1519	8 (5.4%)	10 (5.0%)
97%	2269	14 (9.4%)	21 (10.6%)	2284	7 (4.7%)	14 (7.0%)	2281	10 (6.7%)	14 (7.0%)
96%	3026	18 (12.1%)	26 (13.1%)	3051	8 (5.4%)	14 (7.0%)	3048	10 (6.7%)	16 (8.0%)
95%	3787	24 (16.1%)	30 (15.1%)	3815	12 (8.1%)	15 (7.5%)	3805	12 (8.1%)	25 (12.6%)
94%	4550	28 (18.8%)	32 (16.1%)	4573	18 (12.1%)	19 (9.6%)	4568	15 (10.1%)	28 (14.1%)

Table 7-13 – Scenario D perturbations of specificity with screen detected cancers (SDCs), interval cancers (ICs) and next round cancers (NRCs) as cases – additional cancers detected. *FP: False positive, FRC: Future round cancer, NRIC: Next round interval cancer, Spec: Specificity.*

The results of applying the 94.0-99.0% specificity thresholds for Scenario E where cases are auto recalled if above the threshold and were not recalled by human readers, are shown in Table 7-14.

SDC, IC, NRC as cases ruled in									
	DL-1			DL-2			DL-3		
Spec	%ARR	IC	NRC	%ARR	IC	NRC	%ARR	IC	NRC
99.0%	1.0%	32 (7.3%)	42 (6.1%)	0.9%	20 (4.6%)	36 (5.2%)	0.9%	36 (8.2%)	37 (5.4%)
98.0%	1.9%	47 (10.7%)	67 (9.7%)	1.9%	38 (8.7%)	75 (10.9%)	1.8%	49 (11.2%)	71 (10.3%)
97.0%	2.9%	59 (13.4%)	92 (13.4%)	2.8%	58 (13.2%)	88 (12.8%)	2.8%	60 (13.7%)	90 (13.1%)
96.0%	3.8%	71 (16.2%)	111 (16.1%)	3.8%	65 (14.8%)	107 (15.6%)	3.7%	73 (16.6%)	106 (16.6%)
95.0%	4.8%	82 (18.7%)	132 (19.2%)	4.7%	76 (17.3%)	120 (17.4%)	4.6%	85 (19.4%)	121 (17.6%)
94.0%	5.7%	88 (20.1%)	152 (22.1%)	5.7%	90 (20.5%)	139 (20.2%)	5.6%	99 (22.6%)	132 (19.2%)

Table 7-14 – Scenario E perturbations of specificity with screen detected cancers (SDCs), interval cancers (ICs) and next round cancers (NRCs) as cases. *%ARR: Additional Recall rate, IC: Interval cancer, NRC: Next round cancer, Spec: Specificity.*

Applying this scenario at the 94.0% threshold results in an increase in the number of ICs (20.1-22.6%) and NRCs (19.2-22.1%) detected. A lower number of cases are overall detected at the higher specificity threshold of 99.0% (ICs (4.6-8.2%), NRCs (5.2-6.1%)). However the recall rate also increased at all thresholds. This is potentially further offset by increase in FRCs and NRICs detected shown in Table 7-15.

FP, FRC, NRIC ruled in									
	DL-1			DL-2			DL-3		
Spec	FP	FRC	NRIC	FP	FRC	NRIC	FP	FRC	NRIC
99%	675	9 (6.0%)	8 (4.0%)	668	3 (2.0%)	6 (3.0%)	630	4 (2.7%)	4 (2.0%)
98%	1369	12 (8.1%)	13 (6.5%)	1358	5 (3.4%)	9 (4.5%)	1299	8 (5.4%)	9 (4.5%)
97%	2067	14 (9.4%)	18 (9.1%)	2061	7 (4.7%)	12 (6.0%)	1992	10 (6.7%)	13 (6.5%)
96%	2774	17 (11.4%)	23 (11.6%)	2772	8 (5.4%)	12 (6.0%)	2701	10 (6.7%)	15 (7.5%)
95%	3487	23 (15.4%)	27 (13.6%)	3484	12 (8.1%)	13 (6.5%)	3405	12 (8.1%)	23 (11.6%)
94%	4204	27 (18.1%)	29 (14.6%)	4196	17 (11.4%)	16 (8.0%)	4130	15 (10.1%)	24 (12.1%)

Table 7-15 – Scenario E perturbations of specificity with screen detected cancers (SDCs), interval cancers (ICs) and next round cancers (NRCs) as cases – additional cancers detected. *FP: False positive, FRC: Future round cancer, NRIC: Next round interval cancer, Spec: Specificity.*

The density and violin plots in Figure 7-7 show the distribution of cases and the assigned 94.0% and 99.0% specificity threshold cut-off for each algorithm rule in triage approaches.

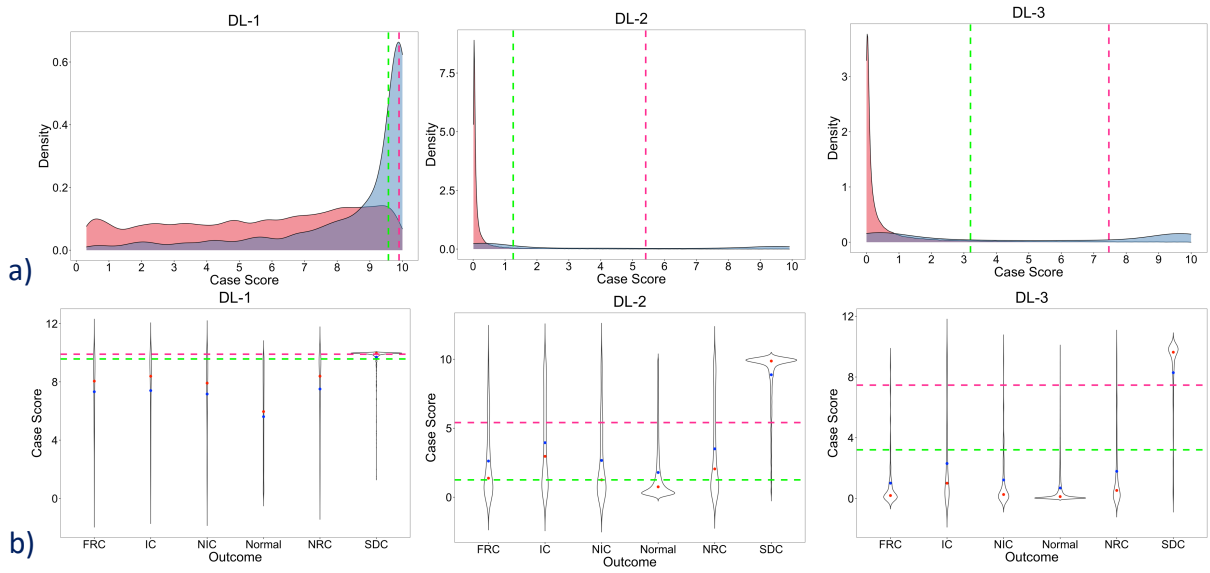


Figure 7-7 – Plots for rule in triage thresholds – Screen detected cancers (SDCs), next round cancers (NRCs) and interval cancers (ICs). *a) Density plot for screen detected, next round and interval cancers as cases in blue and normal cases in red, b) violin plot for all cancer case types, where the blue dot in the violin plot is the mean score and the red is the median score. The green line represents 94.0% specificity, the pink line represents 99.0% specificity. FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NIC: Next round interval cancer, SDC: Screen detected cancer.*

7.4.5 Combined approach

The Combined approach entailed combining Scenario C, at the 99.0% sensitivity (threshold 1), and Scenario E at the 99.0% specificity threshold, as shown in Figure 7-8.

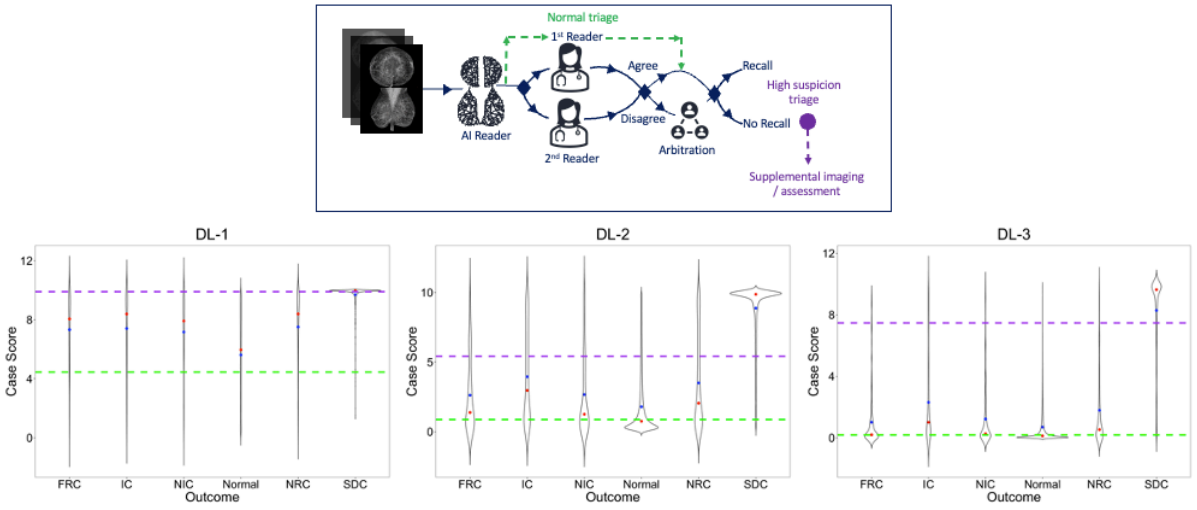


Figure 7-8 – Violin plots for the combined approach of Scenario C and E for both rule in and rule out triage by an artificial intelligence (AI) algorithm. *The normal triage threshold was set a threshold 1 99.0% sensitivity, as shown by the green line on the violin plot. The high suspicious rule in threshold was set at 99.0% specificity, as shown by the purple line on the violin plot. Where the blue dot in the violin plot is the mean score and the red is the median score. FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NIC: Next round interval cancer, SDC: Screen detected cancer.*

	DL-1 + readers	DL-2 + readers	DL-3 + readers
Sensitivity	71.3% [68.7-73.7] p = 0.03516*	70.5% [68.0-73.0] p > 0.05*	71.6% [69.1-74.1] p = 0.01758*
Specificity	96.5% [96.4-96.6] p < 0.01	96.4% [96.3-96.6] p < 0.01	96.5% [96.4-96.6] p < 0.01
Precision	25.8%	25.3%	25.9%
Arbitration	2.1%	1.9%	1.8%
Recall rate	4.7%	4.7%	4.7%
n (%) Detected			
SDC	886 (99.9%)	886 (99.9%)	887 (100%)
IC	59 (13.4%)	49 (11.2%)	63 (14.4%)
NRC	62 (9.0%)	57 (8.3%)	59 (8.6%)
FRC	13 (8.7%)	7 (4.7%)	9 (6.0%)
NRIC	16 (8.0%)	15 (7.5%)	11 (5.5%)
Rule out			
Single reading (%)	27565 (35.0%)	41937 (53.2%)	43869 (55.6%)
Rule in			
% Additional RR	0.97%	0.93%	0.90%

Table 7-16 – Combined approach of Scenario C and E for both rule in and rule out triage by an artificial intelligence (AI) algorithm. *The normal triage threshold was set at 99.0% sensitivity (threshold 1). The high suspicious rule in threshold was set at 99.0% specificity. FRC: Future round cancer, IC: Interval cancer, NRC: Next round cancer, NRIC: Next round interval cancer, RR: Recall rate, SDC: Screen detected cancer. p values are calculated using a one-sided z-test. *Tested for superiority.*

Using this approach the sensitivity was superior ($p < 0.05$) for DL-1 and DL-3. However, this would result trade off in specificity. Overall the proportion of ICs (Δ +5.0%~+8.2%) and NRCs (Δ +5.4%~6.1%) detected was higher, Table 7-16. The recall rate (Δ +1.2%) was observed to be higher and arbitration rate was lower (Δ -0.4%~-0.7%), Table 7-16.

Two settings were applied to calculate the pAUROC for DL-1, DL-2, DL-3, first at the 94.0-99.0% specificity as shown in blue in Figure 7-9, and then at either the 99.0-100% sensitivity (Figure 7-9.a) or 85.0-100% sensitivity (Figure 7-9.b). Overall the AI algorithms achieved a good pAUROC at all thresholds, with DL-3 achieving the highest pAUROC at all settings, Table 7-17.

pAUROC – SDC	DL-1	DL-2	DL-3
94.0-99.0% Specificity	89.6% [88.3-90.8]	89.2% [87.9-90.4]	93.0% [91.9-94.0]
99.0-100% Sensitivity	62.7% [58.1-68.1]	66.8% [61.2-75.3]	70.1% [65.1-77.4]
pAUROC – SDC / NRC / IC	**DL-1**	**DL-2**	**DL-3**
94.0-99.0% Specificity	71.2% [70.1-72.2]	70.7% [69.6-71.8]	72.7% [71.6-73.8]
85.0-100% Sensitivity	62.1% [60.7-63.8]	63.1% [61.7-64.6]	63.6% [62.2-65.0]

Table 7-17 – Partial area under the receiver operator characteristic (pAUROC) curve results. *95.0% CI in square brackets. IC: Interval cancer, NRC: Next round cancer, pAUROC: Partial area under the receiver operator characteristic curve, SDC: Screen detected cancer.*

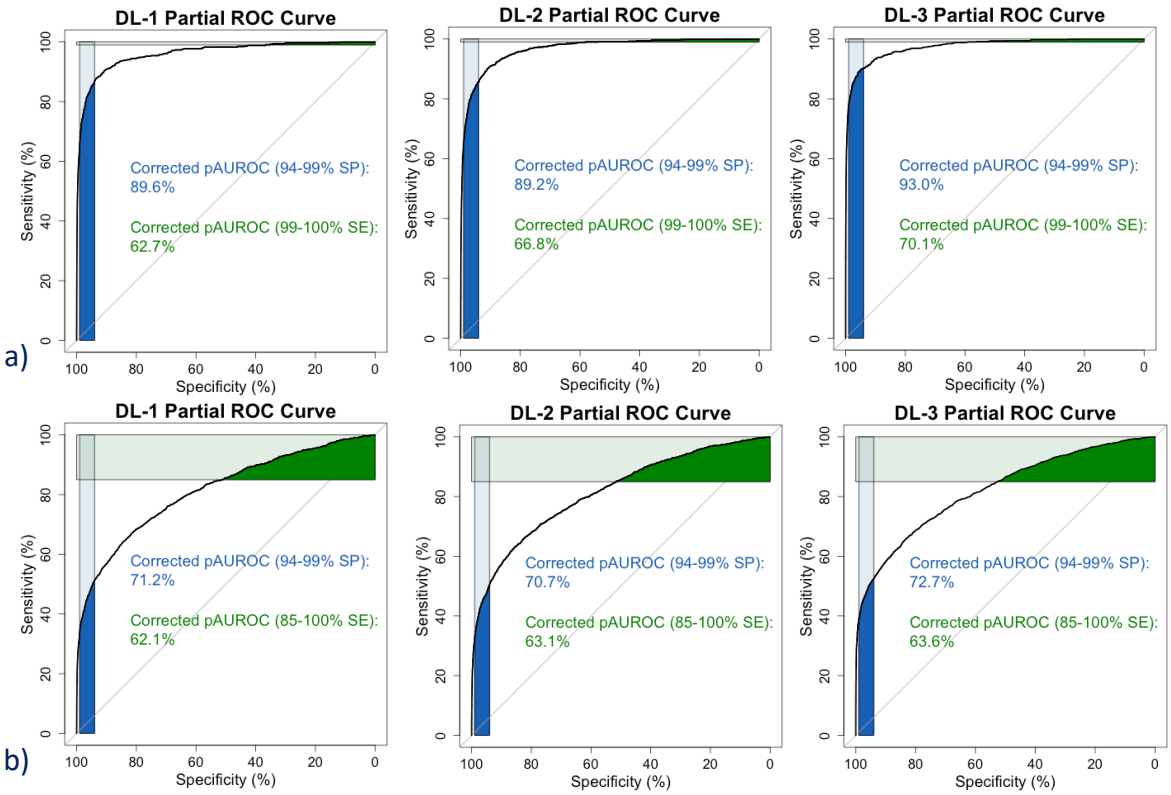


Figure 7-9 – Partial receiver characteristic (pROC) curves. *a) Screen detected cancers as cases applying a 99.0-100% sensitivity to reflect rule out threshold 1 and 2 in the study as shown in green, and a 94.0-99.0% specificity to reflect the rule in thresholds used in this study as shown in blue. b) Screen detected cancers, next round cancers and interval cancers as cases applying 85.0-100% sensitivity to reflect rule out threshold 1, 2 and 3 in the study as shown in green, and a 94.0-99.0% specificity to reflect the rule in thresholds used in this study as shown in blue.*

7.4.6 Sub-group analysis

Performance of the AI algorithms was further assessed at threshold 1 (99.0% sensitivity) for the SDCs that were missed by each AI algorithm at this threshold [n = 9 (1.0%)] using Scenario B, Table 7-18. There was no statistically significant difference in the types of cancers missed relative to the true distribution of cancer cases using Chi squared χ^2 test ($p > 0.05$).

	n	DL-1	DL-2	DL-3
Total Cases n	887	9	9	9
Total Lesions n	921	9	9	9
Total invasive lesions	770	8	5	7
Age at Screening				
< 60	327	3 (0.9%)	4 (1.2%)	4 (1.2%)
>= 60	560	6 (1.1%)	5 (0.9%)	5 (0.9%)
p value	-	0.82696	0.639289	0.639289
FFDM Vendor				
GE	554	5 (0.9%)	7 (1.3%)	7 (1.3%)
Philips	333	4 (1.2%)	2 (0.6%)	2 (0.6%)
p value	-	0.670613	0.34459	0.34459
Invasive Tumour Size$^{\delta}$				
< 15 mm	415	3 (0.7%)	1 (0.2%)	5 (1.2%)
>= 15 mm	326	5 (1.5%)	4 (1.2%)	2 (0.6%)
Missing	29	0 (0.0%)	0 (0.0%)	0 (0.0%)
p value	-	0.294494	0.106186	0.413069
Invasive Tumour Grade $^{\delta}$				
1	184	0 (0.0%)	0 (0.0%)	0 (0.0%)
2	403	7 (1.7%)	4 (1.0%)	5 (1.2%)
3	157	1 (0.6%)	1 (0.6%)	2 (1.3%)
Missing	8	0 (0.0%)	0 (0.0%)	0 (0.0%)
p value	-	-	-	-
Density BI-RADS$^{\alpha}$				
a	124	2 (1.6%)	0 (0.0%)	0 (0.0%)
b	531	4 (0.8%)	7 (1.3%)	5 (0.9%)
c	225	3 (1.3%)	2 (0.9%)	4 (1.8%)
d	7	0 (0.0%)	0 (0.0%)	0 (0.0%)
p value	-	-	-	-
Density BI-RADS $^{\beta}$				
a	47	0 (0.0%)	0 (0.0%)	0 (0.0%)
b	169	2 (1.2%)	1 (0.6%)	1 (0.6%)
c	87	0 (0.0%)	0 (0.0%)	0 (0.0%)
d	35	2 (5.7%)	0 (0.0%)	1 (2.9%)
Missing	549	5 (0.9%)	8 (1.5%)	7 (1.3%)
p value	-	-	-	-

Table 7-18 – Sub group analysis of DL-1, DL-2, DL-3 set at the threshold of 99.0% sensitivity (threshold 1) using Scenario B for the screen detected cancers (SDCs) missed. *BI-RADS: Breast imaging-reporting and data system, FFDM: Full Field Digital Mammography, GE: General Electric.* $^{\delta}$Lesions reported. $^{\alpha}$*DL-3 BI-RADS density scores on processed full field digital mammograms.* $^{\beta}$ *Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data. p values were determined by using Chi squared* χ^2 *test to compare against the detected proportion of cancers cases by the true distribution for each cancer characteristic category. p values < 0.05 were considered statistically significant.*

At the 94.0% specificity threshold applied in Scenario E, for the auto recall of cases with a high suspicion not recalled by double reading, The types of cases detected at the 94.0% specificity threshold are outlined in Table 7-19 for ICs and Table 7-20 for NRCs.

	n	DL-1	DL-2	DL-3
Total Cases n	439	88	90	99
Total Lesions n	445	91	93	101
Total invasive lesions	405	87	85	94
Age at Screening				
< 60	203	30 (14.8%)	28 (13.8%)	38 (18.7%)
>= 60	236	58 (24.6%)	62 (26.3%)	61 (25.9%)
p value	-	0.036197	0.008378	0.155554
FFDM Vendor				
GE	275	72 (26.2%)	56 (20.4%)	55 (20.0%)
Philips	164	16 (9.8%)	34 (20.7%)	44 (26.8%)
p value	-	0.000536	0.940191	0.190878
Invasive Tumour Size$^{\delta}$				
< 15 mm	126	25 (19.8%)	23 (18.3%)	16 (12.7%)
>= 15 mm	228	51 (22.4%)	52 (22.8%)	66 (29.0%)
Missing	51	11 (21.6%)	10 (19.6%)	12 (23.5%)
p value	-	0.904813	0.700864	0.019921
Invasive Tumour Grade $^{\delta}$				
1	55	11 (20.0%)	11 (20.0%)	14 (25.5%)
2	189	48 (25.4%)	49 (25.9%)	46 (24.3%)
3	152	26 (17.1%)	24 (15.8%)	32 (21.1%)
Missing	9	2 (22.2%)	1 (11.1%)	2 (22.2%)
p value	-	0.516076	0.280206	0.933244
Time interval (months)				
0-12	69	13 (18.8%)	12 (17.4%)	19 (27.5%)
13-24	156	31 (19.9%)	25 (16.0%)	38 (24.4%)
25-36	214	44 (20.6%)	53 (24.8%)	42 (19.6%)
p value	-	0.966804	0.21088	0.482711
Radiological classification				
Normal / Benign	337	44 (13.1%)	52 (15.4%)	56 (16.6%)
Uncertain	78	37 (47.4%)	31 (39.7%)	36 (46.2%)
Suspicious	9	5 (55.6%)	5 (55.6%)	4 (44.4%)
Unclassifiable	6	2 (33.3%)	0 (0.0%)	2 (33.3%)
Missing	9	0 (0.0%)	2 (22.2%)	1 (11.1%)
p value	-	< 0.01	-	0.000567

Density BI-RADS	β	α	β	α	β	α	β	α
a	12	25	0 (0.0%)	3 (12.0%)	1 (8.3%)	5 (20.0%)	5 (41.7%)	6 (24.0%)
b	65	216	5 (7.7%)	54 (25.0%)	14 (21.5%)	49 (22.7%)	17 (26.2%)	43 (19.9%)
c	58	185	8 (13.8%)	29 (15.7%)	12 (20.7%)	34 (18.4%)	13 (22.4%)	47 (25.4%)
d	31	13	3 (9.7%)	2 (15.4%)	7 (22.6%)	2 (15.4%)	9 (29.0%)	3 (23.1%)
Missing	273	0	72 (26.4%)	0 (0.0%)	56 (20.5%)	0 (0.0%)	55 (20.1%)	0 (0.0%)
p value	-	-	-	0.2139	0.9271	0.8256	0.6093	0.7744

Table 7-19 – Sub group analysis of DL-1, DL-2, DL-3 set at 96.0% specificity threshold, using Scenario E for the interval cancers (ICs) detected. *BI-RADS: Breast imaging-reporting and data system, FFDM: Full Field Digital Mammography, GE: General Electric.* [δ]Lesions reported. [α]*DL-3 BI-RADS density scores on processed full field digital mammograms.* [β] *Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data. p values were determined by using Chi squared* χ^2 *test to compare against the detected proportion of cancers cases by the true distribution for each cancer characteristic category. p values < 0.05 were considered statistically significant.*

	n		DL-1		DL-2		DL-3	
Total Cases n	688		152		139		132	
Total Lesions n	718		159		144		133	
Total invasive lesions	592		134		122		111	
Age at Screening								
< 60	309		61 (19.7%)		44 (14.2%)		49 (15.9%)	
>= 60	379		91 (24.0%)		95 (25.1%)		83 (21.9%)	
p value	-		0.3208		0.0047		0.104	
FFDM Vendor								
GE	504		132 (26.2%)		106 (21.0%)		75 (14.9%)	
Philips	184		20 (10.9%)		33 (17.9%)		57 (31.0%)	
p value	-		0.0003		0.5262		0.0002	
Invasive Tumour Size[δ]								
< 15 mm	316		67 (21.2%)		54 (17.1%)		39 (12.3%)	
>= 15 mm	214		54 (25.2%)		58 (27.1%)		63 (29.4%)	
Missing	62		13 (21.0%)		10 (16.1%)		9 (14.5%)	
p value	-		0.667276		0.61173		0.00023	
Invasive Tumour Grade[δ]								
1	157		42 (26.8%)		33 (21.0%)		26 (16.6%)	
2	285		58 (20.4%)		65 (22.8%)		63 (22.1%)	
3	110		29 (26.4%)		19 (17.3%)		19 (17.3%)	
Missing	40		5 (12.5%)		5 (12.5%)		3 (7.5%)	
p value	-		0.305361		0.53279		0.224504	
Time Interval								
Median [IQR] (months)	35.7 [35.0-39.2]		35.4 [35.0-37.3]		35.8 [35.0-39.3]		35.6 [34.9-40.4]	
Density BI-RADS	β	α	β	α	β	α	β	α
a	27	100	5 (18.5%)	16 (16.0%)	3 (11.1%)	17 (17.0%)	7 (0.0%)	18 (18.0%)
b	95	401	6 (6.3%)	92 (22.9%)	20 (21.1%)	82 (20.5%)	29 (0.9%)	72 (18.0%)
c	47	186	6 (12.8%)	44 (23.7%)	7 (14.9%)	40 (21.5%)	14 (1.8%)	42 (22.6%)
d	15	1	3 (20.0%)	0 (0.0%)	3 (20.0%)	0 (0.0%)	7 (0.0%)	0 (0.0%)
Missing	504	0	132 (26.2%)	0 (0.0%)	106 (21.0%)	0 (0.0%)	95 (18.9%)	0 (0.0%)
p value	-	-	0.004912	-	0.78437	-	0.0821	-

Table 7-20 – Sub group analysis of DL-1, DL-2, DL-3 set at 96.0% specificity threshold, using Scenario E for the next round cancers (NRCs) detected. *BI-RADS: Breast imaging-reporting and data system, FFDM: Full Field Digital Mammography, GE: General Electric.* [δ]Lesions reported. [α]*DL-3 BI-RADS density scores on processed full field digital mammograms.* [β]*Volpara 5th edition BI-RADS mammographic breast density from raw full field digital mammograms for Cambridge data. p values were determined by using Chi squared χ^2 test to compare against the detected proportion of cancers cases by the true distribution for each cancer characteristic category. p values < 0.05 were considered statistically significant.*

DL-1 detected 88 ICs and 152 NRCs, DL-2 detected 90 ICs and 139 NRCs, and DL-3 detected 99 ICs and 132 NRCs. It is proposed at this threshold, and using the Scenario E approach, that these cases could be recalled for supplemental imaging using a modality such as abbreviated MRI for earlier detection. There was a statistically significant difference for the ICs detected at the 94.0% specificity for the age at screening (DL-1 and DL-2), FFDM vendor (DL-1), invasive tumour size (DL-3) and radiological classification (DL1 and DL-3), Table 7-19. In addition, there was a statistically significant difference ($p < 0.05$) for the NRCs detected for age at screening (DL-2), FFDM vendor (DL-1 and DL-3) as well as invasive tumour size (DL-3) and Volpara mammographic breast density (DL-1), Table 7-20.

7.4.7 Failure analysis

Of the 9 (1.0%) SDCs case missed by each AI algorithm at threshold 1 (99.0% sensitivity), only one case missed overlaps for all algorithms, Figure 7-10.

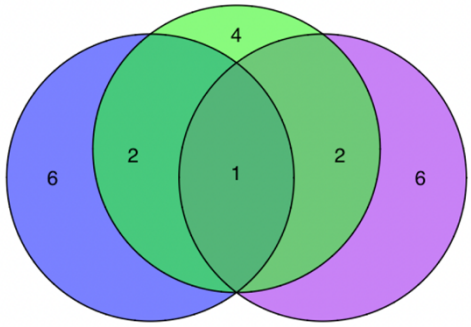


Figure 7-10 – Venn diagram – not proportional, for screen detected cancers (SDCs) missed at threshold 1, Scenario B. *For DL-1 in blue, DL-2 in purple, and DL-3 in green.*

The SDC case that was missed by all AI algorithms was from a 63-year-old patient, diagnosed with a left sided grade 2 16 mm invasive cancer from Cambridge screening, Figure 7-11.

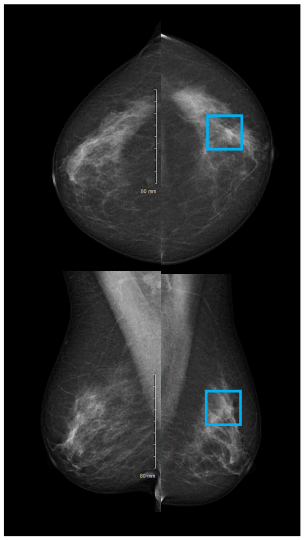


Figure 7-11 – Missing case analysis, case missed by artificial intelligence (AI). *Screening image, with a blue bounding box to show the location of the cancer. The images were annotated by a breast radiologist to show the true location of the cancer.*

Of the ICs case detected by each AI algorithm at the 94.0% threshold 34 cases overlap and for NRCs 57 cases overlap, Figure 7-12.

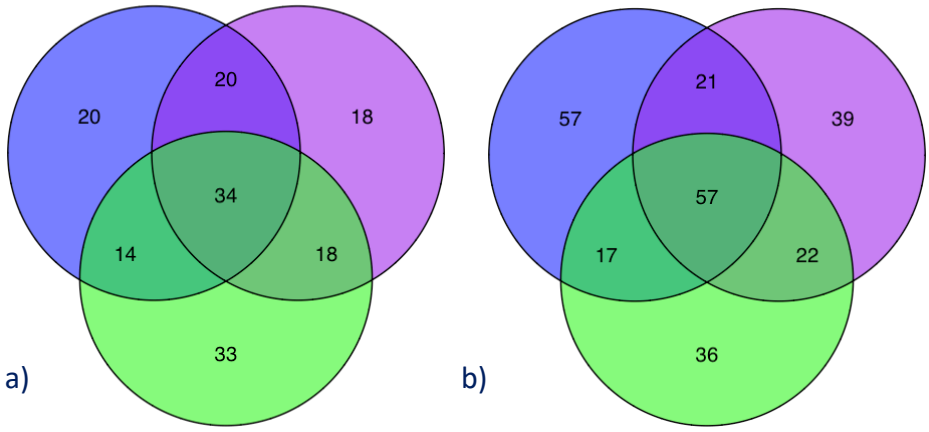


Figure 7-12 – Venn diagram – not proportional, for a) interval cancers (ICs) and b) next round cancers (NRCs) detected at the 94.0% specificity threshold Scenario E. *For DL-1 in blue, DL-2 in purple, and DL-3 in green.*

7.5 Discussion

7.5.1 Overall performance

Implementing a CADt rule out workflow found a large proportion of cases could be read by either no readers (Scenario B) or one reader (Scenario C) whilst missing between 0-34 (0.0-3.8%) SDCs. Simulating the effect on overall screening performance found the specificity and sensitivity remained non-inferior to the double reading performance at all thresholds. Similar results for rule out CADt were found in previously published papers, reporting between 17.0-91.0% cases could be not read by human readers whilst estimated to miss 0-7.0% of the cancer cases[133]. However, many of these previous studies used enriched and small datasets[135,231]. The implementation of DL for CADt could have a positive impact on the efficiency of screening and help in places where there is a shortage of trained expert human readers. Furthermore, this reduction in workload to improve efficiency could also help offset the increase in workload from the rule in triage approach to improve earlier detection of cancers. However, the question still remains as to where an acceptable threshold should be set for triage ruling out CADt applications.

At the highest specificity threshold (99.0%) for Scenario E, auto recall cases with a high suspicion not recalled by human readers, a small proportion of ICs (4.6-8.2%) and NRCs (5.2-6.1%) were recalled for supplemental imaging / assessment and could potentially be detected. However, this is a smaller number than that reported in Dembrower *et al*, 1.0% highest scores 12.0% ICs and 14.0% NRCs[134]. This is possibly due to the method used for threshold identification in our study, where we set each AI algorithm at a 94.0-99.0% specificity as appose to taking the cases with the highest 1.0-5.0% scores. Further guidance is also required for this CADt approach, as to what modality or form of

assessment would be best suited to detect these 'occult' cancers recalled by AI systems only as well as what location prompting should be provided by the AI algorithms to radiologist carrying out the additional review. MRI with the increase sensitivity compared to mammography could be offered for these mammographically 'occult' cancers.

During the running of this study two tools were updated. All the data reported in the study is from the same version of the updated algorithms. It is important in future work to monitor for the changes in performance with these updates in algorithms as well as that these processes must be time efficient to account for these frequent changes.

Alternative CADt workflows are also possible such as sending any suspicious cases back to the second reader only with the AI prompts for review, or even just using the AI algorithms to generate a smart worklist with cases prioritised in order of suspicion so the most suspicious cases are read first when potentially the readers are most alert.

7.5.2 Further analysis

When using both the rule in (Scenario C) and rule out (Scenario E) combination approach the sensitivity was found to be superior, with a trade off in specificity for two out of the three algorithms. In Lauritzen *et al*, the sensitivity was non-inferior ($p = 0.02$) and the specificity was higher ($p < 0.001$)[296]. Although in their study it was proposed normal cases were not read by human readers if the case reached the auto recall out threshold[296]. Keeping a reader in the loop in the first instance when deploying such an automated AI workflow would be beneficial for two reasons; 1) to provide human oversight to AI algorithm decisions acting as a safety net and 2) to build trust and knowledge regarding these systems by radiologists who have not trained with these systems. Interestingly, one case was missed by all AI algorithms at the set auto rule out threshold 1 demonstrating that these systems both detect and miss different cancers. Further work into the use of these systems together should be carried out to see if there is an added benefit.

7.5.3 Limitations

There are several limitations to this study. The data was from one region in the UK. The study was retrospective and so the impact on the reader can only be simulated and the true effect on reading a smaller proportion of cases cannot be evaluated. In addition, it is not definitive that the cancers flagged for supplemental imaging or assessment will be detected. Thresholds for this study were found on the study dataset and not an independent dataset, thus causing bias. All available data was used in study test set to provide a sufficient sample size, thus there was no independent dataset, without overlap with the study cohort, from which to identify thresholds. A proportion of women aged 67-69 were excluded from this study as they did not have sufficient follow-up for the required ground truth. We used a strict ground truth definition for this study. However, in recent studies a

sufficient follow-up time with a no cancer outcome for the case has been used, and is an alternative way of defining a case that would limit the loss of cases from the normal case ground truth.

7.6 Conclusion

CAD triage applications of the latest DL algorithms provide multiple workflow solutions. A large proportion of cases can be triaged out of double reading to either an automated decision of no recall or for single reading, whilst estimated to miss only 0.0-3.8% of SDCs. The potential benefit of efficiency from automated rule out triage could offset the increase in recall rate from an automated rule in approach, which provides the opportunity to improve IC and NRC detection and thus the earlier detection of some cancers. Prospective studies implementing one or more of these workflows are required to further investigate performance and the effect on reader performance. It is important to evaluate the readers acceptability of these thresholds as well as reader interaction with AI systems, as this is not possible to evaluate in simulated studies.

Chapter 8 – Contributions, Future Work and Conclusions

8.1 Contributions to knowledge

This thesis evaluated the use of artificial intelligence (AI) in breast cancer screening.

The major contributions to knowledge from this thesis include a systematic review and meta-analysis of the stand-alone use of AI in breast cancer screening, the creation of a large curated mammographic imaging database (The Cambridge Cohort – Mammography East Anglia Digital Imaging Archive (CC-MEDIA)) which provides multiple representative year data from two National Health Service Breast Screening Programme (NHSBSP) sites, a comparative analysis of three different AI algorithms for the early detection of interval cancers, a study investigating the use of three different AI algorithms as stand-alone screen readers, and an evaluation of three different AI algorithms for normal rule out and high suspicion rule in triage approaches in breast cancer screening.

The systematic review and meta-analysis presented in Chapter 3 highlighted the rapid increase in published literature over the past six years investigating the latest deep learning algorithms performance in breast cancer screening. Two key workflow applications were the focus of this review, stand-alone screen reading and triage. The performance of AI systems was comparable to the human readers. Furthermore, a large proportion of cases could be triaged whilst missing a small proportion of cancers. However, this review also established there is a high level of bias due to the use of internal datasets and no pre-setting of the algorithm threshold. In addition, the evidence was from a limited number of studies using small and enriched datasets. The gaps in evidence and standard methodology within the field identified through this review, such as ground truth classification, were then applied in Chapters 5, 6 and 7.

The creation of CC-MEDIA database outlined in Chapter 4, highlights the governance and technical processes required to create a large medical imaging database. The patient and public involvement (PPI) work carried out during the database creation helped ensure the transparent communication with patients about the use of their data in this research. This database has also been used in three separate research projects from this thesis by researchers at the University of Cambridge Radiology Department. In addition, the image extraction pipeline created, and lessons learned from deploying this method at both sites, has contributed to the ongoing development of an image extraction protocol from Cambridge University Hospitals NHS Foundation Trust PACS to the University of Cambridge Radiology Department.

Chapter 5 demonstrated that AI algorithms are able to detect breast cancer at an earlier time point using the screening mammogram. By comparing the three different algorithms on the same data set

this work identified that the interval cancers detected by each AI algorithms do differ. Fluctuations in performance were identified when translating pre-identified thresholds from other sites, thus the stability of algorithm performance when transferring between sites is an important measure and should be considered when evaluating algorithm performance. The feasibility of installing and running multiple AI systems, processing of data from the CC-MEDIA database, and methods for comparative analysis were established in this chapter and provided the basis from which to carry out the analysis in Chapters 6 and 7.

The comparative study for stand-alone screen reading presented in Chapter 6 adds to the growing body of literature for the use of AI as a stand-alone reader either entirely independent or in combination with a human reader in a double reading system. All three algorithms demonstrated non-inferiority at clinically relevant thresholds, thus reaching the required benchmark for prospective testing. All algorithms were generalisable to the NHSBSP even though less than 10% of training data was from the UK. Furthermore all algorithms were generalisable to both mammographic machine vendors (Philips and GE) included in the data, with no statistically significant difference in performance when comparing the two sites using different machines, despite all algorithms training on less than 1% Philips data. The importance of reporting metrics such as sensitivity, specificity and partial area under the receiver operating characteristic (pAUROC) curve alongside area under the receiver operating characteristic (AUROC) curve was also detailed to account for the class imbalance in screening as well as that the algorithms are operating at high specificity in screening, so as not to increase recall rates.

Chapter 7 presents the first comparative study of AI algorithms for triage applications using the same external dataset. The rule out triage approach demonstrated that all algorithms could class a large proportion of cases as 'normal' whilst missing a very small proportion of screen detected cancers. This loss of screen detected cancers could be offset by the rule in triage approach for the earlier detection of cancers. The acceptable trade-off between these two approaches requires clarification in future work through discussion between breast radiologists and the national screening programme.

8.2 Future work

8.2.1 AI in the NHS

Whilst there have been reports, briefings, and proposals for the adoption of AI into the NHS there is no established pathway for the approval of AI algorithms to be used in the NHS[297–299]. The National Screening Committee (NSC) report published in 2021 concluded that "the current evidence is a long way from the quality and quantity required for implementation into clinical practice" and so does not to support the implementation of AI into the NHSBSP[136]. Thus no algorithms are currently being

used in the programme. The report detailed that further evidence is required from both retrospective and prospective studies.

Overall the aim of the studies in Chapters 5, 6 and 7 was to address the gaps in evidence highlighted in the 2021 NSC report. The results in these three chapters demonstrate the generalisability and acceptable performance from all three AI algorithms using NHSBSP data. However, as these studies were retrospective and required simulation, the performance can only be estimated. This emphasises the need for prospective studies deploying the various workflow approaches applied in these chapters with sufficient follow-up time to account for the earlier detection cancer benefit that could be obtained from these systems.

8.2.2 Retrospective studies

Retrospective studies allow for the faster review of multiple AI algorithms for multiple different workflow applications at clinically relevant thresholds. The speed of these studies is important due to the continuous updates of AI systems as well as that there are now more than fourteen different algorithms approved by the Food and Drug Administration (FDA) for mammographic screening applications[300]. In order to improve the generalisability of results in this thesis the 127,000 case CC-MEDIA database could be used in collaboration with other databases, such as

The Optimam Mammography Image Database (OMI-DB), or additional NHSBSP sites could be added to the database in order to provide a national test set that is more representative of the seventy-five NHSBSP sites. Furthermore testing across continents using the large medical imaging databases that have been established over the past ten years, detailed in Chapter 4 Table 4-1, could allow for broader generalisability testing. This geographical expansion is also important to include a more diverse screening population to investigate for bias in algorithms. The recording of ethnicity and socioeconomic information is an important aspect of this work, however as shown in Chapter 4 Section 4.4.4 this information is often not available. The inclusion of additional sites and databases also provides wider coverage of mammographic manufacturers e.g. Hologic and Siemens and screening programmes e.g. single reader or biennial round length, not evaluated in this work of this thesis. Retrospective testing could play a role in the future benchmarking of AI algorithms to a set programme performance standard for an already approved algorithm application that has proceeded through prospective testing. As it is not feasible for all algorithms to be tested prospectively for all applications.

8.2.3 Prospective studies

Prospective studies allow for the assessment of the impact AI system have on the reader performance, overall effect on programme performance as well as the acceptability of an AI adapted workflow approach by readers. Prospective studies have been funded in the UK to evaluate both the

Kheiron system at up to fifteen NHSBSP sites and Google system at three NHS sites, although for the Google study the "AI system would not be used in patient care during the study"[301–303]. Other prospective studies around the world are taking place in the Spain, South Korea, Sweden, Norway, China and Russia, testing various deployment approaches and algorithms[238,239,304–308]. Ideally these prospective studies should be randomised to provide the highest level of evidence. In Denmark, Transpara has already been incorporated into routine screen reading to help with the Covid-19 pandemic screening backlog, where the system will be used for a triage application of cases with low scores to be read by one reader and high scores continuing to double reading, like the approach shown in Chapter 7 Scenario C of this thesis[309].

The feasibility of implementing AI into the NHSBSP is also important to consider. In this thesis all algorithms were hosted via on premises hardware in a bespoke research environment specifically designed to carry out this work. Installing and maintain such systems within the NHS is an important point to consider as technical expertise and technical infrastructure varies between sites. It has been acknowledged by the NHS that AI systems could be hosted in one of the two approved cloud providers (Microsoft Azure or Amazon Web Services) which could facilitate a more centralised oversight of these systems at each site[310]. Furthermore, the recording of AI outputs in NBSS has not yet been tested. Extracting data from NBSS to create the CC-MEDIA database required the development of unique Crystal Report queries and was a complex process which depended on expertise in this field. Lastly, the NHS Trust information governance procedures that have to be satisfied in order to deploy a new system within the NHS firewall can be extensive and take a long time for approval which should be factored into the planning of any study. As part of our work we have gone through the local Trust governance approvals for one out of the three algorithms included in this thesis to both evaluate the feasibility of this sign off process and for the initial setup of prospective work.

As outlined in Chapter 2, the factors to consider when implementing AI into the clinical workflow extend beyond the technical requirements as the ethical and legal implications should also be clarified. The Royal College of Radiologists have incorporated AI training into the updated curriculum for trainee radiologists[311]. But it is important to establish the type of training that existing radiologists should undertake before using these systems. Alongside this thesis Professor Gilbert and I have developed an online teaching module for the National Breast Imaging Academy titled "Computer-Aided Detection (CAD) and Artificial Intelligence (AI)" to provide an overview of AI in breast cancer screening for healthcare professionals.

Questions that should be addressed in future prospective studies include:

- What is an acceptable performance of an AI system to achieve for the workflow approach?

- What to do when there is a disagreement between an AI algorithm and human reader in each type of workflow deployment?
    - It is likely that these cases should proceed to arbitration for further review using the prompts provided by the AI system.
- What percentage of additional cases is it both feasible and acceptable to triage to supplemental imaging for the earlier detection of cancers?
- Does each algorithm have to be tested prospectively before deployment for each workflow application?
- What is the cost effectiveness of using the AI systems for a specific workflow application?
- Should consent be obtained from all women whose mammograms are read by AI systems?
    - This is unclear as the aim of systems is to provide standard of care if not improve detection. However, as shown by the PPI work undertaken in this thesis clear communication with the women participating in screening is required and centralised clear communication at point of invite would be most appropriate.
- Where does the legal responsibility lie when using systems as stand-alone readers?

8.2.4 Future work - AI research questions

The CC-MEDIA database described in this thesis is approved for the collection of data from 2011 until 2020 at two NHSBSP sites. Thus, the database will continue to be expanded in order to include prior screening rounds for patients which will allow for the evaluation of AI systems that incorporate the prior image, potentially improving AI performance. Other areas of interest for future work include the use of AI systems together either via an ensemble method, as explored in this thesis and in Schaffter *et al,* or in tandem to obtain the benefit from the difference in cancers detected by each system[137].

Whilst this thesis explores the main applications of AI in breast cancer screening there are remaining gaps not addressed in this thesis which were highlighted in the NSC report[136]. These include evaluating the performance for subgroup populations, including cases with breast implants or a previous cancer. Similarly projects looking at the impact of artefacts, more than four views, non-standard views on performance should be investigated. Future work should also incorporate cost effectiveness analysis for the various AI screening approaches explored in this thesis. As screening is a balance between early detection and feasibility cost effectiveness it is important aspect in the evaluation pipeline. The studies in this thesis focused on the performance of stand-alone algorithms and so did not investigate the accuracy of prompt locations provided by the AI algorithms in addition to the continuous output score. As discussed in Chapter 5, studies such as Lång *et al* have further investigated the accuracy of these prompts to establish the likelihood that an occult cancer could be

found if the AI system provided this additional guidance to the radiologists[270]. A researcher at the University of Cambridge is currently using the outputs from the studies in this thesis to establish the accuracy of the prompts provided by the AI algorithms.

The Breast Screening – Risk Adaptive Imaging for Density (BRAID) study is a large prospective multi-centre trial, investigating the use of supplemental imaging for women with Breast Imaging-Reporting and Data System (BI-RADS) classified C and D mammographic breast density[94]. The CC-MEDIA database is being used by researchers as part of this study to investigate the use of risk prediction and mammographic breast density algorithms. These studies aim to establish the best threshold to guide the use of supplemental imaging. In the BRAID study patients also complete the BOADICEA risk questionnaire, designed by researchers at the University of Cambridge[97]. The risk information from the BOADICEA risk questionnaire where possible will also be extracted from the CC-MEDIA database to be used in combination with risk prediction and mammographic breast density algorithms. The inclusion of prior screening rounds in the database allows for the assessment of change in mammographic breast density overtime. In addition, the long-term follow-up information also included allows for the calculation of five-year risk. It has been proposed that the CC-MEDIA database will then be used to build new risk prediction tools. An application has been submitted to the CC-MEDIA Database Access Committee for a team at the University of Cambridge Maths Department to receive secure access to the database in order to facilitate the development of new algorithms.

Digital Breast Tomosynthesis (DBT) is already used in screening in the USA. In the UK a large multi-centre prospective trial is currently underway, Prospective Trial of DBT in Breast Cancer Screening (PROSPECTS) trial, to establish the added benefit from DBT in the NHSBSP[312]. Numerous publications have shown improved reading times and good AI algorithm performance with DBT[313–315]. The studies performed as part of this thesis should be replicated using DBT data in the UK screening programme if DBT is implemented into the NHSBSP in the future.

8.3 Conclusions

1. Stand-alone AI algorithms achieve a similar performance compared to human reader performance, although the evidence is from a small number of studies many of which used small and enriched retrospective cohorts leading to high rates of bias in reported studies.
2. Development of a medical imaging database requires extensive ethical approvals, patient and public involvement, governance procedures as well as technical expertise.
3. AI algorithms are able to detected interval cancers at the previous screening timepoint.

4. In this UK dataset the three AI algorithms tested achieved non-inferior performance compared to the single first human reader at both screening sites when used as a stand-alone reader. In addition, when combined with the single human first reader all AI algorithms achieved a non-inferior performance compared to double reading.
5. Each AI algorithm detected different interval and next round cancers.
6. A high proportion of cases (35.0%-68.9%) can be ruled out as 'normal' or assigned for single reading only by the AI systems whilst missing a small proportion (0.0-3.8%) of screen detected cancers.
7. Up to 20% of interval and next round cancers can be detected at a high specificity threshold which could be recalled for assessment and supplemental imaging.
8. A combined approach using both rule in and rule out triage, led to a superior sensitivity performance with a trade off in specificity. A lower arbitration rate and higher recall rate was observed.

My proposal for the future of AI in clinical practice is that AI will not replace the vital role of radiologists, rather it will enhance early detection of cancer.

References

1. World Health Organization. Breast Cancer. /web/20220321085732/https://www.who.int/news-room/fact-sheets/detail/breast-cancer (2021).
2. Office for National Statistics. Cancer registration statistics, England: 2016. /web/20220321091917/https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/final2016 (2018).
3. Harbeck, N. *et al. Breast cancer*. *Nature Reviews Disease Primers* vol. 5 (2019).
4. Hu, K. *et al.* Global patterns and trends in the breast cancer incidence and mortality according to sociodemographic indices: An observational study based on the global burden of diseases. *BMJ Open* **9**, 1–8 (2019).
5. World Health Organization. Cancer. /web/20220321094701/https://www.who.int/news-room/fact-sheets/detail/cancer (2022).
6. Feng, Y. *et al.* Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes Dis.* **5**, 77–106 (2018).
7. Loibl, S., Poortmans, P., Morrow, M., Denkert, C. & Curigliano, G. Breast cancer. *Lancet* **397**, 1750–1769 (2021).
8. Viale, G. The current state of breast cancer classification. *Ann. Oncol.* **23**, (2012).
9. Berry, D. A. *et al.* Effect of screening and adjuvant therapy on mortality from breast cancer. *NEJM* **353**, 1784–92 (2005).
10. Cancer Research UK. Breast cancer survival statistics. https://web.archive.org/web/20220207071809/https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/survival.
11. Office for National Statistics. Cancer survival in England - adults diagnosed. /web/20220321100459/https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancersurvivalratescancersurvivalinenglandadultsdiagnosed (2019).
12. Berry, D. A., Cronin, K. A. & Plevritis, S. K. Influence of tumour stage at breast cancer detection on survival in modern times: Population based study in 173 797 patients. *BMJ* **351**, (2015).
13. Kalager, M. *et al.* Improved breast cancer survival following introduction of an organized

mammography screening program among both screened and unscreened women: A population-based cohort study. *Breast Cancer Res.* **11**, 1–9 (2009).

14. Tsang, J. Y. S. & Tse, G. M. Molecular Classification of Breast Cancer. *Adv. Anat. Pathol.* **27**, 27–35 (2020).
15. McCart Reed, A. E., Kalinowski, L., Simpson, P. T. & Lakhani, S. R. Invasive lobular carcinoma of the breast: the increasing importance of this special subtype. *Breast Cancer Res.* **23**, 1–16 (2021).
16. Wilson, N., Ironside, A., Diana, A. & Oikonomidou, O. Lobular Breast Cancer: A Review. *Front. Oncol.* **10**, 1–13 (2021).
17. Rakha, E. A. *et al.* Breast cancer prognostic classification in the molecular era: The role of histological grade. *Breast Cancer Res.* **12**, (2010).
18. Tan, P. H. *et al.* The 2019 World Health Organization classification of tumours of the breast. *Histopathology* **77**, 181–185 (2020).
19. World Health Organization. *WHO Classification of Tumours, 5th Edition, Volume 2: Breast Tumours*. (2019).
20. Sinn, H. P. & Kreipe, H. A brief overview of the WHO classification of breast tumors, 4th edition, focusing on issues and updates from the 3rd edition. *Breast Care* **8**, 149–154 (2013).
21. Elston, C. W. & Ellis, I. O. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up. *Histopathology* **19**, 403–410 (1991).
22. IO Ellis *et al. Pathology reporting of breast disease in surgical excision specimens incorporating the dataset for histological reporting of breast cancer*. *The Royal College of Pathologists* https://www.rcpath.org/uploads/assets/7763be1c-d330-40e8-95d08f955752792a/G148_BreastDataset-hires-Jun16.pdf (2016).
23. Dai, X. *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. *Am. J. Cancer Res.* **5**, 2929–2943 (2015).
24. Senkus, E. *et al.* Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **26**, 8–30 (2015).
25. Harbeck, N., Thomssen, C. & Gnant, M. St. Gallen 2013: Brief preliminary summary of the consensus discussion. *Breast Care* **8**, 102–109 (2013).
26. Falck, A. K., Fernö, M., Bendahl, P. O. & Rydén, L. St Gallen molecular subtypes in primary breast cancer and matched lymph node metastases - aspects on distribution and prognosis for patients with luminal A tumours: Results from a prospective randomised trial. *BMC Cancer* **13**, 1–10 (2013).

27. Kalli, S. *et al.* American joint committee on cancer's staging system for breast cancer, eighth edition: What the radiologist needs to know. *Radiographics* **38**, 1921–1933 (2018).
28. Koh, J. & Kim, M. J. Introduction of a new staging system of breast cancer for radiologists: An emphasis on the prognostic stage. *Korean J. Radiol.* **20**, 69–82 (2019).
29. Giuliano, A. E. *et al.* Breast Cancer-Major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA. Cancer J. Clin.* **67**, 290–303 (2017).
30. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
31. Wilson, J. M. G., Jungner, G. & World Health Organization. *Principles and practice of screening for disease*. /web/20220321153328/https://apps.who.int/iris/handle/10665/37650 (1968).
32. World Health Organization. *Screening Programmes: A short guide*. *WHO Press* vol. 1 https://apps.who.int/iris/bitstream/handle/10665/330829/9789289054782-eng.pdf (2020).
33. Public Health England. *Consolidated Standards for NHS Breast Screening Programme*. https://www.gov.uk/government/publications/breast-screening-consolidated-programme-standards/nhs-breast-screening-programme-screening-standards-valid-for-data-collected-from-1-april-2017 (2017).
34. Birnbaum, J. K., Duggan, C., Anderson, B. O. & Etzioni, R. Early detection and treatment strategies for breast cancer in low-income and upper middle-income countries: a modelling study. *Lancet Glob. Heal.* **6**, 885–893 (2018).
35. Duffy, S. W., Chen, T. H.-H., Smith, R. A., Yen, A. M.-F. & Tabar, L. Real and artificial controversies in breast cancer screening. *Breast Cancer Manag.* **2**, 519–528 (2013).
36. Marmot, M. G. *et al.* The benefits and harms of breast cancer screening: An independent review. *Br. J. Cancer* **108**, 2205–2240 (2013).
37. Duffy, S. W. *et al.* Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer* **126**, 2971–2979 (2020).
38. Gilbert, F. J. *et al.* Opportunities in cancer imaging: risk-adapted breast imaging in screening. *Clin. Radiol.* **76**, 763–773 (2021).
39. Clift, A. K. *et al.* The current status of risk-stratified breast screening. *Br. J. Cancer* **126**, 533–550 (2022).
40. Forrest P. *Breast cancer screening. Report to the Health Ministers of England Wales Scotland and N Ireland by a working group chaired by Professor Sir Patrick Forrest. HMSO*. https://webarchive.nationalarchives.gov.uk/ukgwa/20150506221529/http://www.cancerscreening.nhs.uk//breastscreen/publications/forrest-report.html (1986).
41. Advisory Committee on Breast Cancer Screening. Screening for breast cancer in England: Past

and future. *J. Med. Screen.* **13**, 59–61 (2006).

42. Raftery, J. & Chorozoglou, M. Possible net harms of breast cancer screening: Updated modelling of Forrest report. *BMJ* **344**, 1–8 (2012).
43. Public Health England. Achieving and maintaining the 36 month round length. /web/20220321160434/https://www.gov.uk/government/publications/breast-screening-set-and-maintain-round-length/achieving-and-maintaining-the-36-month-round-length-aug19 (2019).
44. European commission. Screening ages and frequencies. https://healthcare-quality.jrc.ec.europa.eu/european-breast-cancer-guidelines/screening-ages-and-frequencies (2022).
45. Schünemann, H. J. *et al.* Breast cancer screening and diagnosis: A synopsis of the european breast guidelines. *Ann. Intern. Med.* **172**, 46–56 (2020).
46. Monticciolo, D. L. *et al.* Breast Cancer Screening Recommendations Inclusive of All Women at Average Risk: Update from the ACR and Society of Breast Imaging. *J. Am. Coll. Radiol.* **18**, 1280–1288 (2021).
47. Lagerlund, M., Åkesson, A. & Zackrisson, S. Population-based mammography screening attendance in Sweden 2017–2018: A cross-sectional register study to assess the impact of sociodemographic factors. *Breast* **59**, 16–26 (2021).
48. National Institute for Public Health and the Enviroment. Breast Cancer Screening Programme. https://www.rivm.nl/en/breast-cancer-screening-programme (2022).
49. Cancer Registry of Norway. BreastScreen Norway. https://www.kreftregisteret.no/en/screening/BreastScreen_Norway/breastscreen-norway/ (2021).
50. Australian Goverment Department of Health. BreastScreen Australia Program. https://www.health.gov.au/initiatives-and-programs/breastscreen-australia-program (2022).
51. National Health Commission of the People's Republic of China. Chinese guidelines for diagnosis and treatment of breast cancer 2018 (English version). *Chinese J. Cancer Res.* **31**, 259–277 (2019).
52. US Preventative Services Taskforce. Breast Cancer: Screening. /web/20220321172647/https://uspreventiveservicestaskforce.org/uspstf/recommendation/breast-cancer-screening (2016).
53. Canadian Taskforce on Preventative Healthcare. Breast Cancer Update (2018). https://canadiantaskforce.ca/guidelines/published-guidelines/breast-cancer-update/ (2018).
54. NHS Digital. *Breast screening programme. England 2018-19*.

/web/20220321174209/https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme/england---2018-19 (2020).

55. Taylor-Phillips, S. & Stinton, C. Double reading in breast cancer screening: Considerations for policy-making. *Br. J. Radiol.* **93**, (2020).
56. Gale, A. G. *PERFORMS - a self assessment scheme for radiologists in breast screening*. (2019).
57. Taylor-Phillips, S. *et al.* Double reading in breast cancer screening: Cohort evaluation in the CO-OPS trial. *Radiology* **287**, 749–757 (2018).
58. The Royal College of Radiologists. *Clinical Radiology UK Workforce Census 2020 Report*. (2021).
59. National Institue for Health and Care Excellence. Familial breast cancer: classification, care and managing breast cancer and related risks in people with a family history of breast cancer. https://www.nice.org.uk/guidance/cg164 (2019).
60. Radiology Café. Production of X-rays. https://www.radiologycafe.com/frcr-physics-notes/x-ray-imaging/production-of-x-rays/ (2021).
61. Radiopaedia. Bremsstrahlung radiation. https://radiopaedia.org/articles/bremsstrahlung-radiation?lang=gb (2022).
62. Public Health England. National Diagnostic Reference Levels (NDRLs) from 19 August 2019. https://www.gov.uk/government/publications/diagnostic-radiology-national-diagnostic-reference-levels-ndrls/ndrl (2019).
63. Public Health England. *NHS Breast Screening Programme Guidance for breast screening mammographers Third edition*. (2017).
64. The Royal College of Radiologists. *Guidance on screening and symptomatic breast imaging: Fourth edition*. (2019).
65. Winkler, N. S., Raza, S., Mackesy, M. & Birdwell, R. L. Breast density: Clinical implications and assessment methods. *Radiographics* **35**, 316–324 (2015).
66. Harvey, J. A. & Bovbjerg, V. E. Quantitative Assessment of Mammographic Breast Density: Relationship with Breast Cancer Risk. *Radiology* **230**, 29–41 (2004).
67. Lian, J. & Li, K. A Review of Breast Density Implications and Breast Cancer Screening. *Clin. Breast Cancer* **20**, 283–290 (2020).
68. Yaffe, M. J. Mammographic density. Measurement of mammographic density. *Breast Cancer Res.* **10**, 1–10 (2008).
69. Sprague, B. L. *et al.* Variation in Mammographic Breast Density Assessments among Radiologists in Clinical Practice: A Multicenter Observational Study. *Ann. Intern. Med.* **165**, 457–464 (2016).

70. D'Orsi C, Sickles EA, M. E. M. *Breast Imaging Reporting and Data System: ACR BI-RADS breast imaging atlas. 5th ed. Reston, Va: American College of Radiology*. (2013).
71. Fowler, E. E., Sellers, T. A., Lu, B. & Heine, J. J. Breast Imaging Reporting and Data System (BI-RADS) breast composition descriptors: Automated measurement development for full field digital mammography. *Med. Phys.* **40**, 1–9 (2013).
72. Alomaim, W. *et al.* Variability of Breast Density Classification Between US and UK Radiologists. *J. Med. Imaging Radiat. Sci.* **50**, 53–61 (2019).
73. Ciatto, S. *et al.* Categorizing breast mammographic density: Intra- and interobserver reproducibility of BI-RADS density categories. *Breast* **14**, 269–275 (2005).
74. Lehman, C. D. *et al.* National performance benchmarks for modern screening digital mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* **283**, 49–58 (2017).
75. Sprague, B. L. *et al.* Prevalence of mammographically dense breasts in the United States. *J. Natl. Cancer Inst.* **106**, (2014).
76. Alonzo-Proulx, O., Mawdsley, G. E., Patrie, J. T., Yaffe, M. J. & Harvey, J. A. Reliability of automated breast density measurements. *Radiology* **275**, 366–376 (2015).
77. Vinnicombe, S. J. Breast density: why all the fuss? *Clin. Radiol.* **73**, 334–357 (2018).
78. Brandt, K. R. *et al.* Measurements: Implications for risk prediction and supplemental screening. *Radiology* **279**, 710–719 (2016).
79. Astley, S. M. *et al.* A comparison of five methods of measuring mammographic density: A case-control study. *Breast Cancer Res.* **20**, 1–13 (2018).
80. Matthews, T. P. *et al.* A multisite study of a breast density deep learning model for full-field digital mammography and synthetic mammography. *Radiol. Artif. Intell.* **3**, (2021).
81. Wu, N. *et al.* Breast density classification with deep convolutional neural networks. *Arxiv [Preprint]* (2017) doi:10.1109/ICASSP.2018.8462671.
82. Lehman, C. D. *et al.* Mammographic breast density assessment using deep learning: Clinical implementation. *Radiology* **290**, 52–58 (2019).
83. Yala, A., Lehman, C., Schuster, T., Portnoi, T. & Barzilay, R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* **292**, 60–66 (2019).
84. Dontchos, B. N., Yala, A., Barzilay, R., Xiang, J. & Lehman, C. D. External Validation of a Deep Learning Model for Predicting Mammographic Breast Density in Routine Clinical Practice. *Acad. Radiol.* **28**, 475–480 (2021).
85. Destounis, S. *et al.* Using volumetric breast density to quantify the potential masking risk of mammographic density. *Am. J. Roentgenol.* **208**, 222–227 (2017).

86. Destounis, S., Arieno, A., Morgan, R., Roberts, C. & Chan, A. Qualitative Versus Quantitative Mammographic Breast Density Assessment: Applications for the US and Abroad. *Diagnostics* **7**, 30 (2017).
87. McCormack, V. A. & Dos Santos Silva, I. Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1159–1169 (2006).
88. Boyd, N. F., Martin, L. J., Yaffe, M. J. & Minkin, S. Mammographic density and breast cancer risk: Current understanding and future prospects. *Breast Cancer Res.* **13**, 1–12 (2011).
89. Boyd, N. F. *et al.* Mammographic signs as risk factors for breast cancer. *Br. J. Cancer* 185 (1982).
90. Melnikow, J. *et al.* Supplemental screening for breast cancer in women with dense breasts: A systematic review for the U.S. Preventive services task force. *Ann. Intern. Med.* **164**, 268–278 (2016).
91. Miles, R. C., Lehman, C., Warner, E., Tuttle, A. & Saksena, M. Patient-Reported Breast Density Awareness and Knowledge after Breast Density Legislation Passage. *Acad. Radiol.* **26**, 726–731 (2019).
92. Mann, R. M. *et al.* Breast cancer screening in women with extremely dense breasts recommendations of the European Society of Breast Imaging (EUSOBI). *Eur. Radiol.* 4036–4045 (2022) doi:10.1007/s00330-022-08617-6.
93. Yala, A. *et al.* Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. *J. Clin. Oncol.* 8–10 (2021) doi:10.1200/jco.21.01337.
94. ClinicalTrials.gov. Breast Screening - Risk Adaptive Imaging for Density (BRAID). /web/20220322123321/https://clinicaltrials.gov/ct2/show/NCT04097366 (2020).
95. Destounis, S. V., Santacroce, A. & Arieno, A. Update on breast density, risk estimation, and supplemental screening. *Am. J. Roentgenol.* **214**, 296–305 (2020).
96. Brentnall, A. R. *et al.* A Case-Control Study to Add Volumetric or Clinical Mammographic Density into the Tyrer-Cuzick Breast Cancer Risk Model. *J. Breast Imaging* **1**, 99–106 (2019).
97. Lee, A. *et al.* BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet. Med.* **21**, 1708–1718 (2019).
98. Pal Choudhury, P. *et al.* Comparative validation of the BOADICEA and Tyrer-Cuzick breast cancer risk models incorporating classical risk factors and polygenic risk in a population-based prospective cohort of women of European ancestry. *Breast Cancer Res.* **23**, 1–5 (2021).
99. Van Veen, E. M. *et al.* Use of single-nucleotide polymorphisms and mammographic density plus classic risk factors for breast cancer risk prediction. *JAMA Oncol.* **4**, 476–482 (2018).

100. Bennett, R. L., Sellars, S. J. & Moss, S. M. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. *Br. J. Cancer* **104**, 571–577 (2011).
101. Public Health England. *NHS Breast Screening Programme Reporting, classification and monitoring of interval cancers and cancers following previous assessment*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/801400/Guidance_on_Interval_cancers_Final.pdf (2017).
102. MacInnes, E. G. *et al.* Radiological audit of interval breast cancers: Estimation of tumour growth rates. *Breast* **51**, 114–119 (2020).
103. Cornford, E. & Sharma, N. Interval Cancers and Duty of Candour, a UK Perspective. *Curr. Breast Cancer Rep.* **11**, 89–93 (2019).
104. Kerlikowske, K. *et al.* Identifying women with dense breasts at high risk for interval cancer a cohort study. *Ann. Intern. Med.* **162**, 673–681 (2015).
105. Wanders, J. O. P. *et al.* Volumetric breast density affects performance of digital screening mammography. *Breast Cancer Res. Treat.* **162**, 95–103 (2017).
106. Wanders, J. O. P. *et al.* The effect of volumetric breast density on the risk of screen-detected and interval breast cancers: A cohort study. *Breast Cancer Res.* **19**, 1–13 (2017).
107. Turing, A. M. Computing machinery and intelligence. *MIND* **LIX,** 433–460 (1950).
108. Moor, J. The Dartmouth College Artificial Intelligence Conference: The next fifty years. *AI Mag.* **27**, 87–91 (2006).
109. van Leeuwen, K. G., Schalekamp, S., Rutten, M. J. C. M., van Ginneken, B. & de Rooij, M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur. Radiol.* **31**, 3797–3804 (2021).
110. International Organization for Standardization [ISO]. ISO/IEC TR 24028:2020(en) Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. /web/20220323154646/https://www.iso.org/obp/ui/ (2020).
111. Chartrand, G. *et al.* Deep Learning: A Primer for Radiologists. *Radiographics* **37**, 2113–2131 (2017).
112. Le, E. P. V., Wang, Y., Huang, Y., Hickman, S. & Gilbert, F. J. Artificial intelligence in breast imaging. *Clin. Radiol.* **74**, 357–366 (2019).
113. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
114. Cheng, P. M. *et al.* Deep learning: An update for radiologists. *Radiographics* **41**, 1427–1445 (2021).
115. NHS England and NHS Improvment. *Diagnostic Imaging Dataset Statistical Release*. *NHS*

England vol. 1 https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2020/01/Provisional-Monthly-Diagnostic-Imaging-Dataset-Statistics-2020-01-23.pdf (2020).

116. Geras, K. J. *et al.* High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. *Arxiv [Preprint]* 1–9 (2017).
117. Shen, L. *et al.* Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Sci. Rep.* **9**, 1–12 (2019).
118. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *NeurIPS Proc* 1–9 (2012).
119. Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y. & Park, S. H. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: Results from recently published papers. *Korean J. Radiol.* **20**, 405–410 (2019).
120. Giger, M. L., Chan, H. P. & Boone, J. Anniversary paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. *Med. Phys.* **35**, 5799–5820 (2008).
121. Boyer, B., Balleyguier, C., Granat, O. & Pharaboz, C. CAD in questions / answers Review of the literature. **69**, 24–33 (2009).
122. Seung, J. K. *et al.* Computer-aided detection in full-field digital mammography: Sensitivity and reproducibility in serial examinations. *Radiology* **246**, 71–80 (2008).
123. Gilbert, F. J. & Lemke, H. Computer-aided diagnosis. *Br. J. Radiol.* **78**, 1–2 (2005).
124. Skaane, P., Kshirsagar, A., Hofvind, S., Jahr, G. & Castellino, R. A. Mammography screening using independent double reading with consensus: Is there a potential benefit for computer-aided detection? *Acta radiol.* **53**, 241–248 (2012).
125. Lehman, C. D. *et al.* Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern. Med.* **175**, 1828–1837 (2015).
126. Keen, J. D., Keen, J. M. & Keen, J. E. Utilization of Computer-Aided Detection for Digital Screening Mammography in the United States, 2008 to 2016. *J. Am. Coll. Radiol.* **15**, 44–48 (2018).
127. Rao, V. M. *et al.* How widely is computer-aided detection used in screening and diagnostic mammography? *J. Am. Coll. Radiol.* **7**, 802–805 (2010).
128. Gilbert, F. J. *et al.* Single Reading with Computer-Aided Detection for Screening Mammography. *N. Engl. J. Med.* **359**, 1675–1684 (2008).
129. Taylor, P. & Potts, H. W. W. Computer aids and human second reading as interventions in

screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. *Eur. J. Cancer* **44**, 798–807 (2008).

130. Khoo, L. A. L., Taylor, P. & Given-Wilson, R. M. Computer-aided detection in the United Kingdom National Breast Screening Programme: Prospective study. *Radiology* **237**, 444–449 (2005).
131. Tchou, P. M. *et al.* Interpretation time of computer-aided detection at screening mammography. *Radiology* **257**, 40–46 (2010).
132. Hupse, R., Samulski, M., Lobbes, M. B. & Ritse M. Mann. Computer-aided detection of masses at mammography: Interactive decision support versus prompts. *Radiology* **266**, 123–9 (2013).
133. Hickman, S. E. *et al.* Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis. *Radiology* **302**, 88–104 (2022).
134. Dembrower, K. *et al.* Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digit. Heal.* **2**, e468–e474 (2020).
135. Rodriguez-Ruiz, A. *et al.* Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur. Radiol.* **29**, 4825–4832 (2019).
136. Freeman, K. *et al. Use of artificial intelligence for image analysis in breast cancer screening - Rapid review and evidence map (UK NSC)*. (2021).
137. Schaffter, T. *et al.* Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw. open* **3**, e200265 (2020).
138. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
139. Hickman, S. E., Baxter, G. C. & Gilbert, F. J. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. *Br. J. Cancer* **125**, 15–22 (2021).
140. The Royal College of Radiologists. *Clinical Radiology UK Workforce Census 2019 Report*. (2020).
141. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
142. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Heal.* **1**, e271–e297 (2019).
143. House of Lords Select Committee on Artificial Intelligence. *AI in the UK: ready, willing and able?* (2018).
144. NHSX. *Artificial Intelligence : how to get it right Holistic guidance for the development and*

deployment of AI in health and care. (2019).

145. Geis, J. R. *et al.* Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Radiology* **293**, 436–440 (2019).
146. Park, S. H. & Han, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **286**, 800–809 (2018).
147. NHSX. *A Buyer's Guide to AI in Health and Care*. (2020).
148. Willemink, M. J. *et al.* Preparing medical imaging data for machine learning. *Radiology* **295**, 4–15 (2020).
149. Salim, M. *et al.* External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncol.* **6**, 1581–1588 (2020).
150. OPTIMAM. OMI-DB Database Information (tabular view). https://medphys.royalsurrey.nhs.uk/omidb/stats_table/ (2020).
151. Health Data Research UK. Health Data Research Innovation Gateway: About. https://www.hdruk.ac.uk/about-us/ (2020).
152. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
153. Suckling, J. *et al.* The Mammographic Image Analysis Society Digital Mammogram Database. *Expert. Medica, Int. Congr. Ser.* **1069**, 375–378 (1994).
154. Lee, R.S., Gimenez, F.L., Hoogi, A., Rubin, D. Curated Breast Imaging Subset of DDSM [Dataset]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY (2020).
155. Newitt, D., Hylton, N. on behalf of the I-SPY 1 Network and ACRIN 6657 Trial Team. Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK (2020).
156. Moreira, I. C. *et al.* INbreast: Toward a Full-field Digital Mammographic Database. *Acad. Radiol.* **19**, 236–248 (2012).
157. Dembrower, K., Lindholm, P. & Strand, F. A Multi-million Mammography Image Dataset and Population-Based Screening Cohort for the Training and Evaluation of Deep Neural Networks—the Cohort of Screen-Aged Women (CSAW). *J. Digit. Imaging* **33**, 408–413 (2020).
158. Wu, N., Phang, J., Park, J., Shen, Y., Kim, S.G., Heacock, L. et al. *The NYU Breast Cancer Screening Dataset v1.0*. https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf (2019).
159. Breast Cancer Digital Repository. More about BCDR. https://bcdr.eu/information/about (2020).

160. Lingle W Erickson BJ Zuley ML Jarosz R Bonaccio E Filippini J et al. Radiology Data from The Cancer Genome Atlas Breast Invasive Carcinoma [TCGA-BRCA] collection [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP (2020).
161. UK National Screening Committe. *Interim guidance for those wishing to incorporate artificial intelligence into the National Breast Screening Programme*. (2019).
162. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H. & Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **18**, 500–510 (2018).
163. Nagendran, M. *et al.* Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ* **368**, 1–12 (2020).
164. NHSX. AI in Health and Care Award winners. https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/ai-health-and-care-award-winners/ (2020).
165. Liu X Cruz Rivera S Moher D Calvert MJ Denniston AK and The SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).
166. Cruz Rivera S Liu X Chan A Denniston AK Calvert MJ The SPIRIT-AI and CONSORT-AI Working Group SPIRIT-AI and Group CONSORT-AI Steering Group and SPIRIT-AI and CONSORT-AI Consensus Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 1351–1363 (2020).
167. Mongan, J., Moy, L. & Kahn, C. E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol. Artif. Intell.* **2**, e200029 (2020).
168. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* **393**, 1577–1579 (2019).
169. Sounderajah, V. *et al.* Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat. Med.* **26**, 807–808 (2020).
170. Halligan, S., Altman, D. G. & Mallett, S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *Eur. Radiol.* **25**, 932–939 (2015).
171. Recht, M. P. *et al.* Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur. Radiol.* **30**, 3576–3584 (2020).
172. Pianykh, O. S. *et al.* Continuous learning AI in radiology: Implementation principles and early applications. *Radiology* **297**, 6–14 (2020).
173. Ghafur, S., Fontana, G., Halligan, J., Shaughnessy, J. O. & Darzi, A. *NHS data: Maximising its impact on the health and wealth of the United Kingdom.* (2020).

174. Gilbert, F. J., Smye, S. W. & Schönlieb, C. B. Artificial intelligence in clinical imaging: a health system approach. *Clin. Radiol.* **75**, 3–6 (2020).
175. Salim, M., Dembrower, K., Eklund, M., Lindholm, P. & Strand, F. Range of Radiologist Performance in a Population-based Screening Cohort of 1 Million Digital Mammography Examinations. *Radiology* **297**, 33–39 (2020).
176. Conant, E. F. *et al.* Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. *Radiol. Artif. Intell.* **1**, e180096 (2019).
177. Rasti, R., Teshnehlab, M. & Phung, S. L. Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks. *Pattern Recognit.* **72**, 381–390 (2017).
178. Dalmış, M. U. *et al.* Fully automated detection of breast cancer in screening MRI using convolutional neural networks. *J. Med. Imaging* **5**, 014502 (2018).
179. Zhou, J. *et al.* Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images. *J. Magn. Reson. Imaging* **50**, 1144–1151 (2019).
180. Reeves, G. K. *et al.* Comparison of the effects of genetic and environmental risk factors on in situ and invasive ductal breast cancer. *Int. J. Cancer Comp.* **131**, 930–7 (2011).
181. Green, J. *et al.* Cohort Profile : the Million Women Study. **48**, 28–29 (2019).
182. Vilmun, B. M. *et al.* Impact of adding breast density to breast cancer risk models: A systematic review. *Eur. J. Radiol.* **127**, (2020).
183. Dench, E. *et al.* Measurement challenge: protocol for international case–control comparison of mammographic measures that predict breast cancer risk. *BMJ Open* **9**, e031041 (2019).
184. Qu, Y. H. *et al.* Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using a deep learning (DL) method. *Thorac. Cancer* **11**, 651–658 (2020).
185. Ravichandran, K., Braman, N., Janowczyk, A. & Madabhushi, A. A deep learning classifier for prediction of pathological complete response to neoadjuvant chemotherapy from baseline breast DCE-MRI. in *Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis* vol. 10575 105750C (2018).
186. Huynh, B. Q., Antropova, N. & Giger, M. L. Comparison of breast DCE-MRI contrast time points for predicting response to neoadjuvant chemotherapy using deep convolutional neural network features with transfer learning. *Med. Imaging 2017 Comput. Diagnosis* **10134**, 101340U (2017).
187. Braman, N. *et al.* Deep learning-based prediction of response to HER2-targeted neoadjuvant chemotherapy from pre-treatment dynamic breast MRI: A multi-institutional validation study. *pre print arXiv2001.08570* (2020).
188. Ha, R. *et al.* Convolutional Neural Network Using a Breast MRI Tumor Dataset Can Predict

Oncotype Dx Recurrence Score. *J. Magn. Reson. Imaging* **49**, 518–524 (2019).

189. Department of Health and Social Care. Code of conduct for data-driven health and care technology. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology (2020).
190. Office for Artifical Intelligence. Department for Digital Culture, Media and Sport. Joint statement from founding members of the Global Partnership on Artificial Intelligence. https://www.gov.uk/government/publications/joint-statement-from-founding-members-of-the-global-partnership-on-artificial-intelligence/joint-statement-from-founding-members-of-the-global-partnership-on-artificial-intelligence (2020).
191. Mudgal, K. S. & Das, N. The ethical adoption of artificial intelligence in radiology. *BJR*|*Open* **2**, 20190020 (2020).
192. Ledford, H. Google health-data scandal spooks researchers. https://www.nature.com/articles/d41586-019-03574-5 (2020).
193. DeCamp, M. & Lindvall, C. Latent bias and the implementation of artificial intelligence in medicine. *J. Am. Med. Informatics Assoc.* **27**, 2020–2023 (2020).
194. Chen, I. Y. *et al.* Ethical Machine Learning in Health. *pre print arXiv2009.10576* (2020).
195. Kahn, C. E. Combatting Bias in Medical AI Systems. https://pubs.rsna.org/page/ai/blog/2020/7/ryai_editorsblog0715 (2020).
196. Department of Health and Social Care. The NHS Constitution for England. https://www.gov.uk/government/publications/the-nhs-constitution-for-england/the-nhs-constitution-for-england (2020).
197. Department of Health and Social Care. Creating the right framework to realise the benefits for patients and the NHS where data underpins innovation. https://www.gov.uk/government/publications/creating-the-right-framework-to-realise-the-benefits-of-health-data/creating-the-right-framework-to-realise-the-benefits-for-patients-and-the-nhs-where-data-underpins-innovation (2020).
198. Legislation.go.uk. Data Protection Act 2018. http://www.legislation.gov.uk/ukpga/2018/12/contents (2020).
199. Intersoft Consulting. General Data Protection Regulation (GDPR). (2020).
200. Information Comissioners Office. What are the rules on special category data? https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-are-the-rules-on-special-category-data/#scd1 (2020).
201. Information Comissioners Office. *Anonymisation: managing data protection risk code of*

practice. https://ico.org.uk/media/1061/anonymisation-code.pdf (2012).

202. HRA. Confidentiality Advisory Group. https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/confidentiality-advisory-group/ (2020).
203. The Wellcome Trust. *The One-Way Mirror: Public attitudes to commercial access to health data*. (2016).
204. NHS Digital. National data opt-out. https://digital.nhs.uk/services/national-data-opt-out (2020).
205. National Data Guardian. Caldicott Principles: a consultation about revising, expanding and upholding the principles. https://www.gov.uk/government/consultations/caldicott-principles-a-consultation-about-revising-expanding-and-upholding-the-principles (2020).
206. Ming, C. *et al.* Machine learning-based lifetime breast cancer risk reclassification compared with the BOADICEA model: impact on screening recommendations. *Br. J. Cancer* **123**, 860–867 (2020).
207. Wachter, R. M. *Making it work: harnessing the power of health information technology to improve care in England*. (2016).
208. Department For Digital Culture Media and Sport. National Data Strategy. https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy#about-the-national-data-strategy (2020).
209. Hern, A. NHS could have avoided WannaCry hack with 'basic IT security', says report. https://www.theguardian.com/technology/2017/oct/27/nhs-could-have-avoided-wannacry-hack-basic-it-security-national-audit-office (2017).
210. Moore, S. M. *et al.* De-identification of medical images with retention of scientific research value. *Radiographics* **35**, 727–735 (2015).
211. NHS. NHS Digital Academy. https://www.england.nhs.uk/digitaltechnology/nhs-digital-academy/ (2020).
212. The Topol Review. *Preparing the healthcare workforce to deliver the digital future*. (2019).
213. The Royal College of Radiologists. *Clinical Radiology Specialty Training Curriculum*. (2020).
214. American College of Radiology Data Science Institute®. FDA Cleared AI Algorithms. https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms. (2020).
215. Sechopoulos, I. & Mann, R. M. Stand-alone artificial intelligence - The future of breast cancer screening? *Breast* **49**, 254–260 (2020).
216. Watanabe, L. The Power of Triage (CADt) in Breast Imaging. *Applied Radiology* https://www.appliedradiology.com/articles/the-power-of-triage-cadt-in-breast-imaging. (2020).

217. DeAngelis, C. D. & Fontanarosa, P. B. US Preventive Services Task Force and Breast Cancer Screening. *JAMA* **303**, 172–173 (2010).

218. Pharoah, P. D. P., Sewell, B., Fitzsimmons, D., Bennett, H. S. & Pashayan, N. Cost effectiveness of the NHS breast screening programme: life table model. *BMJ* **346**, 1–8 (2013).

219. Kohli, A. & Jha, S. Why CAD Failed in Mammography. *J. Am. Coll. Radiol.* **15**, 12–14 (2017).

220. Philpotts, L. E. Can computer-aided detection be detrimental to mammographic interpretation? *Radiology* **253**, 17–22 (2009).

221. McInnes, M. D. F. *et al.* Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies The PRISMA-DTA Statement. *JAMA - J. Am. Med. Assoc.* **319**, 388–396 (2018).

222. Whiting, P. F., Rutjes, A. W. ., Westwood, M. . & Al, E. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann. Intern. Med.* **18**, 529–536 (2011).

223. UK National Screening Committe. *Use of artificial intelligence for image analysis in breast cancer screening. Rapid review and evidence map*. (2021).

224. Wolff, R. F. *et al.* PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **170**, 51–58 (2019).

225. *R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* https://www.r-project.org/.

226. *Philipp Doebler (2020). mada: Meta-Analysis of Diagnostic Accuracy. R package version 0.5.10. https://CRAN.R-project.org/package=mada*.

227. *Angelo Canty and Brian Ripley (2020). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-25. https://cran.r-project.org/web/packages/boot/boot.pdf*.

228. Reitsma, J. B. *et al.* Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J. Clin. Epidemiol.* **58**, 982–990 (2005).

229. Yala, A., Schuster, T., Miles, R., Barzilay, R. & Lehman, C. A deep learning model to triage screening mammograms: A simulation study. *Radiology* **293**, 38–46 (2019).

230. Balta, C., Rodriguez-Ruiz, A., Mieskes, C., Karssemeijer, N. & Heywang-Köbrunner, S. H. Going from double to single reading for screening exams labeled as likely normal by AI: what is the impact? *Proc. SPIE 11513, 15th Int. Work. Breast Imaging* 66 (2020) doi:10.1117/12.2564179.

231. Kyono, T., Gilbert, F. J. & van der Schaar, M. MAMMO: A Deep Learning Solution for Facilitating Radiologist-Machine Collaboration in Breast Cancer Diagnosis. *Arxiv [Preprint]* 1–18 (2018).

232. Kyono, T., Gilbert, F. J. & van der Schaar, M. Improving Workflow Efficiency for Mammography Using Machine Learning. *J. Am. Coll. Radiol.* **17**, 56–63 (2020).

233. Lotter, W. *et al.* Robust breast cancer detection in mammography and digital breast tomosynthesis using annotation-efficient deep learning approach. *Arxiv [Preprint]* 1–16 (2019).
234. Rodríguez-Ruiz, A. *et al.* Detection of breast cancer with mammography: Effect of an artificial intelligence support system. *Radiology* **290**, 1–10 (2019).
235. Rodriguez-Ruiz, A. *et al.* Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *J. Natl. Cancer Inst.* **111**, 916–922 (2019).
236. Kim, H. E. *et al.* Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit. Heal.* **2**, e138–e148 (2020).
237. Kheiron. Press release: Kheiron wins UK Government award to help solve critical challenges in UK breast screening with Mia. https://www.kheironmed.com/news/https/www.nhsx.nhs.uk/news/nhs-ai-lab-speed-cancer-and-heart-care/ (2020).
238. ClinicalTrials.gov. Development of Artificial Intelligence System for Detection and Diagnosis of Breast Lesion Using Mammography. https://clinicaltrials.gov/ct2/show/NCT03708978 (2021).
239. ClinicalTrials.gov. Experiment on the Use of Innovative Computer Vision Technologies for Analysis of Medical Images in the Moscow Healthcare System. https://clinicaltrials.gov/ct2/show/NCT04489992 (2021).
240. IBM Research Editorial Staff. DREAM Challenge results: Can machine learning help improve accuracy in breast cancer screening? https://www.ibm.com/blogs/research/2017/06/dream-challenge-results/ (2020).
241. Heaven, W. D. AI is wrestling with a replication crisis. *MIT Technology Review* https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-science-big-tech-google-deepmind-facebook-openai/?utm_source=pocket-newtab-global-en-GB. (2020).
242. Haibe-Kains, B. *et al.* The importance of transparency and reproducibility in artificial intelligence research. *Nature* **586**, E14-18 (2020).
243. Lowes, S. & Paul, S. British Society of Breast Radiology Virtual Annual Scientific Meeting 2021. *Breast Cancer Res.* **23**, 1–9 (2021).
244. Goldacre, B. & Morley, J. *Better, Broader, Safer: Using Health Data for Research and Analysis*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067053/goldacre-review-using-health-data-for-research-and-analysis.pdf (2022).
245. Prior, F. W. *et al.* TCIA: An information resource to enable open science. *Proc. Annu. Int. Conf.*

IEEE Eng. Med. Biol. Soc. EMBS 1282–1285 (2013) doi:10.1109/EMBC.2013.6609742.

246. Johnson, A. E. W. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 1–8 (2019).
247. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.* **15**, 1–19 (2018).
248. Suckling J, Parker J, Dance D, Astley S, Hutt I, Boggis C, Ricketts I, et al. Mammographic Image Analysis Society (MIAS) database v1.21 [Dataset]. https://www.repository.cam.ac.uk/handle/1810/250394 (2015).
249. Heath M, Bowyer K, Kopans d, M. R. and K. J. P. The Digital Database for Screening Mammography. http://www.eng.usf.edu/cvprg/mammography/database.html#:~:text=The Digital Database for Screening,Medical Research and Materiel Command. (1998).
250. Halling-Brown, M. D. *et al.* OPTIMAM mammography image database: A large-scale resource of mammography images and clinical data. *Radiol. Artif. Intell.* **3**, (2021).
251. Lee, R. S. *et al.* Data Descriptor: A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **4**, 1–9 (2017).
252. Lee, R.S., Gimenez, F.L., Hoogi, A., Rubin, D. Curated Breast Imaging Subset of DDSM [Dataset]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY (2020).
253. Biokeanos. INbreast Database. https://biokeanos.com/source/INBreast (2022).
254. BCDR. Breast Cancer Digital Repository. https://www.bcdr.eu/information/about (2022).
255. Raul Ramon Pollan. Improving multilayer perceptron classifiers AUC performance. (2011).
256. Jeong, J. J. *et al.* The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.5M Screening and Diagnostic Mammograms. *arXiv* 2013–2015 (2022).
257. Moser, K. *et al.* Extending the age range for breast screening in England: Pilot study to assess the feasibility and acceptability of randomization. *J. Med. Screen.* **18**, 96–102 (2011).
258. OFFIS DICOM Toolkit. DCMTK. https://support.dcmtk.org/docs/index.html (2022).
259. DICOM Standards Committee. DICOM PS3.15 2022b - Security and System Management Profiles. https://dicom.nema.org/medical/dicom/current/output/chtml/part15/ps3.15.html (2022).
260. G. van Rossum. *Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam*. (1995).
261. NHS. NHS Breast Screening Programme Central Return Data Set (KC62). https://www.datadictionary.nhs.uk/data_sets/central_return_data_sets/nhs_breast_screening_programme_central_return_data_set__kc62_.html (2022).

262. NHS Digital. Breast Screening Programme. https://digital.nhs.uk/data-and-information/publications/statistical/breast-screening-programme (2022).
263. Public Health England. Interval cancers explained in the NHS Breast Screening Programme. https://www.gov.uk/government/publications/nhs-screening-programmes-duty-of-candour/interval-cancers-explained-in-the-nhs-breast-screening-programme-notes-for-professionals-and-patients (2020).
264. Gov UK. Regional ethnic diversity. https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/regional-ethnic-diversity/latest (2020).
265. Heller, S. L., Hudson, S. & Wilkinson, L. S. Breast density across a regional screening population: Effects of age, ethnicity and deprivation. *Br. J. Radiol.* **88**, (2015).
266. Maroni, R. *et al.* A case-control study to evaluate the impact of the breast screening programme on mortality in England. *Br. J. Cancer* **124**, 736–743 (2021).
267. General Medical Council. The professional duty of candour. https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/candour---openness-and-honesty-when-things-go-wrong/the-professional-duty-of-candour (2022).
268. Mainprize, J. G. *et al.* Prediction of Cancer Masking in Screening Mammography Using Density and Textural Features. *Acad. Radiol.* **26**, 608–619 (2019).
269. Sheth, M. M. & McElligott, S. E. Case-based Review of Subtle Signs of Breast Cancer at Mammography. *Radiographics* **39**, 630–631 (2019).
270. Lång, K., Hofvind, S., Rodríguez-Ruiz, A. & Andersson, I. Can artificial intelligence reduce the interval cancer rate in mammography screening? *Eur. Radiol.* **31**, 5940–5947 (2021).
271. Larsen, M., Aglen, C. F., Lee, M. A. C. I. & Hoff, M. S. S. R. Artificial Intelligence Evaluation of 122 969 Mammography Examinations from a Population-based Screening Program. *Radiology* **000**, 1–9 (2022).
272. Hinton, B. *et al.* Deep learning networks find unique mammographic differences in previous negative mammograms between interval and screen-detected cancers: A case-case study. *Cancer Imaging* **19**, 1–9 (2019).
273. Graewingholt, A. & Rossi, P. G. Retrospective analysis of the effect on interval cancer rate of adding an artificial intelligence algorithm to the reading process for two-dimensional full-field digital mammography. *J. Med. Screen.* **28**, 369–371 (2021).
274. Arkin, C. F., Mitchell, S. & Wachtel, M. How Many Patients Are Necessary to Assess Test Performance? *JAMA J. Am. Med. Assoc.* **264**, 2074–2075 (1990).
275. Cohen, J. F. *et al.* STARD 2015 guidelines for reporting diagnostic accuracy studies:

Explanation and elaboration. *BMJ Open* **6**, 1–17 (2016).

276. *Wickham H, François R, Henry L, Müller K (2022). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.*
277. *Wickham H, Girlich M (2022). tidyr: Tidy Messy Data. https://tidyr.tidyverse.org, https://github.com/tidyverse/tidyr.*
278. Bates D, Mächler M, Bolker B, W. S. Fitting Linear Mixed-Effects Models Using lme. *J. Stat. Softw.* **4**, 1–48 (2015).
279. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, M. M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, (2011).
280. Saito T, R. M. Precrec: fast and accurate precision-recall and ROC curve calculations in R. *Bioinformatics* **33**, 145-147. (2017).
281. Grolemund G, W. H. Dates and Times Made Easy with lubridate. *J. Stat. Softw.* **40**, 1–25 (2011).
282. *Carstensen B, Plummer M, Laara E, Hills M (2022). Epi: A Package for Statistical Analysis in Epidemiology. R package version 2.46*. (2022).
283. *Matt Dowle (2021). data.table. R package version 1.14.2. https://cran.r-project.org/web/packages/data.table/index.html*.
284. Chen, H. & Boutros, P. C. VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, (2011).
285. Social Science Statistics. Easy Fisher Exact Test Calculator. https://www.socscistatistics.com/tests/fisher/default2.aspx (2022).
286. Patel, M. N., Looney, P., Young, K. & Halling-Brown, M. D. Automated collection of medical images for research from heterogeneous systems: trials and tribulations. *Med. Imaging 2014 PACS Imaging Informatics Next Gener. Innov.* **9039**, 90390C (2014).
287. Wanders, A. J. T. *et al.* Interval Cancer Detection Using a Neural Network and Breast Density in Women with Negative Screening Mammograms. *Radiology* (2022) doi:10.1148/radiol.210832.
288. Gilbert, F. J. *et al.* Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom national breast screening program. *Radiology* **241**, 47–53 (2006).
289. Freeman, K. *et al.* Use of artificial intelligence for image analysis in breast cancer screening programmes: Systematic review of test accuracy. *BMJ* **374**, (2021).
290. Sharma, N. *et al.* Large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening. *medRxiv* 2021.02.26.21252537 (2021).

291. Simel, D. L., Samsa, G. P. & Matchar, D. B. Likelihood ratios with confidence: Sample size estimation for diagnostic test studies. *J. Clin. Epidemiol.* **44**, 763–770 (1991).
292. Wu, N. *et al.* Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Trans. Med. Imaging* **39**, 1184–1194 (2020).
293. Lotter, W. *et al.* Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat. Med.* **27**, 244–249 (2021).
294. Taylor-Phillips, S., Clarke, A., Wheaton, M., Kearins, O. & Wallis, M. Fatigue and performance in interpreting breast screening mammograms. *Breast Cancer Res.* **14**, P24–P24 (2012).
295. Lång, K. *et al.* Identifying normal mammograms in a large screening population using artificial intelligence. *Eur. Radiol.* **31**, 1687–1692 (2021).
296. Lauritzen, A. D. & Lynge, E. An Artificial Intelligence – based Mammography Screening Protocol for Breast Cancer : Outcome and. *Radiology* 1–9 (2022).
297. National Institue for Health and Care Excellence. *Artificial intelligence in mammography - medtech innovation briefing (MIB242)*. https://www.nice.org.uk/advice/mib242 (2021).
298. NHS. The National Strategy for AI in Health and Social Care. https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/the-national-strategy-for-ai-in-health-and-social-care/.
299. Harwich, E. & Laycock, K. *Thinking on its own: AI in the NHS*. *Reform* vol. Jan https://reform.uk/research/thinking-its-own-ai-nhs (2018).
300. American college of radiology. ACR Data Science Institute AI Central. https://aicentral.acrdsi.org (2022).
301. Research, N. I. for H. and C. AI in Health and Care Awards - funded projects 2020. https://www.nihr.ac.uk/documents/ai-in-health-and-care-awards-funded-projects-2020/25625#Mia_Mammography_Intelligent_Assessment_-_Kheiron_Medical_Technologies (2020).
302. Kheiron Medical Technologies. The AI in Health and Care Award and Kheiron Medical. https://www.kheironmed.com/nhsx-and-kheiron-medical/ (2022).
303. Imperial College London. AI breast cancer screening project wins government funding for NHS trial. https://www.imperial.ac.uk/news/222653/ai-breast-cancer-screening-project-wins/ (2021).
304. ClinicalTrials.gov. Artificial Intelligence in Large-scale Breast Cancer Screening (ScreenTrustCAD). https://clinicaltrials.gov/ct2/show/NCT04778670?term=screening+artificial+intelligence&cond=breast+cancer&draw=2&rank=3 (2021).

305. ClinicalTrials.gov. Artificial Intelligence in Breast Cancer Screening Programs in Córdoba (AITIC) (AITIC). https://clinicaltrials.gov/ct2/show/NCT04949776?term=screening+artificial+intelligence&cond=breast+cancer&draw=2&rank=2 (20221).
306. ClinicalTrials.gov. Artificial Intelligence for breaST canceR scrEening in mAMmography (AI-STREAM). https://clinicaltrials.gov/ct2/show/NCT05024591?term=screening+artificial+intelligence&cond=breast+cancer&draw=2&rank=4 (2021).
307. ClinicalTrials.gov. Artificial Intelligence in Breast Cancer Screening in Region Östergötland Linkoping (AI-ROL). https://clinicaltrials.gov/ct2/show/NCT05048095?term=AI&cond=breast+cancer+screening&draw=2&rank=1 (2022).
308. ClinicalTrials.gov. Mammography Screening With Artificial Intelligence (MASAI) (MASAI). https://clinicaltrials.gov/ct2/show/NCT04838756?term=AI&cond=breast+cancer+screening&draw=2&rank=10 (2022).
309. ScreenPoint Medical. Transpara® Breast Care AI to help radiologists in Denmark reduce Covid screening backlog. https://www.prnewswire.co.uk/news-releases/transpara-r-breast-care-ai-to-help-radiologists-in-denmark-reduce-covid-screening-backlog-819864219.html (2021).
310. NHS Digital. Cloud products, tools and assets. https://digital.nhs.uk/services/cloud-centre-of-excellence/cloud-products-tools-and-assets (2022).
311. The Royal College of Radiologists. *Clinical Radiology Specialty Training Curriculum*. www.rcr.ac.uk (2021).
312. ClinicalTrials.gov. Prospective Trial of Digital Breast Tomosynthesis (DBT) in Breast Cancer Screening. (PROSPECTS). (2019).
313. van Winkel, S. L. *et al.* Impact of artificial intelligence support on accuracy and reading time in breast tomosynthesis image interpretation: a multi-reader multi-case study. *Eur. Radiol.* **31**, 8682–8691 (2021).
314. Geras, K. J., Mann, R. M. & Moy, L. Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives. *Radiology* **293**, 246–259 (2019).
315. Conant, E. F. *et al.* Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. *Radiol. Artif. Intell.* **1**, e180096 (2019).

Appendix 1

Definitions of commonly used terms in this review:

- Computer-aided detection (CADe) = A system which locates an abnormality within an image and provides a prompt or marker to assist a human reader.

- Computer-aided diagnosis (CADx) = A system which provides a classification for the type of abnormality found in an image. For example, at the level of cancer or no cancer for a case.

- Computer-aided triage (CADt) = A system which automatically assigns cases to normal or abnormal category. Providing a possible final case decision for normal cases and highlights abnormal cases for further human reader review.

- Stand-alone = An algorithm that interprets the whole mammogram case / exam and provides an outcome independent of human interaction or interpretation.

- Reader = A breast clinician, radiologists or reporting radiographer who reports mammographic images.

- 2D standard-view mammography = An x-ray image of breast tissue which includes two views (mediolateral oblique and cranial caudal views) for each breast (right and left).

- Adapted screening = The adjustment of radiological screening workflow by changing reading protocols. Such as using a CADt algorithm for machine only reading of normal cases and presenting a proportion of suspicious cases to a single or double reader system. Other adjustments include the possibility of using a CADe and CADx algorithm as a stand-alone system to substitute one of the readers in a double reading system.

- Testing = The evaluation of an algorithm's performance.

- Development = The training, tuning and validation of an algorithm.

- Pre-assigned thresholds = ML algorithm test performance levels (e.g., sensitivity and specificity) which are determined in the protocol and specified according to current evidence or national performance. This is in contrast to thresholds that are altered to find the optimum performance following the completion of the test.

- Clinically relevant thresholds = are the current screening programme targets (sensitivity and specificity) as well as current screening reader performance,

which ML algorithm performance is required to reach or provide a workflow solution where these standards are met. For example, in a double reading system if ML is to be used as a stand-alone reader alongside another human reader, then the thresholds for the ML algorithm could be set at current single reader performance.

- *Open database = “Neither login nor registration are required for these data collections”. We have defined this also as a public database.

- *Safeguarded database = “The safeguards include knowing who is using the data and for what purpose. The EUL outlines the restrictions on use for a particular data collection”.

- *Controlled database = “These data are only available to users who have been accredited and their data usage has been approved by the relevant Data Access Committee”.

- Private database = This is a controlled or safeguarded database as outlined above.

- External testing = When an algorithm is tested by an independent third party who has not been involved in the development of the algorithm.

- Internal testing = When an algorithm is tested by the company / academic institution that developed it.

- External dataset = A dataset that is from a different dataset to the dataset that was used for development (training and validation). This can be either geographically (from a different site or country), temporally (from a different time period) or both geographically and temporally different.

- Internal dataset = A dataset that is from the same dataset as the dataset that was used for development (training and validation), which is used for testing.

- **Gray literature = “evidence not published in commercial publications”.

*UK Data Service. Data access policy. https://www.ukdataservice.ac.uk/get-data/data-access-policy. Accessed 6 January 2021.
**Paez A. Gray literature: An important resource in systematic reviews. J Evid Based Med. 2017;10(3):233–240.

Appendix 2

Protocol registration
PROSPERO (CRD42019156016)
https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=156016

Link to protocol https://www.crd.york.ac.uk/PROSPEROFILES/156016_PROTOCOL_20200909.pdf

Registered amendments

1. 29/10/2019 submitted initial application following completion of preliminary searches
2. 12/05/2020 updated time period for the review and fields for extraction, submitted prior to final search execution
3. 12/05/2020 updated authors, submitted prior to final search execution
4. 09/09/2020 submitted an update to the final search execution in protocol

Deviations from the protocol

5. Data collection – additional items collected which were not included in the protocol, through this may introduce bias in these fields (e.g. processed mammography adjusted from processed / raw) it was felt that these fields added significant information to the review.
6. Data collected – certain data collected as part of this review is not reported in paper, however this is available on request for access to the originally extracted raw data from the authors.
7. Data collected – study authors were not contacted for further information as this was felt that it could possibly bias the results of reporting as well as confuse the metrics used to evaluate quality of reporting (CLAIM, QUADAS, PROBAST). Therefore, we have reported based on what was available in the original manuscript and supplemental material only. To ensure data extraction was robust this was checked by a third reviewer with a computer science background.
8. Meta-analysis – this was conducted only for external studies as this allowed for consistency in reporting and a larger enough number of studies to be compared. The methods from Liu et al were used to direct this analysis.

Conflicts of interest
FJG undertakes consulting for technology companies, and both FJG and SEH have research collaborations with technology companies as detailed in the conflicts of interest statement. None of these organizations had any role in the funding, conduct, or publication of the study.

Appendix 3

Digital Literature Database Search:

EMBASE (EXCERPTA MEDICA DATABASE)

Database: Embase <1996 to 2020 Week 35>
Search Strategy:
---
1 (breast* adj2 (cancer* or carcino* or tumour* or tumor* or malignan*)).ti,ab.
2 (breast* adj2 (lump* or lesion* or mass*)).ti,ab.
3 exp breast cancer/
4 (Breast adj2 (screen* or imag*)).ti,ab.
5 mammogra*.ti,ab.
6 (mammo-graph* or mastograph*).ti,ab.
7 exp mammography/
8 ((convolutional or transfer or ensemble or deep or machine*) adj2 learning).ti,ab.
9 ((deep or artificial or convolutional or neural) adj2 net*).ti,ab.
10 "artificial intelligence".ti,ab.
11 ("computer assisted diagnosis" or "computer assisted detection" or "computer aided detection" or "computer aided diagnosis").ti,ab.
12 (CNN or CAD).ti,ab.
13 exp machine learning/
14 exp artificial intelligence/
15 (Radiolo* or radiographer* or reader* or expert* or expertise or specialist* or clinician* or physician* or practitioner* or human* or doctor* or person*).ti,ab.
16 (workflow* or "clinical practice" or standalone or stand-alone or independent* or automat* or "screening tool" or "triage tool" or comput*).ti,ab.
17 1 or 2 or 3
18 4 or 5 or 6 or 7
19 8 or 9 or 10 or 11 or 12 or 13 or 14
20 15 or 16
21 17 and 18 and 19 and 20
22 limit 21 to yr="2012 - 2020"

MEDLINE (MEDICAL LITERATURE ANALYSIS AND RETRIEVAL SYSTEM ONLINE)

Database: Ovid MEDLINE(R) and Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Daily and Versions(R) <1946 to September 02, 2020>
Search Strategy:
---
1 (breast* adj2 (cancer* or carcino* or tumour* or tumor* or malignan*)).ti,ab.
2 (breast* adj2 (lump* or lesion* or mass*)).ti,ab.
3 exp breast cancer/
4 (Breast adj2 (screen* or imag*)).ti,ab.
5 mammogra*.ti,ab.
6 (mammo-graph* or mastograph*).ti,ab.
7 exp mammography/
8 ((convolutional or transfer or ensemble or deep or machine*) adj2 learning).ti,ab.
9 ((deep or artificial or convolutional or neural) adj2 net*).ti,ab.

10 "artificial intelligence".ti,ab.
11 ("computer assisted diagnosis" or "computer assisted detection" or "computer aided detection" or "computer aided diagnosis").ti,ab.
12 (CNN or CAD).ti,ab.
13 exp machine learning/
14 exp artificial intelligence/
15 (Radiolo* or radiographer* or reader* or expert* or expertise or specialist* or clinician* or physician* or practitioner* or human* or doctor* or person*).ti,ab.
16 (workflow* or "clinical practice" or standalone or stand-alone or independent* or automat* or "screening tool" or "triage tool" or comput*).ti,ab.
17 1 or 2 or 3
18 4 or 5 or 6 or 7
19 8 or 9 or 10 or 11 or 12 or 13 or 14
20 15 or 16 (6353106)
21 17 and 18 and 19 and 20
22 limit 21 to yr="2012 - 2020"

SCOPUS

((TITLE-ABS-
KEY (breast* W/2 (cancer* OR carcino* OR tumour* OR tumor* OR malignan*))) OR (TITLE -ABS-KEY (breast* W/2 (lump* OR lesion* OR mass*)))) AND ((TITLE-ABS-
KEY (breast* W/2 (screen* OR imag*))) OR (TITLE-ABS-KEY (mammogra*)) OR (TITLE-ABS-KEY (mammo-graph* OR mastograph*))) AND ((TITLE-ABS-
KEY ((convolutional OR transfer OR ensemble OR deep OR machine*) W/2 learning)) OR (TI TLE-ABS-KEY ((deep OR artificial OR convolutional OR neural) W/2 net*)) OR (TITLE-ABS-KEY ("artificial intelligence")) OR (TITLE-ABS-KEY ("computer assisted diagnosis" OR "computer assisted detection" OR "computer aided detection" OR "computer aided diagnosis")) OR (TITLE-ABS-KEY (cnn OR cad))) AND ((TITLE-ABS-
KEY (radiolo* OR radiographer* OR reader* OR expert* OR expertise OR specialist* OR clinici an* OR physician* OR practitioner* OR human* OR doctor* OR person*)) OR (TITLE-ABS-KEY (workflow* OR "clinical practice" OR standalone OR stand-
alone OR independent* OR automat* OR "screening tool" OR "triage
tool" OR comput*))) AND (LIMIT-TO (PUBYEAR , 2020) OR LIMIT-
TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-
TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-
TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-
TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012))

WEB OF SCIENCE (CORE COLLECTION)

# 18 1,998	#16 AND #15 AND #14 AND #13 Refined by: PUBLICATION YEARS: (2020 OR 2012 OR 2019 OR 2018 OR 2017 OR 2016 OR 2015 OR 2014 OR 2013) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years
# 17 3,395	#16 AND #15 AND #14 AND #13 Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 16 11,039,655 #12 OR #11
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 15 645,943 #10 OR #9 OR #8 OR #7 OR #6
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 14 59,602 #5 OR #4 OR #3
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 13 561,062 #2 OR #1
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 12 5,354,666 TS = (workflow* or "clinical practice" or standalone or stand-alone or independent* or automat* or "screening tool" or "triage tool" or comput*)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 11 6,491,265 TS = (Radiolo* or radiographer* or reader* or expert* or expertise or specialist* or clinician* or physician* or practitioner* or human* or doctor* or person*)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 10 107,204 TS = (CNN or CAD)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 9 13,595 TS = ("computer assisted diagnosis" or "computer assisted detection" or "computer aided detection" or "computer aided diagnosis")
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 8 53,218 TS = ("artificial intelligence")
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 7 388,724 TS = ((deep or artificial or convolutional or neural) NEAR/2 net*)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

# 6	202,363	TS = ((convolutional or transfer or ensemble or deep or machine*) NEAR/2 learning) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years
# 5	30	TS = (mammo-graph* or mastograph*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years
# 4	46,265	TS = (mammogra*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years
# 3	25,523	TS = (Breast NEAR/2 (screen* or imag*)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years
# 2	17,412	TS = (breast* NEAR/2 (lump* or lesion* or mass*)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years
# 1	553,266	TS = (breast* NEAR/2 (cancer* or carcino* or tumour* or tumor* or malignan*)) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

CENTRAL (COCHRANE CENTRAL REGISTER OF CONTROLLED TRIALS)

ID Search

#1 ((breast* near/2 (cancer* or carcino* or tumour* or tumor* or malignan*))):ti,ab,kw (Word variations have been searched)

#2 (breast* near/2 (lump* or lesion* or mass*))

#3 (Breast near/2 (screen* or imag*))

#4 (mammogra*)

#5 (mammo-graph* or mastograph*)

#6 ((convolutional or transfer or ensemble or deep or machine*) near/2 learning)

#7 ((deep or artificial or convolutional or neural) near/2 net*)

#8 ("artificial intelligence")

#9 ("computer assisted diagnosis" or "computer assisted detection" or "computer aided detection" or "computer aided diagnosis")

#10 (CNN or CAD)

#11 (Radiolo* or radiographer* or reader* or expert* or expertise or specialist* or clinician* or physician* or practitioner* or human* or doctor* or person*) 1090834

#12 (workflow* or "clinical practice" or standalone or stand-alone or independent* or automat* or "screening tool" or "triage tool" or comput*) 167454

#13 #1 OR #2

#14 #3 OR #4 OR #5

#15 #6 OR #7 OR #8 OR #9 #10

#16 #11 OR #12

#17 #13 AND #14 AND #15 AND #16 with Publication Year from 2012 to 2020, in Trials

Grey Database Search:

DBLP (DATABASE SYSTEMS AND LOGIC PROGRAMMING)

Machine learning Breast cancer Mammography

Then separate search for:

Deep Learning Breast cancer Mammography

ACM (ASSOCIATION FOR COMPUTER MACHINERY, FULL TEXT COLLECTION)

[[All: machine learning] AND [All: breast cancer] AND [All: mammography]] OR [[All: deep learning] AND [All: breast cancer] AND [All: mammography]] AND [Publication Date: (01/01/2012 TO 31/12/2020)]

IEEE

(("All Metadata":Machine learning AND Breast cancer AND Mammography) OR "All Metadata":Deep Learning AND Breast cancer AND Mammography)

arXiv

Query: order: -announced_date_first; size: 200; date_range: from 2012-01-01 to 2020-12-31; include_cross_list: True; terms: AND all=Machine learning AND Breast cancer AND Mammography; OR all=Deep Learning AND Breast cancer AND Mammography

Appendix 4

Fields included in the data extraction:

Table 1

Study details

1. Journal
2. Year
3. Author
4. Title

Study design

5. Design (Retrospective/ prospective)
6. Algorithm name
7. Traditional ML / Deep ML
8. Workflow application
9. Decision level

Study population (train + validation dataset)

10. Total number of cases
11. Total number of images
12. Number of normal cases (*not reported in main tables, please contact authors for the extraction tables)
13. Number of cancer cases (*not reported in main tables, please contact authors for the extraction tables)
14. Number of benign cases (*not reported in main tables, please contact authors for the extraction tables)
15. Vendor (*not reported in main tables, please contact authors for the extraction tables)
16. Country (*not reported in main tables, please contact authors for the extraction tables)

Human readers

17. Readers (number + experience)
18. Single / double / multi-reader
19. Clinical information available to readers (*not reported in main tables, please contact authors for the extraction tables)
20. Prior mammogram available to readers (*not reported in main tables, please contact authors for the extraction tables)
21. Reader reading as part of real time workflow / reader study
22. Ground truth

Algorithm performance

23. Internal / external
24. Algorithm threshold set

25. Randomised / non-randomised data split (*not reported in main tables, please contact authors for the extraction tables)
26. Bootstrapping / cross validation (resampling) (*updated to include other types of study format)
27. %normals (CI)
28. Negative Predictive Value (NPV)
29. False Negatives (FN)
30. Area Under the Curve (AUC) (CI)
31. Sensitivity (CI)
32. Specificity (CI)

Other

33. Data augmentation (flip / rotate / synthetic images) (*not reported in main tables, please contact authors for the extraction tables)
34. Handling missing data (*not reported in main tables, please contact authors for the extraction tables)
35. Compute time (*not reported in main tables, please contact authors for the main extraction tables)
36. Interpretability - e.g. heatmap / locator (*not reported in main tables, please contact authors for the extraction tables)
37. Algorithm code available
38. Funding Source (*not reported in main tables, please contact authors for the extraction tables)
39. Additional information relevant to testing (*not reported in main tables, please contact authors for the extraction tables)

Table 2

Study details

1. Journal
2. Year
3. Author
4. Title

Study population (test dataset)

5. Dataset name
6. Country where mammograms were taken
7. No. Centres
8. Year of studies
9. Vendor
10. Screen / Diagnostic
40. Digital / Film
41. Raw / Processed (*adjusted field to algorithm processing, raw and processed reported in main tables, please contact authors for the extraction tables)

11. Public / Private
12. Internal / External test set
13. Dataset Size cases
14. Dataset Size images
15. Proportion of cancers
16. Proportion of cancers that are (screen detected + subsequent round + interval) (*not reported in main tables, please contact authors for the extraction tables)

Training, validation and testing

17. Used for testing (*not reported in main tables, please contact authors for the extraction tables)
18. Dataset for testing same as train + validation (*not reported in main tables, please contact authors for the extraction tables)
19. Train / validation / test split (*not reported in main tables, please contact authors for the extraction tables)
20. Density measure
21. Average lesion size (*not reported in main tables, please contact authors for the extraction tables)
22. Age

*For clarity a refined selection of fields was included in the main extraction tables (table 1,2,3 and 4). For the details of the additional fields extracted please contact authors for these extraction tables.

Varying terminology in reported studies made the identification of data for extraction challenging. Studies included in this review were allowed to focus on ML development, validation, or both.

Appendix 5

Further description of methods for primary meta-analysis

Studies were included in the primary study level meta-analysis if they were conducted with an external dataset, the ground-truth was similar to the set standard of histopathology plus follow-up of more than one year, and enough information was provided to produce contingency tables for both the algorithm and reader (tested on the same dataset). If a study reported exams only then this was used as the case number for analysis. When a simulated case cohort (e.g. using bootstrapping) was reported, this was used for the total and cancer case size. If the same algorithm was reported in different articles for the same workflow application, then the most recent version of the algorithm was included. If a study reported multiple algorithms, then the highest performing algorithm (at the test stage) defined by AUROC was used. If multiple results for the same algorithm or reader were available in the same article, then only the highest reported study result by either AUROC or if AUROC was not available then by positive prediction (total number of true positives and true negatives) was used (from the test stage).

Appendix 6

Included articles references:

1. *+***Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. J Natl Cancer Inst. 2019;111(9):916–922.
2. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol. European Radiology. 2019;29(9):4825–4832.
3. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A Deep learning model to triage screening mammograms: a simulation study. Radiology. 2019;293(1):38–46.
4. Kyono T, Gilbert FJ, van der Schaar M. Improving workflow efficiency for mammography using machine learning. J Am Coll Radiol. 2020;17(1):56–63.
5. *+***McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94.
6. **Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit Heal. 2020;2(3):e138–e148.
7. *+***Schaffter T, Buist DSM, Lee CI, et al. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw open. 2020;3(3):e200265.
8. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology. 2019;290(3):1–10.
9. *+***Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using annotation-efficient deep learning approach. Arxiv [Preprint]. 2019;1–16. http://arxiv.org/abs/1912.11027.
10. Kyono T, Gilbert FJ, van der Schaar M. MAMMO: a deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. Arxiv [Preprint]. 2018;1–18. http://arxiv.org/abs/1811.02661.
11. Geras KJ, Wolfson S, Shen Y, et al. High-resolution breast cancer screening with multi-view deep convolutional neural networks. Arxiv [Preprint]. 2017;1–9. http://arxiv.org/abs/1703.07047.
12. *+***Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol. 2020;6(10):1581–1588.
13. Dembrower K, Wåhlin E, Liu Y, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Heal. 2020;2(9):e468–e474.
14. Balta C, Rodriguez-Ruiz A, Mieskes C, Karssemeijer N, Heywang-Köbrunner SH. Going from double to single reading for screening exams labeled as likely normal by AI: what is the impact? Proc SPIE 11513, 15th International Workshop on Breast Imaging (IWBI2020), 115130D (22 May 2020).

*Studies included in primary meta-analysis
**Studies included in secondary meta-analysis

Appendix 7

Tabular presentation for Prediction model Risk Of Bias ASsessment Tool (PROBAST) results

Study	RISK OF BIAS			APPLICABILITY		ROB
	PARTICIPANTS	OUTCOME	ANALYSIS	PARTICIPANTS	OUTCOME	OVERALL
McKinney (2020)	☹	☺	☺	?	☺	☹
Kim (2020)	☹	?	☺	☹	?	☹
Rodriguez-Ruiz [1] (2019)	?	☺	☺	☹	☺	?
Rodriguez-Ruiz [2] (2019)	?	☺	☹	☹	☺	☹
Yala (2019)	☺	☺	☹	☺	☺	☹
Kyono [1] (2019)	?	☹	☹	☹	☹	☹
Schaffter (2020)	☺	☺	☺	☺	☺	☺
Kyono [2] (2018)	?	☹	☹	☹	☹	☹
Rodriguez-Ruiz [3] (2019)	☹	☺	☹	☹	☺	☹
Geras (2017)	☹	?	☹	☹	?	☹
Lotter (2019)	☹	☺	☺	☹	☺	☹
Dembrower (2020)	☹	☺	☹	☺	☺	☹
Salim (2020)	☹	☺	☺	☺	☺	☹
Balta (2020)	☺	☹	☹	☺	☹	☹

Note.- ☺Low Risk ☹High Risk ? Unclear Risk

Tabular presentation for Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) results

Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
McKinney (2020)	☹	☺	☺	☺	?	☺	☺
Kim (2020)	☹	☹	?	?	☹	☹	?
Rodriguez-Ruiz [1] (2019)	?	☹	?	☺	☹	☹	☺
Rodriguez-Ruiz [2] (2019)	?	☹	?	☺	☹	☹	☺
Yala (2019)	☹	☺	☺	☺	☺	☺	☺
Kyono [1] (2019)	☹	☺	☹	☹	☹	☺	☹
Schaffter (2020)	☺	☺	☺	☺	☺	☺	☺
Kyono [2] (2018)	☹	☺	☹	☹	☹	☺	☹
Rodriguez-Ruiz [3] (2019)	☹	☹	?	☺	☹	☹	☺
Geras (2017)	☹	☹	?	?	☹	☹	?
Lotter (2019)	☹	☹	?	☺	☹	☹	☺
Dembrower (2020)	☹	☹	☺	☺	☺	☺	☺
Salim (2020)	☹	☺	☺	☺	☺	☺	☺
Balta (2020)	☺	☹	☹	?	☺	☺	☹

Note.- ☺Low Risk ☹High Risk ? Unclear Risk

Appendix 8

Meta-analysis results

Appendix 8 – Table 8.1 - Primary analysis – contingency table

Study	Cases	Cancer	ML						Reader					
			Sens	Spec	TP	FN	FP	TN	Sens	Spec	TP	FN	FP	TN
Rodriguez-Ruiz [1] (2019)	199	79	0.800	0.790	63	16	25	95	0.770	0.790	61	18	25	95
Schaffer (2020)	68 008	780	0.771	0.925	601	179	5042	62186	0.771	0.967	601	179	2219	65009
Lotter (2019)	285	131	0.820	0.909	107	24	14	140	0.820	0.669	107	24	51	103
Salim (2020)	113 663	739	0.819	0.966	605	134	3839	109085	0.774	0.966	572	167	3839	109085
McKinney (2020)	3 097	686	0.562	0.843	386	300	379	2032	0.481	0.808	330	356	463	1948

Note.- FN = False Negative, FP = False Positive, ML = Machine Learning, Sens = Sensitivity, Spec = Specificity, TP = True Positive, TN = True Negative

Appendix 8 – Table 8.2 - Secondary analysis – contingency table

Study	Cases	Cancer	ML						Reader					
			Sens	Spec	TP	FN	FP	TN	Sens	Spec	TP	FN	FP	TN
Rodriguez-Ruiz [1] (2019)	199	79	0.800	0.790	63	16	25	95	0.770	0.790	61	18	25	95
Rodriguez-Ruiz [1] (2019)	129	40	0.850	0.490	34	6	45	44	0.840	0.490	34	6	45	44
Rodriguez-Ruiz [1] (2019)	469	68	0.850	0.670	58	10	132	269	0.770	0.670	52	16	132	269
Rodriguez-Ruiz [1] (2019)	298	49	0.860	0.540	42	7	115	134	0.820	0.540	40	9	115	134
Rodriguez-Ruiz [1] (2019)	326	104	0.810	0.510	84	20	109	113	0.830	0.510	86	18	109	113
Rodriguez-Ruiz [1] (2019)	585	113	0.860	0.680	97	16	151	321	0.840	0.680	95	18	151	321
Rodriguez-Ruiz [1] (2019)	179	75	0.750	0.750	56	19	26	78	0.760	0.750	57	18	26	78
Rodriguez-Ruiz [1] (2019)	204	82	0.810	0.730	66	16	33	89	0.830	0.730	68	14	33	89
Kim (2020)	320	160	0.888	0.819	142	18	29	131	0.753	0.720	120	40	45	115
Schaffer (2020)	68 008	780	0.771	0.925	601	179	5042	62186	0.771	0.967	601	179	2219	65009
Schaffer (2020)	68 008	780	0.771	0.880	601	179	8067	59161	0.839	0.985	654	126	1008	66220
Lotter (2019)	285	131	0.962	0.669	126	5	51	103	0.820	0.669	107	24	51	103
Lotter (2019)	285	131	0.820	0.909	107	24	14	140						
Salim (2020)	113 663	739	0.819	0.966	605	134	3839	109085	0.774	0.966	572	167	3839	109085
Salim (2020)	113 663	739	0.670	0.966	495	244	3839	109085	0.850	0.985	628	111	1694	111230
Salim (2020)	113 663	739	0.674	0.967	498	241	3726	109198						
McKinney (2020)	3 097	686	0.562	0.843	386	300	379	2032	0.481	0.808	330	356	463	1948

Note.- FN = False Negative, FP = False Positive, ML = Machine Learning, Sens = Sensitivity, Spec = Specificity, TP = True Positive, TN = True Negative

Appendix 8 – Table 8.3 - Heterogeneity

Study	N studies	Cases	Cancer	Heterogeneity		Sens (95% CI)	Spec (95% CI)	AUROC (95% CI)
				I^2	Cochrane Q – p value			
Primary - Algorithm	5	185 252	2 415	0.000%	0.621	0.754 (0.656-0.832)	0.906 (0.829-0.950)	0.892 (0.838-0.982)
Primary - Reader	5	185 252	2 415	0.000%	0.609	0.730 (0.607-0.826)	0.886 (0.724-0.958)	0.849 (0.779-0.971)
Secondary – Algorithm	17	185 572	2 575	0.625%	0.446	0.804 (0.755-0.846)	0.821 (0.727-0.888)	0.864 (0.841-0.901)
Secondary - Reader	15	185 572	2 575	0.000%	0.783	0.785 (0.738-0.825)	0.826 (0.692-0.909)	0.836 (0.814-0.876)

Note.- AUROC = Area Under the receiver operating characteristic curve, N = Number, Sens = Sensitivity, Spec = Specificity

Appendix 9 – z-test results

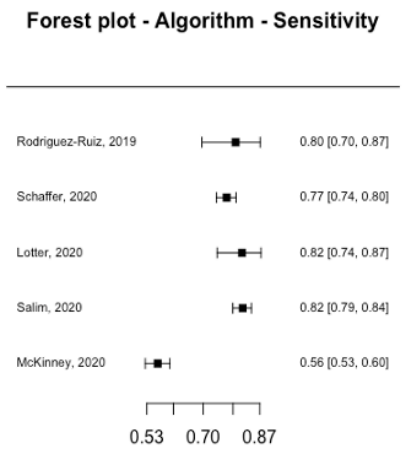
A z-test was applied to the pooled AUROC results for comparison between the ML algorithms and readers in both the primary and secondary meta-analysis, with a p-value <.05 indicating a statistically significant result.

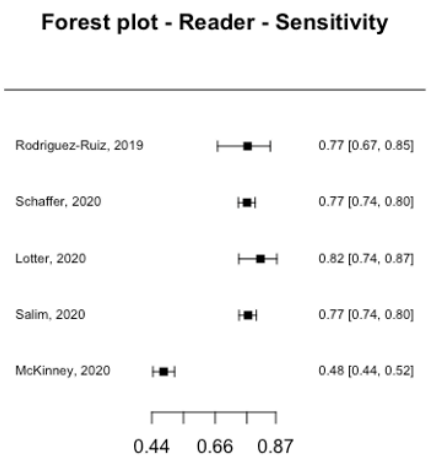
Primary analysis pooled AUROC of ML algorithm compared to pooled AUROC of readers
p-value = .53

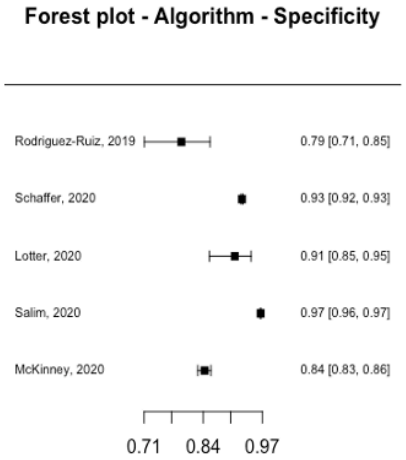
Secondary meta-analysis pooled AUROC of ML algorithm compared to pooled AUROC of readers
p-value = .84

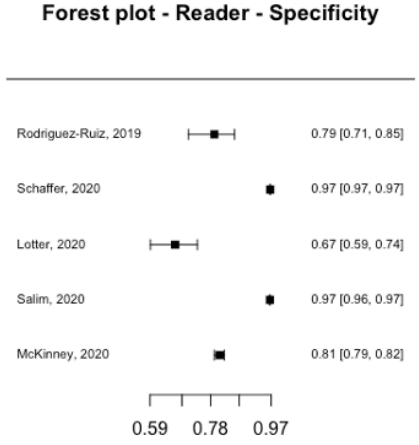
Appendix 10

Forest plots

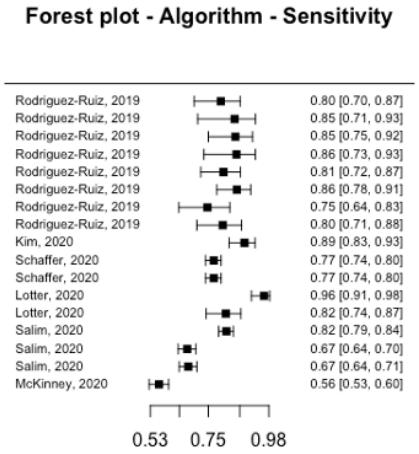


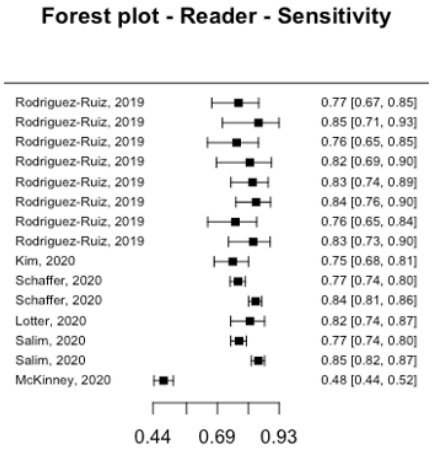


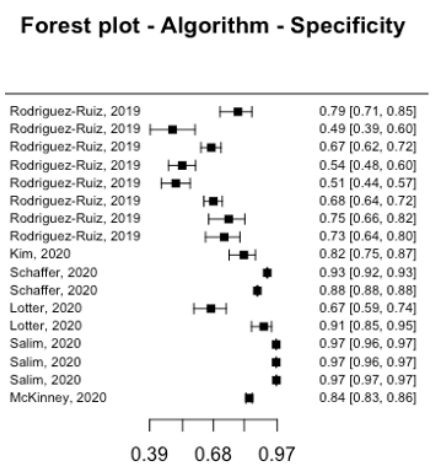


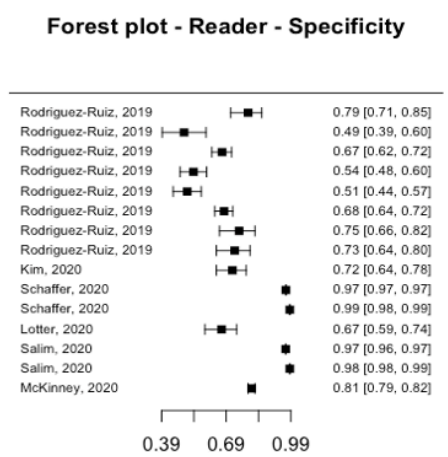


Supplemental Figure 1 - Primary analysis – Forest plot





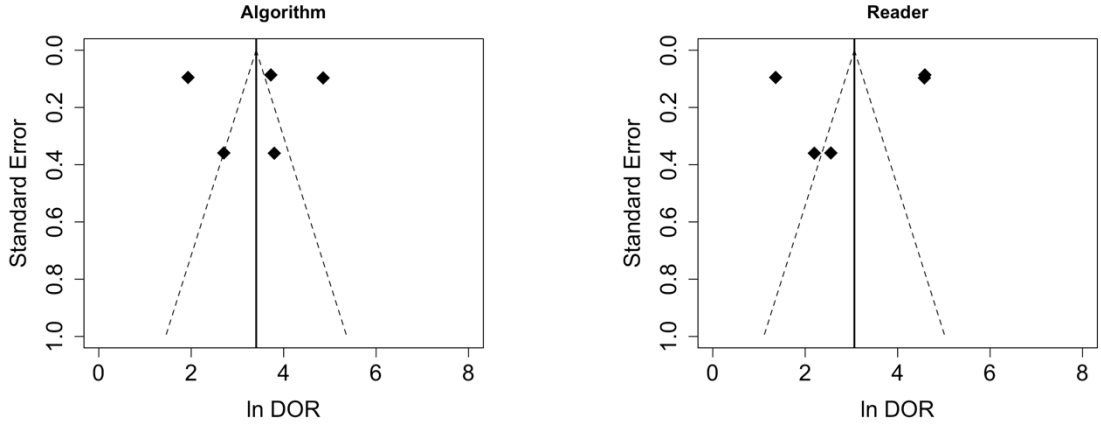




Supplemental Figure 2 - Secondary analysis – Forest plot

Appendix 11

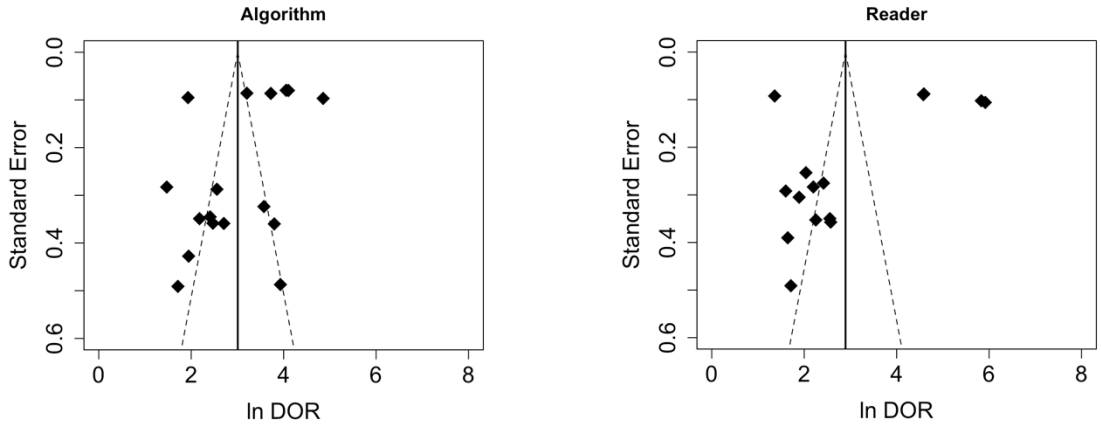
Funnel plot



Supplemental Figure 3 - Primary analysis – Funnel plots

Each algorithm and reader study result that is included in the primary meta-analysis is represented by a diamond shape. The log of diagnostic odds ratio (DOR) is plotted against standard error, with a vertical line for the median and dashed lines for the 95% confidence intervals.

For the primary analysis there are an insufficient number of studies to assess for funnel asymmetry.



Supplemental Figure 4 - Secondary analysis – Funnel plots

Each algorithm and reader study result that is included in the secondary meta-analysis is represented by a diamond shape. The log of diagnostic odds ratio (DOR) is plotted against standard error, with a vertical line for the median and dashed lines for the 95% confidence intervals.

Visual assessment of the secondary analysis funnel plots did not show asymmetry and thus does not suggest publication bias.

Appendix 12

All private DICOM tags were removed as part of the anonymisation process. The table below details the de-identification process for each DICOM tag.

(0002,0000)	FileMetaInformationGroupLength	Keep
(0002,0001)	FileMetaInformationVersion	Keep
(0002,0002)	MediaStorageSOPClassUID	Keep
(0002,0003)	MediaStorageSOPInstanceUID	Hash
(0002,0010)	TransferSyntaxUID	Keep
(0002,0012)	ImplementationClassUID	Keep
(0002,0013)	ImplementationVersionName	MATLAB
(0002,0016)	SourceApplicationEntityTitle	Keep
(0008,0005)	SpecificCharacterSet	Keep
(0008,0008)	ImageType	Keep
(0008,0016)	SOPClassUID	Keep
(0008,0018)	SOPInstanceUID	Hash
(0008,0020)	StudyDate	01/MM/YYYY
(0008,0023)	ContentDate	Blank
(0008,0030)	StudyTime	Blank
(0008,0033)	ContentTime	Blank
(0008,0050)	AccessionNumber	Exam ID
(0008,0060)	Modality	Keep
(0008,0068)	PresentationIntentType	Keep
(0008,0070)	Manufacturer	Keep
(0008,0080)	InstitutionName	Blank
(0008,0090)	ReferringPhysicianName	Blank
(0008,1030)	StudyDescription	Keep
(0008,1032)	ProcedureCodeSequence	Blank
(0008,0100)	CodeValue	Blank
(0008,0102)	CodingSchemeDesignator	Keep
(0008,0103)	CodingSchemeVersion	Keep
(0008,0104)	CodeMeaning	Keep
(0008,103E)	SeriesDescription	Keep
(0008,1090)	ManufacturerModelName	Keep
(0008,2218)	AnatomicRegionSequence	Blank
(0008,0100)	CodeValue	Keep
(0008,0102)	CodingSchemeDesignator	Keep
(0008,0104)	CodeMeaning	Keep
(0010,0010)	PatientName	Trial ID
(0010,0020)	PatientID	Trial ID
(0010,0030)	PatientBirthDate	01/01/YYYY
(0010,0040)	PatientSex	Blank
(0010,1010)	PatientAge	Keep
(0012,0062)	PatientIdentityRemoved	Yes
(0018,0015)	BodyPartExamined	Keep
(0018,0060)	KVP	Keep
(0018,1020)	SoftwareVersions	Kee

(0018,1110)	DistanceSourceToDetector	Keep
(0018,1111)	DistanceSourceToPatient	Keep
(0018,1114)	EstimatedRadiographicMagnificationFactor	Keep
(0018,1130)	TableHeight	Keep
(0018,1150)	ExposureTime	Keep
(0018,1151)	XRayTubeCurrent	Keep
(0018,1152)	Exposure	Keep
(0018,1153)	ExposureInuAs	Keep
(0018,1164)	ImagerPixelSpacing	Keep
(0018,1191)	AnodeTargetMaterial	Keep
(0018,11A0)	BodyPartThickness	Keep
(0018,11A2)	CompressionForce	Keep
(0018,1405)	RelativeXRayExposure	Keep
(0018,1508)	PositionerType	Keep
(0018,1510)	PositionerPrimaryAngle	Keep
(0018,5101)	ViewPosition	Keep
(0018,7004)	DetectorType	Keep
(0018,7005)	DetectorConfiguration	Keep
(0018,700C)	DateOfLastDetectorCalibration	Blank
(0018,700E)	TimeOfLastDetectorCalibration	Blank
(0018,7020)	DetectorElementPhysicalSize	Keep
(0018,7022)	DetectorElementSpacing	Keep
(0018,7050)	FilterMaterial	Keep
(0020,000D)	StudyInstanceUID	Hash
(0020,000E)	SeriesInstanceUID	Hash
(0020,0010)	StudyID	Exam ID
(0020,0011)	SeriesNumber	Keep
(0020,0013)	InstanceNumber	Keep
(0020,0020)	PatientOrientation	Keep
(0020,0052)	FrameOfReferenceUID	Keep
(0020,0062)	ImageLaterality	Keep
(0020,1040)	PositionReferenceIndicator	Blank
(0028,0002)	SamplesPerPixel	Keep
(0028,0004)	PhotometricInterpretation	Keep
(0028,0006)	PlanarConfiguration	Keep
(0028,0010)	Rows	Keep
(0028,0011)	Columns	Keep
(0028,0100)	BitsAllocated	Keep
(0028,0101)	BitsStored	Keep
(0028,0102)	HighBit	Keep
(0028,0103)	PixelRepresentation	Keep
(0028,0106)	SmallestImagePixelValue	Keep
(0028,0107)	LargestImagePixelValue	Keep
(0028,0301)	BurnedInAnnotation	Keep
(0028,1040)	PixelIntensityRelationship	Keep
(0028,1041)	PixelIntensityRelationshipSign	Keep
(0028,1052)	RescaleIntercept	Keep
(0028,1053)	RescaleSlope	Keep
(0028,1054)	RescaleType	Keep
(0028,1300)	BreastImplantPresent	Keep

(0028,1350)	PartialView	Keep
(0028,2110)	LossyImageCompression	Keep
(0040,0316)	OrganDose	Keep
(0040,0318)	OrganExposed	Keep
(0040,8302)	EntranceDoseInmGy	Keep
(0054,0220)	ViewCodeSequence	Blank
(0008,0100)	CodeValue	Keep
(0008,0102)	CodingSchemeDesignator	Keep
(0008,0104)	CodeMeaning	Keep
(0054,0222)	ViewModifierCodeSequence	Blank
(2050,0020)	PresentationLUTShape	Keep
(7FE0,0010)	PixelData	Blank