medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

# Large-scale evaluation of an AI system as an independent reader for double reading in breast cancer screening

Nisha Sharma, FRCR[1]*, Annie Y. Ng, Ph.D.[2†*], Jonathan J. James, FRCR[3], Galvin Khara, Ph.D.[2], Eva Ambrozay, M.D.[4], Christopher C. Austin, M.D.[2], Gabor Forrai, M.D. Ph.D.[5], Ben Glocker, Ph.D.[6], Andreas Heindl, Ph.D.[2], Edit Karpati, M.D.[2,4], Tobias M. Rijken, M.Sc.[2], Vignesh Venkataraman, Ph.D.[2], Joseph E. Yearsley, M.Sc.[2], Peter D. Kecskemethy, Ph.D.[2]

## Abstract

Screening mammography with two human readers increases cancer detection and lowers recall rates, but high resource requirements and a shortage of qualified readers make double reading unsustainable in many countries. The use of AI as an independent reader may yield more objective, accurate and outcome-based screening. Clinical validation of AI requires large-scale, multi-site, multi-vendor studies on unenriched cohorts.

This retrospective study evaluated the performance of the Mia™ version 2.0.1 AI system from Kheiron Medical Technologies on an unenriched sample (275,900 cases from 177,882 participants) collected across seven screening sites in two countries and four hardware vendors, and is representative of a real-world screening population over 10 years. Performance was determined for standalone AI and simulated double reading to assess non-inferiority and superiority on relevant screening metrics.

Standalone AI showed superiority on sensitivity and non-inferiority on specificity while detecting 29.7% of cancers found within three years after screening, and 29.8% of missed interval cancers. Double reading with AI was at least non-inferior compared to human double reading at every metric, with superiority for recall rate, specificity and positive predictive value (PPV). AI as an independent reader reduced the workload, but increased arbitration rate from 3.3% to 12.3%. Applying the AI system under investigation would have reduced the overall number of human reads required by 44.8%. The recall rate was reduced by a relative 4.1%, suggesting there could be fewer follow-up procedures, reduced stress for patients, and less administrative and clinical work.

Using the AI system as an independent reader maintains the standard of care of double reading, detects cancers missed by human readers, while automating a substantial part of the workflow, and could therefore bring significant clinical and operational benefits.

[1]The Leeds Teaching Hospital NHS Trust, Leeds, UK. [2]Kheiron Medical Technologies, London, UK. [3]Nottingham Breast Institute, City Hospital, Nottingham University Hospitals NHS Trust, Nottingham, UK. [4]Medicover, Hungary. [5]Duna Medical Center, Budapest, Hungary; European Society of Breast Imaging (EUSOBI). [6]Department of Computing, Imperial College London, London, UK. *These authors contributed equally to this work. †Corresponding author: annie@kheironmed.com

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

# Introduction

Despite improvements in therapy, breast cancer remains the leading cause of cancer-related mortality among women worldwide, accounting for approximately 600,000 deaths annually (1). Randomised trials and incidence–based mortality studies have demonstrated that population-based screening programmes substantially reduce breast cancer mortality (2-6).

Full-field digital mammography (FFDM) is the most widely used device for breast cancer screening globally, but involves a complex interpretative task (7). It has been shown that using two readers (double reading), with arbitration, increases cancer detection rates by 6-15%, whilst keeping recall rates low (8-10). The model is standard practice in at least 27 countries in Europe, Japan, Australia and the Middle East (11). The high cost of two expert readers to interpret every mammogram, alongside growing shortages of qualified screening readers, means double reading is difficult to sustain in many countries (12-14).

Breast radiology has experience using computer-aided detection (CAD) software to automate the analysis of screening mammograms. Whilst widely adopted by over 83% of US facilities (15), recent studies question its benefit to screening outcomes (16-17). When tested in the United Kingdom National Health Service Breast Screening Programme (UK NHSBSP) as an alternative to double reading, traditional CAD led to a reduction in specificity with a significant increase in recall rates (18).

Modern artificial intelligence (AI) has emerged as a promising alternative. Recent works suggest that the current generation of AI-based algorithms may interpret mammograms at least at the level of human readers (19-23). Evaluations were based on small-scale reader studies (19-21) and larger scale retrospective evaluations (21-23) performed on artificially enriched sets, often involving resampling in an attempt to approximate a more representative screening population, and with dataset images significantly skewed towards a single mammography hardware vendor in each case. AI and its true potential to positively transform clinical practice on real-world screening populations remains to be confirmed.

There is a need for rigorous large-scale studies to assess the performance of AI for mammography in double reading on diverse cohorts of women across multiple screening sites and programmes, and on unenriched data representative of a true screening population. Such studies should evaluate the AI's performance on images from various hardware vendors, using the most relevant clinical screening metrics. The aim of this study was to evaluate the ability of a novel AI system to act as a reliable independent reader in a double reading workflow, as well as demonstrating its standalone performance compared to the historical results. In this context, the study makes an important contribution, providing evidence that using AI maintains the standard of care of double reading, detects cancers missed by human readers, while automating a substantial part of the workflow.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

# Methods

## Study design

The AI system was evaluated through two separate tests, both comparing performance to the historical standard of care. The first compared the standalone performance of the AI system to the historical first human reader. This first reader opinion was selected because it was the only guaranteed independent read at all participating sites. The second test compared simulated double reading performance, using AI as an independent second reader, to the historical human double reading.

All comparisons were determined on the same unenriched cohorts, representative of a real-world screening population. Performance was measured in terms of sensitivity, specificity, recall rate, cancer detection rate (CDR), positive predictive value (PPV), and arbitration rate (rate of disagreement between the first and second readers) (see Appendix 3 for more details).

The statistical plan (see Statistical Methods) for analyses was developed, agreed with and executed by an external speciality Clinical Research Organisation (CRO) (Veristat LLC). All results presented for the listed metrics are CRO-verified.

The study had UK National Health Service (NHS) Health Research Authority (HRA) (REC reference: 19/HRA/0376) and ETT-TUKEB (Medical Research Council, Scientific and Research Ethics Committee, Hungary) approval (Reg no: OGYÉI/46651-4/2020).

## Study population and samples

All analyses were conducted on a historical cohort of de-identified cases from seven European sites representing four centers: three from the UK and one in Hungary (HU), between 2009 and 2019. The three UK centers were Leeds Teaching Hospital NHS Trust (LTHT), Nottingham University Hospitals NHS Trust (NUH), and United Lincolnshire Hospitals NHS Trust (ULH). All participate in the UK NHSBSP overseen by Public Health England (PHE) and adhere to a three-year screening interval, with women between 50 and 70 years old invited to participate with a small cohort between 47 and 49 years, and 71 and 73 years added under the UK age extension trial (Age X) (24). The Hungarian center, MaMMa Klinika (MK), included four sites and corresponding mobile screening units, which all follow a two-year screening interval, and invite eligible women aged 45 to 65. At all sites, women outside the regional screening programme age range can choose to participate as per standard of care (opportunistic screening). Screening cases were acquired from the dominant hardware vendor at each site, which included Hologic (at LTHT), GE Healthcare (NUH), Siemens (ULH), and IMS Giotto (MK).

A total of 453,599 cases were extracted from all sites. Cases were excluded in three steps, creating a highly representative sample with minimal bias in the study data (figure 1A). The largest portion of exclusions were cases outside of routine screening. This resulted in a final cohort of 275,900 eligible cases from 177,882 participants. All eligible cases were used for analysis, allowing multiple cases per participant, representative of a multi-interval real-world setting.

## Standard of care double reading and double reading with an AI system

At all sites, the first reader opinion was made in isolation, and the second reader had access, at their discretion, to the opinion of the first. When both opinions agreed, a definitive “recall” or “no recall” decision was reached. In cases of disagreement, an arbitration, performed by one or two radiologists, made the definitive “recall” or “no recall” decision (figure 1B).

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

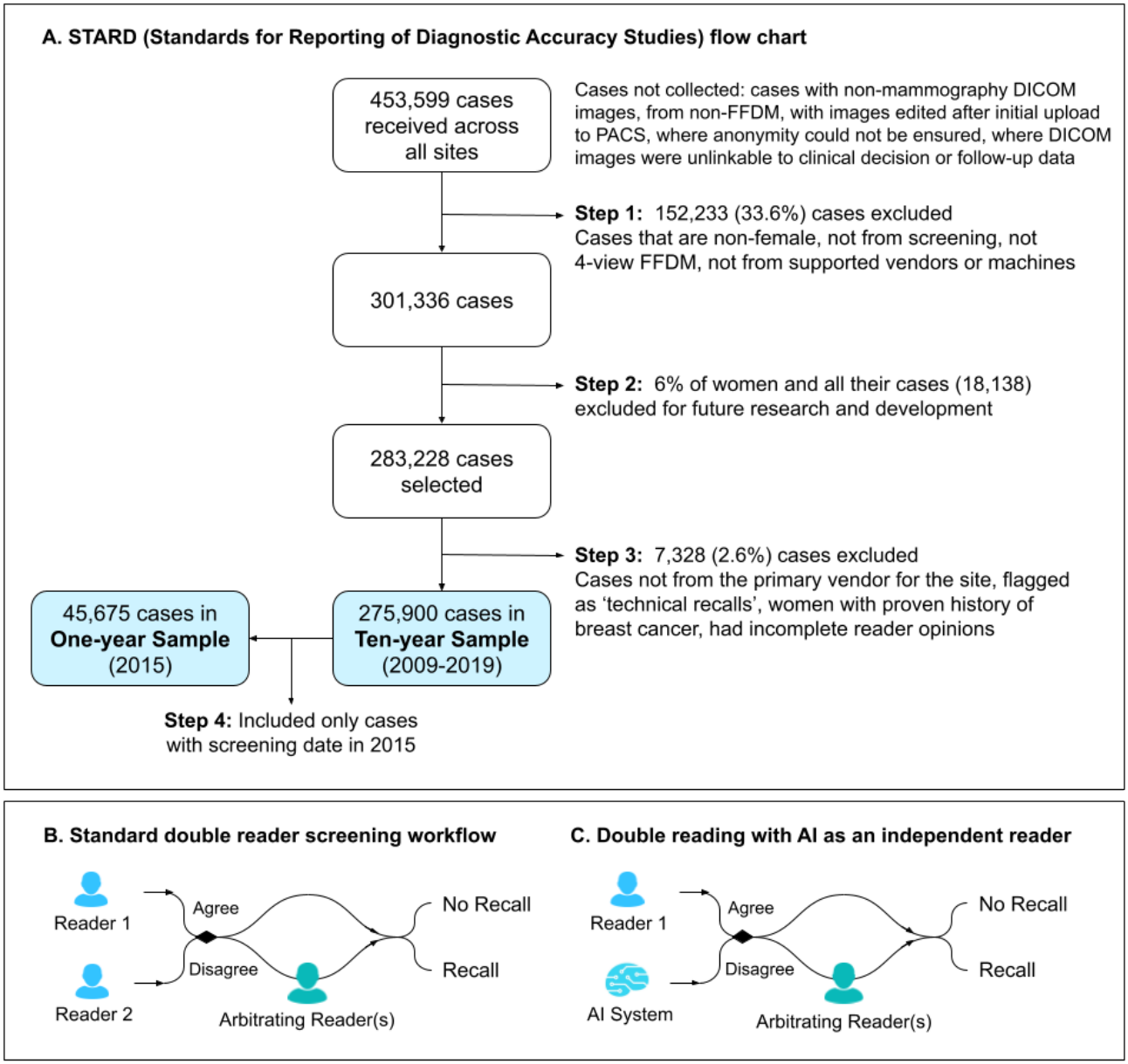


**Figure 1. A.** STARD (Standards for Reporting of Diagnostic Accuracy Studies) flow chart describing case eligibility and the final two study samples, 'ten-year' and 'one-year'. **B.** Standard double reader screening workflow. **C.** Double reading with AI as an independent reader

---

Double reading with the AI system (figure 1C) was simulated by combining the opinion of the historical first reader with the AI system. When both agreed, a definitive "recall" or "no recall" decision was made. Upon disagreement, if available, the historical arbitration opinion was used, otherwise the historical second reader opinion was chosen. This is an approximation since the second reading, although not always independent from the first read, has a different performance to arbitration.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

## AI System

All study cases were analysed by the Mia™ version 2.0.1 'AI system', developed by Kheiron Medical Technologies at its primary operating point. The AI system works with standard DICOM (Digital Imaging and Communications in Medicine) cases as inputs, analyses four images with two standard FFDM views per breast, and generates a binary suggestion of "recall" (for further assessment due to suspected malignancy) or “no recall” (until the next screening interval). The AI system’s output is deterministic, and is based on a single prediction per case avoiding methods requiring specialised compute resources. The version of the AI software was fixed prior to the study. All study data came from participants whose data was never used in any aspect of algorithm development and was separated from and inaccessible for research and development.

## Determining ground truth, subsample definitions and metrics

Sensitivity, CDR, and PPV were calculated with positives defined as ‘screen-detected positives’ and ‘three-year subsequent cancers’, collectively. Screen-detected positives were screening cases correctly identified by the historical double reader workflow, with a pathology-proven malignancy confirmed by cytology, core biopsy and/or histology of the surgical specimen within 180 days of the screening exam. Three-year subsequent cancers were defined as a screening case with a pathology-proven cancer arising within 1,095 days following the original screening date, aligned with the definition of interval cancers (IC) for three-year screening interval programmes such as in the UK. Given the two-year screening interval in practice at MK, this means that all ICs within the two-year screening interval (‘two-year ICs’) and additional cancers detected at the next screening round were also included as ‘three-year subsequent cancer’ cases. Recognising the importance of screening interval differences, regional analyses for UK and HU were also performed, using two-year ICs in place of three-year subsequent cancers for HU.

Specificity was calculated on negatives defined as any screening case with evidence of a negative follow-up result that includes a mammography reading at least 1,035 days (i.e. two months less than a three-year screening interval) after the original screening date with no proof of malignancy in between. PPV, CDR, recall rate, and arbitration rate were calculated on all 275,900 eligible cases. No ground truthing was required for recall rate or arbitration rate (see Appendix 3).

For cases that were read between 2016 and 2019, sufficient time had not elapsed to ensure complete IC data collection. To mitigate this limitation, a ‘one-year’ sample of 45,675 cases from calendar year 2015 was also analysed. This provided more reliable information on historical missed cancers, at the cost of a smaller sample size of 45,675 cases (see Appendix 2).

## Statistical methods

A 95% confidence level was used for all confidence intervals (CIs), non-inferiority and superiority testing. Non-inferiority and superiority were tested using relative differences, as standard. Non-inferiority was defined to rule out a relative difference of more than 10% in the direction of reduced performance with a 97.5% confidence. The 10% margin has been previously used for the assessment of mammography screening with CAD systems but the 97.5% non-inferiority confidence (from the use of two-sided 95% CIs) is stricter than the 95% commonly used (18). Superiority was tested when non-inferiority was passed.

Each vendor (and corresponding study center) had an equal contribution to the observed metrics in this evaluation, for point estimates, confidence intervals and hypothesis tests. Multiple cases were allowed per participant in the ten-year sample, while 99.98% of participants had one case in the one-year sample.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

**Table 1: Characteristics of ten-year and one-year samples.**

| Characteristics | | Ten-year sample (2009-2019) | | One-year sample (2015) | |
|---|---|---|---|---|---|
| | | Number of cases | Proportion of study population | Number of cases | Proportion of study population |
| **Total** | | 275,900 | 100.0% | 45,675 | 100.0% |
| Center / Vendor | MK / IMS Giotto | 83,410 | 30.2% | 10,462 | 22.9% |
| | NUH / GE | 69,045 | 25.0% | 10,983 | 24.0% |
| | LTHT / Hologic | 64,645 | 23.4% | 10,717 | 23.5% |
| | ULH / Siemens | 58,800 | 21.3% | 13,513 | 29.6% |
| Age | <40 | 483 | 0.2% | 5 | <0.1% |
| | 40 - 49 | 37,696 | 13.7% | 5,575 | 12.2% |
| | 50 - 59 | 114,524 | 41.5% | 19,399 | 42.5% |
| | 60 - 69 | 98,289 | 35.6% | 16,772 | 36.7% |
| | 70 - 79 | 23,359 | 8.5% | 3,702 | 8.1% |
| | 80 - 89 | 1,534 | 0.6% | 221 | 0.5% |
| | >90 | 15 | <0.1% | 1 | <0.1% |
| Positives | Total positives[1] | 2,792 | 1.01% | 493 | 1.08% |
| | Screen-detected positives | 2,310 | 0.84% | 365 | 0.80% |
| | Three-year-subsequent cancer | 482 | 0.17% | 128 | 0.28% |
| | Three-year ICs from UK[2] | 289 | 0.10% | 80 | 0.18% |
| | Two-year ICs from HU[2] | 84 | 0.03% | 12 | 0.03% |

See Appendix 1 for annual breakdown of samples.

1 Used for sensitivity, CDR, and PPV calculations.

2 Used for regional analyses

# Results

Table 1 presents characteristics of the study population. There were 2792 (1.0%) positives overall, made up of 2310 (0.8%) screen-detected positives (in-line with screening expectations) and 482 (0.17%) three-year subsequent cancers. For the one-year sample, the percentage of three-year subsequent cancers was significantly higher, comprising 26.0% of all positives, up from 17.3% in the ten-year sample. The IC rates in both the overall and one-year sample were below expectations, which limit the number of positives in the sample (see Appendix 2).

## Standalone AI behaviour

While the AI system is not aimed to operate as a standalone reader in clinical practice, assessing the standalone behaviour characterises the contribution the AI system could have as an independent reader in the overall double reading workflow. Table 2 presents the comparison results between the standalone AI system and the historical first reader. When measuring the AI system performance on historically screen-detected positives without three-year subsequent cancers, the sensitivity was 88.0% (86.7%, 89.3%).

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

**Table 2: Standalone AI behavior compared with the first reader – results pooled across regions.**

| Performance Metric | Historical first reader (%) | Standalone AI (%) | Test outcome for AI[1] |
|---|---|---|---|
| On ten-year sample: with incomplete IC data available | | | |
| Sensitivity[2] | 76.4 (74.9,78.0) | 78.1 (76.6, 79.7) | **Superior** |
| Specificity | 96.0 (95.9, 96.2) | 91.2 (91.0, 91.4) | **Non-inferior** |
| On one-year sample: with more complete IC data available | | | |
| Sensitivity[2] | 70.1 (66.1, 74.1) | 75.2 (71.3, 79.0) | **Superior** |
| Specificity | 96.6 (96.3, 97.0) | 91.4 (91.0, 91.9) | **Non-inferior** |

95% confidence intervals are presented in parentheses.

1. All test outcomes were based on the relative difference with a two-sided 95% CI. A 10% margin was used for non-inferiority testing (see Statistical Methods for details).
2. The positive pool for sensitivity includes screen-detected positives and 'three-year subsequent cancers' (i.e. three-year ICs for the UK plus two-year ICs and additional cancers detected at the next screening round for HU).

When compared to historical first reader performance, the AI system showed an absolute difference of 1.7% (0.1%, 3.3%) for sensitivity (including three-year subsequent cancers) and -4.8% (-5.1%, -4.6%) for specificity. With relative differences of 2.5% (0.4%, 4.6%) on sensitivity and -5.0% (-5.3%, -4.7%) on specificity, the AI passed the superiority test on sensitivity and non-inferiority on specificity, well within the defined 10% margin for non-inferiority.

The AI system flagged 143 of the 482 (29.7%) historically not detected three-year subsequent cancers, and of the 373 historical ICs (three-year ICs in the UK and two-year ICs in HU) the AI system found 111 (29.8%).

Using the one-year sample, where more complete IC data is available, the superiority on sensitivity and non-inferiority on specificity held, with the AI system flagging 46 of the 128 (35.9%) historically not detected three-year subsequent cancers, which would have led to a 17.1% relative reduction of missed cancers.

## Performance in the double reading workflow

The performance of double-reading with AI was estimated using a simulation (see Methods). The statistical tests show that double reading with the AI system compared to historical double reading was at least non-inferior at every metric, with superiority tested and passed for recall rate, specificity and PPV (table 3).

Regional analyses for UK and HU show that at least non-inferiority held for all metrics at both regions well within the 10% margin, with superiority passed for specificity in the UK and superiority passed for RR, PPV and specificity in HU (table 3).

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

**Table 3: Performance of double reading with and without AI**

| **A) Results pooled across regions on the ten-year and one-year samples** | | | |
|---|---|---|---|
| **Performance Metric** | **Historical double reading (%)** | **Double reading (DR) with AI (%)** | **Test outcome for DR with AI[1]** |
| On ten-year sample: with incomplete IC data available | | | |
| Recall rate | 5.2 (5.1, 5.3) | 4.8 (4.7, 4.9) | **Superior** |
| CDR[2] | 8.6 (8.4, 8.7) | 8.4 (8.2, 8.5) | **Non-inferior** |
| Sensitivity[2] | 84.2 (82.9, 85.6) | 82.4 (81.0, 83.8) | **Non-inferior** |
| Specificity | 96.5 (96.3, 96.6) | 96.8 (96.7, 96.9) | **Superior** |
| PPV[2,4] | 20.5 (20.1, 20.8) | 20.4 (20.1, 20.8) | **Superior** |
| On one-year sample: with more complete IC data available | | | |
| Recall rate | 4.8 (4.6, 5.0) | 4.5 (4.3, 4.6) | **Superior** |
| CDR[2] | 8.1 (7.7, 8.5) | 8.0 (7.6, 8.4) | **Non-inferior** |
| Sensitivity[2] | 76.1 (72.4, 79.8) | 75.1 (71.3, 78.8) | **Non-inferior** |
| Specificity | 97.0 (96.6, 97.3) | 97.3 (97.0, 97.6) | **Superior** |
| PPV[2,4] | 20.3 (19.3, 21.3) | 20.7 (19.6, 21.8) | **Superior** |
| **B) Regional breakdown on the ten-year sample** | | | |
| **Performance metric** | **Historical double reading (%)** | **Double reading (DR) with AI (%)** | **Test outcome for DR with AI[1]** |
| Regional breakdown for UK | | | |
| Recall rate | 3.8 (3.8, 3.9) | 3.8 (3.7, 3.9) | **Non-inferior** |
| CDR (3Y)[2] | 8.8 (8.6, 9.0) | 8.6 (8.4, 8.7) | **Non-inferior** |
| Sensitivity (3Y)[2] | 86.1 (84.5, 87.6) | 83.9 (82.3, 85.6) | **Non-inferior** |
| Specificity | 97.1 (96.9, 97.2) | 97.1 (97.0, 97.3) | **Superior** |
| PPV (3Y)[2,4] | 24.5 (24.0, 25.0) | 24.0 (23.5, 24.4) | **Non-inferior** |
| Regional breakdown for HU | | | |
| Recall rate | 9.2 (9.0, 9.4) | 7.8 (7.7, 8.0) | **Superior** |
| CDR (2Y)[3] | 7.7 (7.1, 8.3) | 7.6 (7.0, 8.2) | **Non-inferior** |
| Sensitivity (2Y)[3] | 88.8 (86.2, 90.9) | 87.5 (84.9, 89.7) | **Non-inferior** |
| Specificity | 95.8 (95.4, 96.1) | 95.4 (95.0, 95.7) | **Superior** |
| PPV (2Y)[3,4] | 8.3 (7.7, 9.0) | 9.6 (8.9, 10.4) | **Superior** |

95% confidence intervals are presented in parentheses.

1. All test outcomes were based on the relative difference with a two-sided 95% CI. A 10% margin was used for non-inferiority testing (see Statistical Methods for details).
2. The positive pool for CDR, sensitivity, and PPV include screen-detected positives and 'three-year subsequent cancers', which are the standard three-year ICs for the UK.
3. The positive pool for CDR, sensitivity, and PPV include screen-detected positives and two-year ICs only, which are relevant for HU.
4. Due to the definition of PPV being over all cases recalled, the figures here represent a lower bound of PPV.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

## Performance comparison of pathological features

Table 4 presents stratifications by pathological features for positive cases, characterising the spectrum of cancers detected by double reading with and without AI, with the maximum absolute percentage difference being 0.7%.

**Table 4: Pathological features of positive cases recalled in double reading with and without AI.**

| Feature | Positive cases from the ten-year sample | | | |
|---|---|---|---|---|
| | Recalled by historical double reading | | Recalled by double reading with AI | |
| | Number of cases | Proportion of positives | Number of cases | Proportion of positives |
| Histological type | | | | |
| Invasive | 1770 | 75.6% | 1742 | 76.0% |
| In-situ | 345 | 14.7% | 327 | 14.3% |
| Unknown[1] | 226 | 9.7% | 222 | 9.7% |
| Pathological size (invasive only) | | | | |
| <=10 mm | 480 | 27.1% | 460 | 26.4% |
| >10 mm | 754 | 42.6% | 750 | 43.1% |
| Unknown[1] | 536 | 30.3% | 532 | 30.5% |
| Lymph node status (invasive only) | | | | |
| Positive | 364 | 20.6% | 363 | 20.8% |
| Negative | 1263 | 71.4% | 1238 | 71.1% |
| Unknown[1] | 143 | 8.1% | 141 | 8.1% |
| Histology grade (invasive only) | | | | |
| 1 | 439 | 24.8% | 428 | 24.6% |
| 2 | 937 | 52.9% | 924 | 53.0% |
| 3 | 333 | 18.8% | 330 | 18.9% |
| Unknown[1] | 61 | 3.4% | 60 | 3.4% |

All positives, i.e. screen-detected cancers and 'three-year subsequent cancers' are included.

1. At UK sites, 0.42% of histological type of screen-detected positives were unknown or unavailable.

## Operational performance

When used as an independent reader in a double-reading workflow, the AI system automates the second read. This workflow reduction was offset by an increased proportion of cases requiring arbitration from 3.3% (3.2%, 3.3%) to 12.3% (12.2%, 12.5%) when using the AI system as an independent reader. These results suggest that 251,014 (44.8%) less case assessments would have been required by human readers for the study period by using the AI system.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

# Discussion

To achieve high cancer detection rates whilst maintaining low recall rates, many European countries rely on double reading, further exacerbating workforce pressures. An AI system that can serve as a robust and reliable independent reader in breast cancer screening addresses both clinical and socio-economic needs, and helps to make high quality care more widely available.

The aim of this large-scale, retrospective observational study was to evaluate the first commercially available AI system for use as an independent reader in the double reading breast cancer screening workflow.

Double reading performance with the AI compared to historical double reading showed superior recall rate (4.8% vs 5.2%) and specificity (96.8% vs 96.5%) and non-inferior cancer detection rate (8.4 vs 8.6 per thousand) and sensitivity (82.4% vs 84.2%). It is worth highlighting that the AI system performed particularly well in detecting ICs, and the comparative cancer detection performance improved when more complete IC data was available. Under the assumptions of the expected IC proportions (see Appendix 2), the AI system's sensitivity is likely to increase in real-world deployment. Importantly, the spectrum of cancers detected in double reading with AI did not change from historical screening results, indicating that the use of AI does not require downstream changes to the existing clinical pathway. Saving the workload of second reading resulted in an estimated 45% reduction of the total workload, while accounting for an increased arbitration rate (12.3% vs 3.3%). Such a workload reduction would significantly reduce the pressure on health services.

When assessed on its own, the AI system showed an absolute 1.7% to 5.1% improvement on sensitivity and found 30% to 36% of historical ICs, indicating that cancer detection could be significantly improved with the AI system. The specificity of the AI system was non-inferior to the historical first human reader but lower, which contributed to the increased arbitration in double reading.

Past studies have compared the performance of AI systems to individual human readers (19-23). Some employed small-scale reader studies (19-21) with enriched samples of 320 to 720 cases, and larger retrospective evaluations (22-23) with 8,805 to 28,853 cases. While reported performances in the small reader studies are encouraging, it remains unclear whether the results on enriched test sets and samples generalise to real-world screening populations, and only Kim et al (20) evaluated performance on multiple vendors. The larger retrospective studies (22-23) provide a more reliable comparison, including information on the impact on double reading. McKinney et al (22) demonstrated non-inferiority on both sensitivity and specificity when simulating double reading with an AI system, while Salim et al (23) showed an AI system paired with a single human reader (without arbitration) detects more cancers than two human readers at the cost of significantly higher recall rates. 95% to 100% of cases in both evaluations came from a single hardware vendor, and Salim et al (23) required resampling to approximate a screening population.

The major strength of this study is that the AI system was evaluated directly for double reading on a diverse, heterogeneous, large-scale and representative screening population with data collected across two national screening programmes with a variety of demographic differences. The authors believe this is the first large-scale study that does not rely on informed sampling to approximate a screening cohort. This is significant as resampling can introduce unwanted biases and is not guaranteed to faithfully represent a target population. The historical reader results represented the practical standard of care, with no influence on reader behavior resulting from participation in the study and no enrichment for positives or any subgroups.

The retrospective nature of the evaluation means a number of limitations exist. In the simulation, the historical second reader opinion was used as the arbitrator when the historical arbitration opinion was unavailable. This is an approximation as the second reader and arbitrator perform different tasks, with different expected performance. Also, the ten-year span of the cases did not contain complete IC data, and while the one-year sample was closer to expectations, this came at the expense of cohort size. Estimating the extent of their

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

impact will be the subject of future work, along with further studies to assess performance in subgroups (e.g. different ethnicities).

While this study demonstrated efficacy in sites already employing double reading, the results suggest the performance standards of double reading could be achieved in programmes currently employing single reading, with a fraction of the resources traditionally required. Countries with single reading as the standard may significantly improve the standard of care with the use of AI, leading to potentially better patient outcomes due to fewer missed cancers.

The results demonstrate that the evaluated AI system can be an effective solution acting as an independent reader in the double reading workflow. Standard of care is at least preserved on all relevant screening metrics, for both standalone and simulated double reading comparisons. The scale and diversity of samples support that the findings are generalisable to many screening programmes and the use of practical metrics ensures that the impact of introducing AI into everyday screening is reliably estimated and of clinical relevance.

Reducing the overall double reading workload by 45% can enable staff redeployment and service improvements such as increased patient interaction, more time for training, an extended programme age range, more focus on complex cases and, during a time of workforce crisis, supporting the sustainability of breast cancer screening.

## Acknowledgements

The UK arm of the study was supported by funding from Innovate UK via an NHS England and Improvement, Office of Life Sciences (OLS) Wave 2 Test Bed Programme and a Medical Research Council (MRC) Biomedical Catalyst award.

We thank M. Bidlek, K. Borbély, G. Di Leo, R. Fülöp, K. Giese, F. Gilbert, T. Helbich, S. Hofvind, B. Joe, K. Keresztes, E. Kovács, M. Milics, Z. Pentek, É. Szabó, L. Tabar, T. Tasnádi, C. Yau for expert input and guidance.

We thank D. Dinneen, S. Kerruish, G. Mehmert, D. Pribil and F. van Beers for technical and management support.

We thank the staff at MaMMa Egészségügyi Zrt., the EMRAD Imaging Network, Nottingham University Hospital Trust and Breast Screening Programme (BSP), The Leeds Teaching Hospital and Leeds/Wakefield BSP, and United Lincolnshire Hospital Trusts and BSP.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

# References


1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2018;68(6):394-424.
2. Tabar L, Yen M, Vitak B, et al. Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. The Lancet. 2003;361(9367):1405-10.
3. Duffy SW, Tabár L, Yen AM, et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. Cancer. 2020;126(13):2971-9.
4. Lauby-Secretan B, Scoccianti C, Loomis D, et al. Breast-Cancer Screening — Viewpoint of the IARC Working Group. N Engl J Med. 2015;372(24):2353-8.
5. Tabár L, Dean PB. Recommendations for breast cancer screening. The Lancet Oncology. 2020;21(11):e511.
6. Zielonke N, Kregting LM, Heijnsdijk EAM, et al. The potential of breast cancer screening in Europe. Int J Cancer. 2021;148(2):406-18.
7. Hoff SR, Myklebust T, Lee CI, et al. Influence of Mammography Volume on Radiologists' Performance: Results from BreastScreen Norway. Radiology. 2019;292(2):289-96.
8. Harvey SC, Geller B, Oppenheimer RG, et al. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. Am J Roentgenol 2003;180:1461-1467.
9. Ciatto S, Ambrogetti D, Bonardi R, et al. Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme. J Med Screen 2005;12:103-106.
10. Blanks RG, Wallis MG, Moss SM. A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme. J Med Screen 1998;5:195-201.
11. Perry N, Broeders M, de Wolf C, et al. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. Annals of Oncology. 2008;19(4):614-22.
12. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ. 2017;10(11);359:j4683.
13. National Health Institutes England, Public Health England, British Society of Breast Radiology, Royal College of Radiologists. The Breast Imaging and Diagnostic Workforce in the United Kingdom. 2017. https://www.rcr.ac.uk/publication/breast-imaging-and-diagnostic-workforce-united-kingdom. Accessed December 28, 2018.
14. Gulland A. Staff shortages are putting UK breast cancer screening "at risk," survey finds. BMJ. 2016;10.1136:i2350.
15. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. JAMA Intern Med. 2015;175(11):1828-37.
16. Keen JD, Keen JM, Keen JE. Utilization of Computer-Aided Detection for digital screening mammography in the United States, 2008-2016. J Am Coll Radiol 2018;15:44-48.
17. Lehman CD, Wellman RD, Buise DS et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med. 2015;175:1828-1837.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252537; this version posted March 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

18. Gilbert FJ, Astley SM, Gillan MGC et al. Single reading with computer-aided detection of screening mammography. N Engl J Med 2008;359:1675-1684.

19. Wu N, Phang J, Park J, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans Med Imaging. 2020;39(4):1184-94.

20. Kim H, Kim HH, Han B, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. The Lancet Digital Health. 2020;2(3):e138-e148.

21. Lotter W, Diab AR, Haslam B, et al. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat Med. 2021;27(2):244-9.

22. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020 Jan 2;577(7788):89-94.

23. 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.

24. AgeX trial, University of Oxford trial protocol, September 2020. Protocol available to download from https://www.agexuk/links/