Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study

*Karin Dembrower, Alessio Crippa, Eugenia Colón, Martin Eklund, Fredrik Strand, and the ScreenTrustCAD Trial Consortium**

Summary

Background Artificial intelligence (AI) as an independent reader of screening mammograms has shown promise, but there are few prospective studies. Our aim was to conduct a prospective clinical trial to examine how AI affects cancer detection and false positive findings in a real-world setting.

Methods ScreenTrustCAD was a prospective, population-based, paired-reader, non-inferiority study done at the Capio Sankt Göran Hospital in Stockholm, Sweden. Consecutive women without breast implants aged 40–74 years participating in population-based screening in the geographical uptake area of the study hospital were included. The primary outcome was screen-detected breast cancer within 3 months of mammography, and the primary analysis was to assess non-inferiority (non-inferiority margin of 0·15 relative reduction in breast cancer diagnoses) of double reading by one radiologist plus AI compared with standard-of-care double reading by two radiologists. We also assessed single reading by AI alone and triple reading by two radiologists plus AI compared with standard-of-care double reading by two radiologists. This study is registered with ClinicalTrials.gov, NCT04778670.

Findings From April 1, 2021, to June 9, 2022, 58 344 women aged 40–74 years underwent regular mammography screening, of whom 55 581 were included in the study. 269 (0·5%) women were diagnosed with screen-detected breast cancer based on an initial positive read: double reading by one radiologist plus AI was non-inferior for cancer detection compared with double reading by two radiologists (261 [0·5%] *vs* 250 [0·4%] detected cases; relative proportion 1·04 [95% CI 1·00–1·09]). Single reading by AI (246 [0·4%] *vs* 250 [0·4%] detected cases; relative proportion 0·98 [0·93–1·04]) and triple reading by two radiologists plus AI (269 [0·5%] *vs* 250 [0·4%] detected cases; relative proportion 1·08 [1·04–1·11]) were also non-inferior to double reading by two radiologists.

Interpretation Replacing one radiologist with AI for independent reading of screening mammograms resulted in a 4% higher non-inferior cancer detection rate compared with radiologist double reading. Our study suggests that AI in the study setting has potential for controlled implementation, which would include risk management and real-world follow-up of performance.

Funding Swedish Research Council, Swedish Cancer Society, Region Stockholm, and Lunit.

Lancet Digit Health 2023

Published **Online**
September 8, 2023
https://doi.org/10.1016/S2589-7500(23)00153-X

See **Online/Comment**
https://doi.org/10.1016/S2589-7500(23)00176-0

*Members are listed in the appendix (p 3)

Breast Imaging Unit, Department of Radiology, Capio Sankt Göran Hospital, Sankt Göransplan, Stockholm, Sweden (K Dembrower MD); **Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden** (K Dembrower, F Strand MD); **Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden** (A Crippa PhD, Prof M Eklund PhD); **Department of Pathology, Unilabs, Capio Sankt Göran Hospital, Sankt Göransplan, Stockholm, Sweden** (E Colón MD); **Breast Radiology Unit, Medical Diagnostics Karolinska, Karolinska University Hospital, Stockholm, Sweden** (F Strand)

Correspondence to:
Dr Karin Dembrower, Breast Imaging Unit, Department of Radiology, Capio Sankt Göran Hospital, Sankt Göransplan, Stockholm 112 81, Sweden
karin.dembrower@ki.se

See **Online** for appendix

Introduction

Mammography screening has been a cornerstone of early detection of breast cancer since the 1980s. Among its challenges is a marked variability between radiologists in diagnostic accuracy, which leads to unnecessary recalls and missed cancer.[1] Additionally, there is a global shortage of breast radiologists that is exacerbated by increasing demands for precision diagnostics from both providers and patients.[2,3] Artificial intelligence (AI) has the potential to address these challenges.

Multiple retrospective studies suggest that AI has sufficient diagnostic accuracy to make radiological reads as an independent reader of screening mammograms.[4–10] Although these studies provided encouraging results, the retrospective study designs have high risk for biases and do not assess the integration of AI in existing screening workflows.[11,12] We aimed to prospectively assess whether double reading of mammography images by AI and one radiologist can achieve non-inferior cancer detection compared with double reading by two radiologists.[13–15]

Methods

Study design and participants

ScreenTrustCAD was a prospective, population-based, paired-reader, non-inferiority study that compared subsequent breast cancer detection based on initial positive read by any of two blinded radiologists and independent AI. The fully blinded, three-reader paired design, with two radiologists and AI, allowed us to examine the downstream effect of any combination of initial readers:[13,14] double reading by two radiologists, double reading by one radiologist plus AI, single reading by AI, and triple reading by two radiologists plus AI.[13,14]

Research in context

Evidence before this study

For 30 years, population-based screening mammography has been a cornerstone in reducing breast cancer mortality. However, mammography screening is associated with several challenges, including a severe shortage of breast radiologists and inconsistent reads between radiologists. Several retrospective studies have indicated that artificial intelligence (AI) can perform at a similar diagnostic accuracy as radiologists in identifying mammograms that show signs of breast cancer in the initial independent read. However, the retrospective design of these studies does not allow assessment of what happens with signs of breast cancer identified by AI in subsequent diagnostic steps—the consensus discussion and clinical investigation. Our systematic search for studies published from Jan 1, 2010, to Feb 1, 2021, in ClinicalTrials.gov (condition: breast cancer; other terms: artificial intelligence screening; study type: interventional studies; study start: before Feb 1, 2021) and PubMed ("artificial intelligence"[MeSH Major Topic]) AND ("mammography"[MeSH Major Topic]) AND ("prospective studies"[MeSH Terms]) with no language restrictions, did not find any started trials or published results describing the implementation of AI in a population-based prospective study of screening mammography. The absence of prospective studies is a barrier to widespread adoption of AI in breast cancer screening programmes.

Added value of this study

This prospective study reports that the subsequent breast cancer detection rate was non-inferior for initial reading performed by one radiologist plus AI compared with double reading by two radiologists. In addition, the detection rate was non-inferior for single reading by AI compared with double reading by two radiologists.

Implications of all the available evidence

This study provides prospective evidence that using AI for the initial reading of screening mammograms increased breast cancer detection rates when AI was implemented in an actual screening workflow. Our findings validate that the results in previous retrospective studies translate to increased cancer detection in a real-world setting.

The trial was done at the Capio Sankt Göran Hospital in Stockholm, Sweden. The study population consisted of women aged 40–74 years living in the hospital's geographical uptake area (Western Stockholm County and part of the inner city) who attended regular mammography screening. For the purpose of screening, sex data were defined by Swedish personal identity number. Women with breast implants were excluded, as the AI software had not been validated for that subgroup. Three groups of women not attending regular screening were also excluded: those who carried a known genetic mutation (*BRCA 1, BRCA 2, PTEN, TP53, STK 11,* and *CDH1*), those assessed by the hereditary cancer clinic as having very high lifetime risk (who participate in special surveillance programmes), and women with a personal history of breast cancer. Additional design details are available in the appendix (p 4). Except for a 3-week pause in the invitations for screening in May, 2020, there was no effect on the primary study endpoint due to the COVID-19 pandemic.

The trial was designed by the authors, and data were collected by trial consortium members (appendix p 3). The authors assume responsibility for the accuracy and completeness of the data and for the fidelity of the trial to the protocol. The study protocol was approved by the ethical review authority of Sweden, which waived the need for individual informed consent.

For the **study protocol** see https://storage.googleapis.com/ctgov2-large-docs/70/NCT04778670/Prot_SAP_002.pdf

Procedures

The standard-of-care radiological workflow in place at Capio Sankt Göran Hospital follows the Swedish National Guidelines for mammography screening. Briefly, craniocaudal and mediolateral oblique mammographic views of each breast are acquired using Philips Microdose SI Universal equipment (Philips, Eindhoven, Netherlands). In the first stage of the workflow, resulting images are independently assessed by two radiologists, each of whom are blinded to the other's read. Following Swedish practice, if a sign leading to suspicion of cancer is seen, the examination is assessed as abnormal; if no suspicious sign is seen, the examination is assessed as normal. If both readers assess the examination as normal, notification of the negative result is sent to the screening participant. If at least one read is abnormal, the case proceeds to a consensus discussion, in which two radiologists discuss the images, deciding whether or not to recall for further investigation. The selection of which radiologists should make the initial read and which radiologists should perform the consensus discussion is based on who is available at the time and there is no selection rule. Recalled patients are examined by additional imaging (eg, special mammography views, tomosynthesis, and ultrasonography). If suspicion of cancer remains, a biopsy sample is acquired; otherwise, the patient is notified of a negative result. Finally, biopsies are analysed by a pathologist who makes a definitive diagnosis of breast cancer (screen-detected) or benign tissue.

All 11 breast radiologists at Capio Sankt Göran Hospital, who had a median 17 years (range 5–32) of experience, participated in the study. One of the radiologists was KD. The existing standard-of-care regarding assigning readers to screening mammograms was not altered due to this study. As before the study, readers at the hospital were assigned the next unread examination without any specific pairing strategy for reader one and two. To explain the meaning of the AI scores and image markings, the investigators held an initial workshop for the radiologists. After that, on-the-job training was provided by KD.

For the purposes of the study, AI was implemented as an independent reader, running in the background, at the first stage of the radiological workflow. Radiologists were blinded to the reads of the other radiologist and of AI. If any of the three readers made an abnormal read, the examination proceeded to the consensus discussion (appendix p 5). The two radiologists in the consensus discussion were not necessarily the same as in the preceding independent read. In the consensus discussion, radiologists had full access to all AI information for all cases: the examination-level AI score, and for any localised image finding, a graphical outline and the corresponding AI abnormality score (example shown in appendix p 6).

We evaluated four strategies for the initial mammogram reading and examined the actual downstream diagnostic outcomes for each: (1) double reading by two radiologists (standard of care); (2) double reading by one radiologist plus AI; (3) single reading by AI; and (4) triple-reading by two radiologists plus AI. For the strategy of double reading by one radiologist and AI, we consistently included the read of the first (initial) reader, as per study protocol.

For the reading by AI, the Insight MMG AI system (version 1.1.6; Lunit, Seoul, South Korea)[4] was used (appendix p 4). In an independent comparison (performed in 2020 by KD, ME, and FS) of three commercial AI systems for cancer detection on images from Karolinska University Hospital in Sweden,[10] Insight MMG was the top performer and was consequently chosen for prospective evaluation in ScreenTrustCAD. For each image, the AI system generated a continuous score related to the estimated degree of abnormality. The examination-level score was the highest of the image scores. The abnormality threshold was calibrated using an enriched retrospective dataset from Dembrower and colleagues[15] containing 6625 mammography examinations acquired during 2012–15 on Philips mammography equipment at two Stockholm hospitals (Capio Sankt Göran Hospital and Southern General Hospital). The threshold examination level score of 53·4 for the binary AI system decision was determined by the level at which double reading by the AI system plus one radiologist achieved a 2% higher cancer detection rate compared with double reading by two radiologists (appendix p 7). The threshold value was predefined before the first patient was assessed and remained unchanged. Images from these two hospitals were not used in the development of the AI system.

Radiologist and AI reads were recorded in the breast radiology system along with tumour measurements (if applicable) and the presence or absence of microcalcifications. Biopsy results were recorded in the electronic health records.

Outcomes

The primary outcome was diagnosis with screen-detected breast cancer, invasive or ductal in situ (or both), within 3 months of undergoing mammography. Secondary outcomes were the number of examinations that ultimately were not positive for cancer at each stage of radiological read (independent reading [abnormal or normal interpretation], consensus discussion [recall or not], and continued investigation [biopsy or not]); reader flagging for consensus discussion; recall decision by consensus; biopsy acquired; and process failures in generating the AI score or in transferring the AI score to the radiological information system.

All prespecified outcomes were assessed in the intention-to-treat population, defined as all consecutive participants regardless of missing reads by AI or a radiologist.

Statistical analysis

The sample size calculation followed the methods for paired screen-positive designs described by Alonzo and colleagues.[16] We assumed a 0·5% prevalence of breast cancer in the screening population, a true positive fraction of 0·70 for a screening mammogram, and a relative true positive fraction of double reading using AI and one radiologist compared with two radiologists of 1·02. The non-inferiority margin was set to 15% relative reduction with a one-sided α of 0·025. The non-inferiority margin was based on the finding in Salim and colleagues[1] that the least sensitive quartile of radiologists had 15% lower sensitivity compared with the most sensitive quartile, and was agreed at a consensus group meeting that included mammography radiologists and statisticians (KD, ME, and FS). Furthermore, we assumed that all patients recommended for biopsy would be compliant. Under these assumptions, we estimated a power of 87% to show non-inferiority with 55 000 participants.

In accordance with standard methodology for screening and diagnostic trials, ScreenTrustCAD followed a screen-positive design, meaning that only participants who tested positive were, after additional investigation, subject to disease verification by biopsy. A consequence of all screen-positive designs is that absolute measures of sensitivity and specificity cannot be estimated because the disease status for participants who screen negative is not verified (appendix pp 73–74). However, comparisons of relative measures of cancer detection rate of different reader combinations can be estimated under a screen-positive design to answer the primary research question of whether double reading with AI and one radiologist finds a non-inferior number of breast cancers as double reading with two radiologists in screening mammography. This design can also answer secondary research questions about number of consensus discussions, recalled participants, and participants with biopsy samples taken under each reader combination. Differences in the detection of breast cancer using double

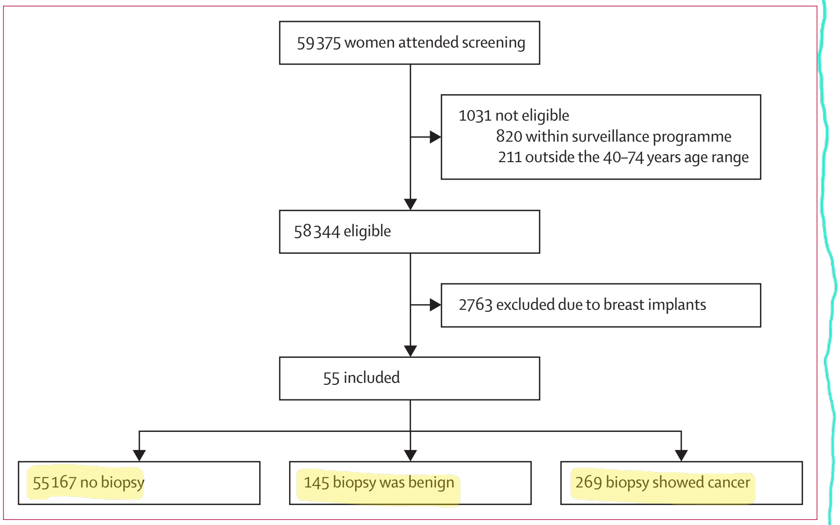

59 375 women attended screening
1031 not eligible
820 within surveillance programme
211 outside the 40–74 years age range
58 344 eligible
2763 excluded due to breast implants
55 included
55 167 no biopsy
145 biopsy was benign
269 biopsy showed cancer

Figure 1: **Study population selection**
Women who attended screening but were part of a special surveillance programme were excluded, as were women who were not in the age range of population-based invitations. Among the remaining eligible women, mammography examinations that were marked as containing breast implants were excluded.

reading by two radiologists versus double reading by one radiologist plus AI were consequently assessed by the relative cancer detection rate, hereafter referred to as the relative true positive fraction, and computed by dividing the number of cancers detected by an experimental reader combination (AI plus one radiologist, AI alone, or triple reading with AI plus two radiologists) by the number of cancers detected by the standard-of-care reader combination (two radiologists). Differences in false positive results at each stage of radiological workflow (independent read, consensus discussion, and biopsy) were assessed by the corresponding relative false positive fractions (computed by dividing the number of participants referred to each stage of the radiological workflow and subsequently declared to have a negative breast cancer screening result by an experimental reader combination by the number of participants referred to each stage of the radiological workflow and subsequently declared to have a negative breast cancer screening result by the standard-of-care reader combination). The 95% CI was calculated by exponentiating the normal-based CI limits on the log scale, using asymptotic standard errors. We report one-sided p values for non-inferiority for breast cancer (non-inferiority margin of 0·15 relative reduction and an α of 0·025) and two-sided p values for superiority (α of 0·05). We compared the standard-of-care two radiologists double reading with the alternative in terms of relative proportion of abnormal interpretations, recall decisions, and biopsy acquisitions. Positive rates were calculated by dividing the number of positive results at each workflow stage by the total number of mammography examinations in the study. Per the statistical analysis plan, missing information for radiologist two was replaced by the AI read if the missingness was less than 1% of the examinations. This was selected to follow the intention-to-treat design and be conservative in terms of assessing non-inferiority.

Prespecified subgroup analyses were performed to examine results in different age groups, mammographic density categories (determined by the AI system, mimicking the four categories of the Breast Imaging Reporting and Data System by the American College of Radiology), and cancer characteristics including invasiveness, presence of axillary lymph node metastasis, categories of tumour size, and molecular subtypes. We also performed a post-hoc analysis stratified by calendar period (quarters). According to Swedish practice, women who report a new lump at the time of screening and are diagnosed with cancer count as having screen-detected cancer. However, practices in other countries might vary, and therefore we performed a prespecified sensitivity analysis for each reader strategy in which all participants who reported a lump to the radiographer at the time of the examination would be excluded from the screening population.

The comparative numbers of women undergoing screening, consensus discussion, biopsies, and being diagnosed with cancers by each reader strategy were normalised to a population of 100 000 screened women. Statistical analyses were done using R (version 4.2.0). All the analyses were prespecified (unless explicitly stated as post hoc) and performed according to the statistical analysis plan (appendix pp 65–85). In accordance with the statistical analysis plan, reported p values and CIs have not been adjusted for multiplicity and should be interpreted with caution. There were no major changes to the study protocol or statistical analysis plan after the start of inclusion on April 1, 2021. Minor revisions to the study protocol and statistical analysis plan were made blinded to the data and results. The complete revision history of the protocol and statistical analysis plan are in the appendix (pp 86–88). This study is registered at ClinicalTrials.gov, NCT04778670.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

From April 1, 2021, to June 9, 2022, 58 344 women aged 40–74 years underwent regular mammography screening, of whom 55 581 were included in the study (figure 1, table 1). Median age was 55 years (IQR 46–65). All participants had AI results reported. The reading of a second radiologist (radiologist two) was missing for 124 (0·2%) examinations, less than 1% of the total number, and so were replaced with the AI read per the statistical analysis plan. 6002 (10·8%) women had an examination assessed as abnormal by reader one, reader two, or AI and were subject to consensus discussion (table 1). Following consensus discussions, 1716 (3·1%) were

	Overall population (n=55 581)	Participants diagnosed with cancer (n=269)
Examination date		
2021: quarter 2	12 426 (22·4%)	56 (20·8%)
2021: quarter 3	8056 (14·5%)	32 (11·9%)
2021: quarter 4	12 609 (22·7%)	73 (27·1%)
2022: quarter 1	12 833 (23·1%)	66 (24·5%)
2022: quarter 2	9657 (17·4%)	42 (15·6%)
Median age, years	55 (46–65)	63 (55–69)
Age category, years		
40–49	18 308 (32·9%)	41 (15·2%)
50–59	15 883 (28·6%)	55 (20·4%)
60–69	14 756 (26·5%)	113 (42·0%)
70–79	6634 (11·9%)	60 (22·3%)
Independent assessment		
Normal interpretation	49 579 (89·2%)	NA
Abnormal interpretation by at least one reader	6002 (10·8%)	NA
Consensus discussion		
Not subject to consensus discussion	49 579 (89·2%)	NA
No recall	4286 (7·7%)	NA
Recall	1716 (3·1%)	NA
Investigation		
Not subject to investigation	53 865 (96·9%)	NA
No biopsy acquired	1302 (2·3%)	NA
Biopsy acquired	414 (0·7%)	NA
Biopsy		
Not subject to biopsy	55 167 (99·3%)	NA
No breast cancer	145 (0·3%)	NA
Breast cancer	269 (0·5%)	NA
Invasiveness		
In situ	NA	63 (23·4%)
Invasive	NA	200 (74·3%)
Axillary lymph node metastasis		
Negative	NA	241 (89·6%)
Positive	NA	28 (10·4%)
Median tumour size, mm	NA	17 (11–30)
Tumour size category, invasive and in situ		
1–20 mm	NA	167 (62·1%)
21–50 mm	NA	76 (28·3%)
≥51 mm	NA	26 (9·7%)
Molecular subtype (n=169)		
Luminal A-like	NA	112 (66·3%)
Luminal B-like	NA	28 (16·6%)
HER2-overexpressing (luminal and non-luminal)	NA	14 (8·3%)
Basal (triple negative)	NA	15 (8·9%)
Nottingham histological grade: invasive (n=185)		
Grade 1	NA	41 (22·1%)
Grade 2	NA	120 (64·9%)
Grade 3	NA	24 (13·0%)
Nottingham histological grade: in situ (n=61)		
Grade 1	NA	5 (8·2%)
Grade 2	NA	22 (36·0%)
Grade 3	NA	34 (55·7%)

(Table 1 continues in next column)

(Continued from previous column)

Data are n (%) or median (IQR). NA=not applicable.

Table 1: **Study population characteristic**

recalled for further investigation, of whom 414 (0·7%) had a biopsy sample taken and 269 (0·5%) were diagnosed with screen-detected breast cancer. 63 (23·4%) participants had ductal cancer in situ and 200 (74·3%) had invasive cancers; 28 (10·4%) women had lymph node metastases. Median tumour size was 17 mm (IQR 11–30). 112 (66·3%) of 169 had luminal A molecular subtype, 28 (16·6%) luminal B, 14 (8·3%) HER2-overexpressing, and 15 (8·9%) basal.

250 (0·45%) women had breast cancer detected by double reading with two radiologists compared with 261 (0·47%) detected by double reading with AI plus one radiologist, a relative proportion of 1·04 (95% CI 1·00–1·09; p<0·0001), showing non-inferiority (table 2, figure 2). Furthermore, because the lower boundary of the two-sided 95% CI was greater than 1, the experimental strategy of double reading with one radiologist and AI was deemed superior to double reading by two radiologists (p=0·017). Single reading by AI detected breast cancer in 246 (0·44%) participants with a relative proportion of 0·98 (95% CI 0·93–1·04) compared with double reading by two radiologists. Single reading by AI was deemed non-inferior (p<0·0001), but not superior (p=0·73), to double reading with two radiologists. Triple reading by two radiologists and AI resulted in detecting breast cancer in 269 (0·48%) participants with a relative proportion of 1·08 (95% CI 1·04–1·11) compared with double reading by two radiologists. Triple reading by two radiologists and AI was deemed superior to double reading with two radiologists (p<0·0001).

Our prespecified subgroup analysis of potential heterogeneity between reader strategies in terms of age, mammographic density, and cancer characteristics, showed no significant differences (appendix pp 8–9). Further information on the underlying count of concordant and discordant reads for each reader combination are in the appendix (pp 11–12). For diagnosed cancer where AI was the only positive initial reader (n=19), the prompted localisation by AI corresponded, in all cases, to the actual localisation of the biopsy-confirmed cancer.

Relative to the two-radiologist approach, the proportion of abnormal interpretations for patients who were not diagnosed with breast cancer was higher with double reading with one radiologist plus AI and with triple reading with two radiologists plus AI, and was lower with AI alone (table 2). Relative to the two-radiologist approach, the proportion of patients whose examinations went to consensus discussion and were not diagnosed with breast cancer was lower with the double reading with AI plus one radiologist and with AI alone, but was higher for AI plus two radiologists. The proportion of no cancer finding among the patients undergoing biopsies was lower for single reading by AI compared with all other reader combinations.

The sensitivity analysis in which all independent reads for patients reporting a lump at time of screening were assumed to be positive screening examinations mostly affected the relative false positive fraction for the AI-only strategy at the initial independent reading stage, and did not change the proportion for double reading with AI plus a radiologist (appendix p 10).

The comparative numbers of women undergoing screening, consensus discussion, biopsies, and being diagnosed with cancers by each reader strategy, normalised to a population of 100 000 screened women are shown in figure 3. Thanks to the efficacy of the consensus discussion, the elevated abnormal screens in the radiologist and AI strategy did not translate into an increase of recalls.

The post-hoc subgroup analysis of potential heterogeneity between reader strategies in terms of calendar time period showed no significant differences (appendix pp 8–9).

Discussion

With the growing body of evidence from retrospective studies that the use of AI can enhance the performance of mammography screening, several prospective evaluations are underway. These use AI in different capacities, for instance to triage mammograms for more or less scrutiny by radiologists (NCT04949776, NCT04838756, and Larsen and colleagues[17]) or to improve the accuracy of radiologists' reads (NCT05024591).

Here, we evaluated AI as an independent reader of mammograms within an established mammography screening workflow. In this prospective, population-based trial, double reading by one radiologist plus AI resulted in a 4% (11/250) increase in screen-detected cancers. Subgroup analysis showed no marked or significant systematic differences for any cancer or patient characteristic between AI and radiologist detection. The proportion of participants diagnosed with breast cancer was thus similar to what was assumed in the sample size calculations (observed 0·5% versus assumed 0·5%).

Double reading by one radiologist plus AI caused a 21% (868/4104) increase in the number of examinations with abnormal interpretation. This suggests that AI and human readers perceive somewhat different image features as suspicious for cancer, and thus that a human reader and AI provide synergism to increase the sensitivity for detecting breast cancers in mammograms.[18] Subsequent consensus discussions that reviewed mammograms, medical history, and AI information, resulted in a 4% (73/1629) lower recall rate for double reading by one radiologist plus AI compared with double reading by two radiologists. Thus, the consensus discussion was effective in ensuring that the higher abnormal interpretation rate for AI plus one radiologist did not translate into an increased recall rate. The cancer detection rate and abnormal interpretation rate are in line with previous retrospective studies.[4,10] In a screening population of 100 000 women, replacing one radiologist with AI would save 100 000 radiologist reads while

	Double reading by two radiologists	Double reading by AI and one radiologist		Single reading by AI		Triple reading by two radiologists and AI	
	Number of women	Number of Women	Relative proportion* (95% CI)	Number of Women	Relative proportion† (95% CI)	Number of Women	Relative proportion‡ (95% CI)
All screened	55 581 (100%)	55 581 (100%)	NA	55 581 (100%)	NA	55 581 (100%)	NA
Abnormal interpretation	4104 (7·38%)	4972 (8·95%)	1·21 (1·18–1·24)	3162 (5·69%)	0·77 (0·74–0·80)	6002 (10·80%)	1·46 (1·44–1·49)
Abnormal interpretation, no cancer	3854 (6·93%)	4711 (8·48%)	1·22 (1·19–1·25)	2916 (5·25%)	0·76 (0·73–0·79)	5733 (10·31%)	1·49 (1·46–1·51)
Recall after consensus discussion	1629 (2·93%)	1556 (2·80%)	0·96 (0·94–0·97)	861 (1·55%)	0·53 (0·50–0·56)	1716 (3·09%)	1·05 (1·04–1·06)
Recall after consensus discussion, no cancer	1379 (2·48%)	1295 (2·33%)	0·94 (0·92–0·96)	615 (1·11%)	0·45 (0·42–0·48)	1447 (2·60%)	1·05 (1·04–1·06)
Biopsy, all	386 (0·69%)	403 (0·73%)	1·04 (1·01–1·08)	349 (0·63%)	0·90 (0·86–0·95)	414 (0·74%)	1·07 (1·05–1·10)
Biopsy, no cancer	136 (0·24%)	142 (0·26%)	1·04 (0·99–1·10)	103 (0·19%)	0·76 (0·67–0·85)	145 (0·26%)	1·07 (1·02–1·11)
Cancer, all	250 (0·45%)	261 (0·47%)	1·04 (1·00–1·09)	246 (0·44%)	0·98 (0·93–1·04)	269 (0·48%)	1·08 (1·04–1·11)
Cancer, invasive	195 (0·35%)	200 (0·36%)	1·03 (0·98–1·07)	187 (0·34%)	0·96 (0·91–1·01)	206 (0·37%)	1·06 (1·02–1·09)
Cancer, invasive >2 cm or lymph node metastasis	77 (0·14%)	81 (0·15%)	1·05 (0·99–1·12)	78 (0·14%)	1·01 (0·94–1·09)	82 (0·15%)	1·06 (1·01–1·13)

AI=artificial intelligence. *Double reading by two radiologists/double reading by AI and one radiologist. †Single reading by AI/double reading by AI and one radiologist. ‡Triple reading by two radiologists and AI/double reading by AI and one radiologist.

Table 2: **Number of women reaching each stage of the screening workflow and the relative proportion for each experimental reader strategy compared with standard-of-care double reading by two radiologists**

increasing consensus discussions by 1562. Even if the consensus discussions would take five times longer than an independent read, the workload reduction would be considerable.

We also examined AI single reading, a strategy that decreased cancer detection compared with standard of care in a previous retrospective study.[10] Here, it showed non-inferior cancer detection compared with standard of care. Because false positives did not accumulate between two readers, the total number of recalls decreased by 47% (768/1629; n=1382 in a screening population of 100 000 women), implying a major reduction in unnecessary worry for women involved. This reduction in harms would increase the overall value of screening. Furthermore, although radiologists would still be involved in consensus discussions and subsequent diagnostic investigation, AI single reading would result in large workload reductions throughout the screening workflow. A requirement before considering implementation of AI-only reading is that previous mammograms and relevant clinical information, such as current symptoms, previous cancer, and previous investigations, are considered. However, it would mean that a large proportion of mammograms would never be assessed by a board-certified physician. This poses important questions around medical responsibility, public acceptability, radiologist training, and certification of AI systems.[12]

Using the most readers, a triple reading strategy using two radiologists plus AI did result in a slightly higher cancer detection rate than double reading by two radiologists, but nearly 50% more consensus discussions. Compared with the proportion of recalls for double reading by two radiologists, triple reading increased recalls by 5% (relative proportion 1·05), whereas double-reading by one radiologist plus AI decreased recalls by 4% (relative proportion 0·96). The additional cost in terms of workload for radiologists and worry for women must be weighed against the incremental increase in cancer detection.

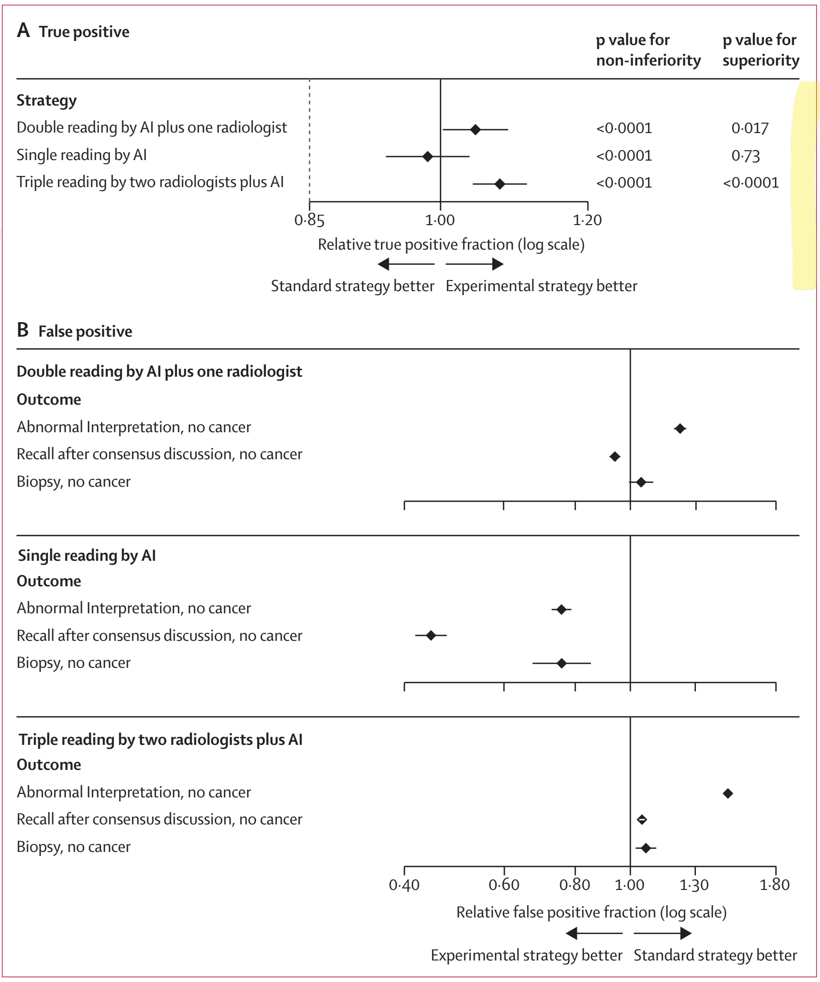
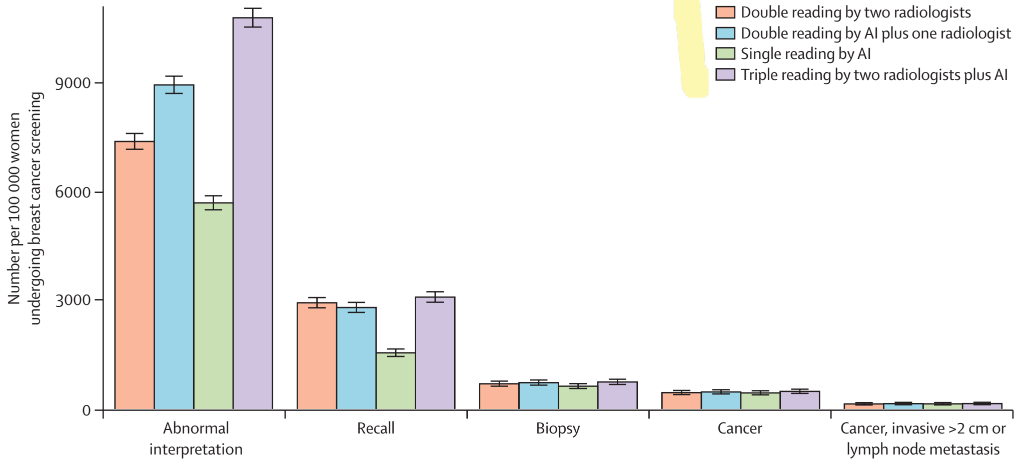


Figure 2: **Relative true and false positive fractions**
(A) The relative true positive fraction (ie, the number of abnormal assessments for cancer examinations by each strategy divided by the number of true abnormal assessments by standard-of-care double reading by two radiologists). (B) The relative false positive fraction (ie, the number of abnormal assessments for women not diagnosed with breast cancer by each strategy divided by the number by standard-of-care double reading by two radiologists). The false positive fraction is reported for each stage of the screening workflow: at the first independent assessment stage (all examinations), at the second consensus discussion stage (the examinations with abnormal independent assessment by any reader), and at the further investigation stage (the women who received a recall decision in the consensus discussion). The x axes are on the log scale. AI=artificial intelligence.

Among the strengths of our study are the full integration of AI as an independent reader into an existing screening workflow affecting which women go on to consensus discussion and, ultimately, who are diagnosed with cancer. Another strength is that the radiologists with whom the AI system is compared had extensive experience (median 17 years). The paired design, where any positive examination at initial independent reading advanced in the workflow, established a flexible analytical framework that enabled assessment of different reader strategies without risk for confounding based on downstream diagnostic outcomes. However, the information available in the consensus discussion differs from standard of care without AI because AI results were available in the consensus discussion. Because the proportion of participants flagged by AI who were recalled after the consensus discussion was lower than for standard of care, the results from ScreenTrustCAD suggest that consensus readers observing an examination with an initial read that is positive by AI but negative by the two radiologists, are nudged towards a negative decision, knowing that two colleagues have already reviewed the images without finding anything suspicious. This is likely to lead to an underestimation of the ability of AI, rather than an overestimation, in terms of detecting cancer. Furthermore, before the start of the study, we calibrated the AI abnormality threshold on the basis of retrospective data. Our initial calibration of the threshold value aimed for a 2% increase in true positive fraction, which led to an actual 6% increase in the current study. This shows that setting the threshold on the basis of retrospective data might not always be sufficient, and that



	Abnormal interpretation	Recall	Biopsy	Cancer	Cancer, invasive >2 cm or lymph node metastasis
Double reading by two radiologists, n (95% CI)	7384 (7168–7605)	2931 (2793–3075)	694 (628–768)	450 (397–510)	139 (110–174)
Double reading by AI plus one radiologist, n (95% CI)	8946 (8710–9187)	2800 (2665–2941)	725 (657–800)	470 (415–531)	146 (116–182)
Singe reading by AI, n (95% CI)	5689 (5499–5886)	1549 (1449–1656)	628 (565–698)	443 (390–502)	140 (112–176)
Triple reading by two radiologists plus AI, n (95% CI)	10 799 (10 542–11 060)	3087 (2946–3235)	745 (676–821)	484 (429–546)	148 (118–184)

***Figure 3:* Number of positive assessments normalised to a screening population of 100 000 women**

The numbers are reported at each stage of the screening workflow: abnormal interpretation by any reader in the independent assessment, recall decision in the consensus discussion, and biopsy decision in the further diagnostic investigation of recalled women. AI=artificial intelligence.

repeated calibration in clinical use might be necessary to maintain a desired operating point. There are currently no quality assurance protocols to detect and correct data drift affecting the AI system performance. Indeed, this is one of the key barriers to implementation of AI in health care. The main limitation of the single-arm paired design arises at a later stage. It does not permit future comparisons of potential differences in interval cancer rates between reader strategies. A consequence of ScreenTrustCAD following the standard protocol for screening trials in which only participants who test positive on the screening test are triaged for further investigation and potential disease status verification using tissue sampling (a screen-positive design) is that absolute estimates of sensitivity and negative predictive value cannot be directly calculated, since we cannot with enough certainty corroborate that no cancer was present for women who did not have a biopsy sample taken. For ScreenTrustCAD, these metrics can be approximated once we have conducted the planned 23-month follow-up study. The specific study setting implies limitations with respect to generalisability of the study results. In particular, the results were obtained within a double reading followed by consensus discussion workflow, using Philips mammography equipment and the AI system INSIGHT MMG from Lunit. Finally, we had to exclude women with breast implants because the AI system had not been validated in this population.

Recently, the prospective MASAI trial reported on outcomes of using AI in screening mammography.[19] In the current paper, we found an increase in cancer detection, compared with standard-of-care, when using AI in combination with radiologists, whereas this was not observed in the MASAI trial. The observed workload reduction in terms of initial reads was similar in the two studies, with a 50% reduction for the superior AI-plus-one-reader strategy in our study compared with 44% for the non-inferior reader strategy in the MASAI trial. The workload reduction would be even greater, 100%, using the non-inferior AI-only reader strategy in our study. Taken together, the results of these two studies have shown the clinical validity—ie, diagnostic accuracy—of AI in real-world mammography screening ranging from superior to non-inferior. Regarding clinical utility, both show large workload reductions between 44% and 100%.

This study has shown that a strategy of double reading by one radiologist plus AI resulted in an increased cancer detection rate compared with double reading by two radiologists. Two combined reasons contributed to the results: the ability of AI to detect cancer with sufficient sensitivity, and the ability of consensus readers to increase specificity by dismissing AI false positives. We also showed that single reading by AI would have a similar cancer detection rate, but a markedly lower recall rate compared with double reading by two radiologists. Unsurprisingly, the triple reading by two radiologists plus AI detected the most cancers, which must be weighed against the increased costs, participant worry due to increased recalls, and failing to address the shortage of breast radiologists. We intend to perform a later follow-up study to examine interval cancers that might have received an initial positive read by any radiologist or AI, and were then later dismissed by the consensus discussion. In addition to workload reduction, implementing AI might also help reducing intrareader and inter-reader variability, improving consistency of assessments. As an independent and external evaluation of AI for screening mammography, our study shows that the AI in the study setting is ready for controlled implementation, which would include risk management, post-market surveillance, and systematic real-world follow-up of performance.

Contributors

FS and ME conceived and designed the study. KD, EC, and FS collected the data. KD, FS, AC, and ME analysed and interpreted the results. KD and FS prepared the draft manuscript. All authors reviewed and verified the data and results. KD and FS accessed and verified the data in the study. All authors approved the final version of the manuscript and had final responsibility for the decision to submit for publication.

Declaration of interests

FS is supported by Medtechlabs (funded by Region Stockholm, Karolinska Institutet, and the Royal Institute of Technology), Region Stockholm, the Swedish Research Council, the Swedish Cancer Society, and the Breast Cancer Association; and has received speaker fees from Lunit and Pfizer. ME is supported by the Swedish Research Council and the Swedish Cancer Society. All other author declare no competing interests.

Data sharing

Individual patient data will be shared to the extent that anonymity can be maintained, that the recipient has an ethical approval to conduct the research, and with a data transfer agreement. A request to obtain study data can be sent to the principal investigator (fredrik.strand@ki.se) who will coordinate with the study hospital to ensure that any data sharing is in compliance with General Data Protection Regulations, Swedish law, and other legal agreements.

Acknowledgments

This study was supported by cost reimbursement to the study hospital from Lunit and grants from the Swedish Research Council, the Swedish Cancer Society, and Region Stockholm. We thank the trial participants, trial site staff, and Lunit for their support of this trial. We would especially like to thank Jonathan Waldenström for managing the information technology aspects, Lennart Skärblom for database queries, and Anders Byström for organisational support. We thank Jeff Matthews for detailed and critical review of the submitted manuscript.

References

1 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* 2020; **297:** 33–39.

2 Kwee TC, Kwee RM. Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence. *Insights Imaging* 2021; **12:** 88.

3 The Royal College of Radiologists. RCR clinical radiology census report 2021. https://www.rcr.ac.uk/clinical-radiology/rcr-clinical-radiology-census-report-2021 (accessed Feb 27, 2023).

4 Kim H-E, Kim HH, Han B-K, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* 2020; **2:** e138–48.

5 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:** 244–49.

6 McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577:** 89–94.

7 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 Health* 2020; **2:** e468–74.

8 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:** 916–22.

9 Leibig C, Brehmer M, Bunk S, Byng D, Pinker K, Umutlu L. Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. *Lancet Digit Health* 2022; **4:** e507–19.

10 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:** 1581–88.

11 Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* 2021; **374:** n1872.

12 Hickman SE, Baxter GC, Gilbert FJ. Adoption of artificial intelligence in breast imaging: evaluation, ethical constraints and limitations. *Br J Cancer* 2021; **125:** 15–22.

13 Pepe MS, Alonzo TA. Comparing disease screening tests when true disease status is ascertained only for screen positives. *Biostatistics* 2001; **2:** 249–60.

14 Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press, 2003.

15 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* 2020; **33:** 408–13.

16 Alonzo TA, Pepe MS, Moskowitz CS. Sample size calculations for comparative studies of medical tests for detecting presence of disease. *Stat Med* 2002; **21:** 835–52.

17 Larsen M, Aglen CF, Lee CI, et al. Artificial intelligence evaluation of 122 969 mammography examinations from a population-based screening program. *Radiology* 2022; **303:** 502–11.

18 Lee SE, Han K, Yoon JH, Youk JH, Kim E-K. Depiction of breast cancers on digital mammograms by artificial intelligence-based computer-assisted diagnosis according to cancer characteristics. *Eur Radiol* 2022; **32:** 7400–08.

19 Lång K, Josefsson V, Larsson A-M, et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol* 2023; **24:** 936–44.