Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study

Veronica Hernström, Viktoria Josefsson, Hanna Sartor, David Schmidt, Anna-Maria Larsson, Solveig Hofvind, Ingvar Andersson, Aldana Rosso, Oskar Hagberg, Kristina Lång

Summary

Background Emerging evidence suggests that artificial intelligence (AI) can increase cancer detection in mammography screening while reducing screen-reading workload, but further understanding of the clinical impact is needed.

Methods In this randomised, controlled, parallel-group, non-inferiority, single-blinded, screening-accuracy study, done within the Swedish national screening programme, women recruited at four screening sites in southwest Sweden (Malmö, Lund, Landskrona, and Trelleborg) who were eligible for mammography screening were randomly allocated (1:1) to AI-supported screening or standard double reading. The AI system (Transpara version 1.7.0 ScreenPoint Medical, Nijmegen, Netherlands) was used to triage screening examinations to single or double reading and as detection support highlighting suspicious findings. This is a protocol-defined analysis of the secondary outcome measures of recall, cancer detection, false-positive rates, positive predictive value of recall, type and stage of cancer detected, and screen-reading workload. This trial is registered at ClinicalTrials.gov, NCT04838756 and is closed to accrual.

Findings Between April 12, 2021, and Dec 7, 2022, 105 934 women were randomly assigned to the intervention or control group. 19 women were excluded from the analysis. The median age was 53·7 years (IQR 46·5–63·2). AI-supported screening among 53 043 participants resulted in 338 detected cancers and 1110 recalls. Standard screening among 52 872 participants resulted in 262 detected cancers and 1027 recalls. Cancer-detection rates were 6·4 per 1000 (95% CI 5·7–7·1) screened participants in the intervention group and 5·0 per 1000 (4·4–5·6) in the control group, a ratio of 1·29 (95% CI 1·09–1·51; p=0·0021). AI-supported screening resulted in an increased detection of invasive cancers (270 *vs* 217, a proportion ratio of 1·24 [95% CI 1·04–1·48]), wich were mainly small lymph-node negative cancers (58 more T1, 46 more lymph-node negative, and 21 more non-luminal A). AI-supported screening also resulted in an increased detection of in situ cancers (68 *vs* 45, a proportion ratio of 1·51 [1·03–2·19]), with about half of the increased detection being high-grade in situ cancer (12 more nuclear grade III, and no increase in nuclear grade I). The recall and false-positive rate were not significantly higher in the intervention group (a ratio of 1·08 [95% CI 0·99–1·17; p=0·084] and 1·01 [0·91–1·11; p=0·92], respectively). The positive predictive value of recall was significantly higher in the intervention group compared with the control group, with a ratio of 1·19 (95% CI 1·04–1·37; p=0·012). There were 61 248 screen readings in the intervention group and 109 692 in the control group, resulting in a 44·2% reduction in the screen-reading workload.

Interpretation The findings suggest that AI contributes to the early detection of clinically relevant breast cancer and reduces screen-reading workload without increasing false positives.

Funding Swedish Cancer Society, Confederation of Regional Cancer Centres, and Swedish governmental funding for clinical research.

Lancet Digit Health 2025

Published Online
February 3, 2025
https://doi.org/10.1016/S2589-7500(24)00267-X

See Online/Comment
https://doi.org/10.1016/j.landig.2025.01.004

Diagnostic Radiology, Translational Medicine, Lund University, Lund, Sweden (V Hernström MD, V Josefsson MD, H Sartor MD PhD, D Schmidt MD, I Andersson MD PhD, A Rosso MSc PhD, O Hagberg MSc PhD, K Lång MD PhD); **Unilabs: Mammography Unit, Skåne University Hospital, Malmö, Sweden** (H Sartor, D Schmidt, I Andersson, K Lång); **Radiology Department, Skåne University Hospital, Malmö, Sweden** (V Hernström, V Josefsson); **Oncology, Clinical Sciences Lund, Lund University, Lund, Sweden** (A-M Larsson MD PhD); **Section for Breast Cancer Screening, Cancer Registry of Norway, Oslo, Norway** (Prof S Hofvind PhD); **Health and Care Sciences, Faculty of Health Sciences, The Arctic University of Norway, Tromsø, Norway** (Prof S Hofvind)

Correspondence to:
Dr Kristina Lång, Division of Diagnostic Radiology, Department of Translational Medicine, Lund University, Malmö 20502, Sweden
kristina.lang@med.lu.se

Introduction

Use of artificial intelligence (AI) in mammography screening has the potential to reduce the screen-reading workload and increase cancer detection, which could affect patient outcomes. These claims need to be evaluated in prospective trials and the potential benefits weighed against possible negative effects.[1]

The double reading of screening examinations is standard of care in European screening programmes.[2] Reducing the double-reading workload by replacing part of human screen reading with AI would have an immediate positive effect on staffing of breast radiologists, who are in short supply in many countries. The reduction in screen-reading workload should not,

Research in context

Evidence before this study

Mammography screening has been shown to be effective in reducing breast cancer mortality. However, cancers are still missed in screening despite the double screen-reading procedure recommended by European guidelines. Some of these cancers are rapidly progressive and can appear as interval cancers before the next screening round. Artificial intelligence (AI) could potentially be used to support radiologists in screen reading. We searched MEDLINE for studies published in English between Jan 1, 2015, and Dec 31, 2020, which included "breast cancer screening" or "mammography screening", and "artificial intelligence" or "machine learning" in the title or abstract. No prospective trials were identified. We found no systematic reviews on test accuracy. A first report on clinical safety from the Mammography Screening with Artificial Intelligence trial (MASAI) showed that an AI-supported screen-reading procedure, involving triage and detection support, was considered safe because the cancer-detection rate did not decline despite a substantial reduction in the screen-reading workload.

Added value of this study

To our knowledge, this is the first randomised controlled trial investigating the use of AI in mammography screening. In this protocol-defined analysis, the objective was to study early screening performance measures and screen-reading workload together with a characterisation of the type and stage of detected cancers in the entire trial population. Characterisation of detected cancers is important for our improved understanding of the clinical impact of AI-supported mammography screening. The AI-supported screen-reading procedure resulted in a significant increase in cancer detection compared with standard double reading, without increasing the false-positive rate while reducing the screen-reading workload. The increased detection was predominantly of small, lymph-node negative, invasive cancers, and in addition to luminal A, included more detected triple-negative, human epidermal growth factor receptor 2 positive, and luminal B cancers. There was no increased detection of low-grade ductal carcinoma in situ. The results indicate that an AI-supported screen reading procedure can contribute to the early detection of breast cancer likely to be clinically progressive.

Implications of all the available evidence

Taken together, results of this randomised controlled trial indicate that an AI-supported screen-reading procedure can safely be used to reduce the screen-reading workload and that the significant increase in cancer detection probably contributes to the early detection of clinically relevant breast cancer. Assessment of the primary endpoint of the interval cancer rate will provide further insight into the prognostic implications of use of AI in mammography screening.

however, be at the cost of an increase in consensus meetings or false-positive recalls.

Breast cancer is a heterogeneous disease ranging from indolent to aggressive types.[3] The characterisation of cancers on the basis of morphology, immunohistochemical biomarkers, and molecular subtype, together with size, lymph node involvement, and distant metastases, yields prognostic and predictive information used in treatment planning and the follow-up of patients with breast cancer.[3,4] Retrospective studies suggest that AI has a high mammographic sensitivity and might therefore reduce the number of cancers overlooked in screening.[5–8]

Results of a few prospective studies published before the start of this study indicate increased cancer detection when AI is used in mammography screening, but little is known of the types and stages of the cancers detected.[9–11] Detecting more cancers with AI support should not come at the expense of an unacceptable increase in false positives or predominantly identify indolent cancers. The disproportionately increased detection of indolent cancers, such as low-grade in situ cancers, would add to the harm of overdiagnosis and overtreatment. The use of AI should therefore ideally lead to the increased detection of cancers that are clinically relevant in terms of morbidity and mortality. Nevertheless, the field is advancing rapidly, with AI already being implemented in some screening programmes and several randomised trials in planning or starting phases. Results from prospective trials can thus provide timely and informative evidence on how the use of AI, along with the choice of screening protocol, affects overall screening performance, detection across cancer subtypes, and potential shifts in cancer stage.

The Mammography Screening with Artificial Intelligence trial (MASAI) is a randomised controlled trial investigating AI-supported screening compared with standard double reading without AI. In the trial, AI was used to triage examinations to single or double reading, depending on the AI risk score, and as detection support for radiologists, with AI highlighting suspicious findings in the image. In a previous report from this trial, the clinical safety of AI-supported screening was assessed in the first 80 000 enrolled participants. AI-supported screening was considered safe since the cancer-detection rate did not decline despite a 44% reduction in screen-reading workload.[12] In this second protocol-defined analysis of the MASAI trial, early screening performance measures and the type and stage of detected cancers have been assessed in the entire trial population. The characterisation of detected cancers can advance our understanding of the clinical impact of the use of AI in mammography screening.

Methods

Study design and participants

The MASAI trial was designed as a randomised, parallel-group, non-inferiority, single-blinded, controlled, screening-accuracy study, aimed at comparing AI-supported mammography screening with standard double reading without AI. The study is described in more detail elsewhere.[12] Within the Swedish national screening programme, participants were recruited at four screening sites in southwest Sweden (Malmö, Lund, Landskrona, and Trelleborg). The inclusion criterion was women eligible to participate in population-based mammography screening, which included general screening for women aged 40–74 years at 1·5–2-year screening intervals and annual screening for those with moderate hereditary risk of breast cancer or a history of breast cancer (for 10 years after surgery, with an upper age limit of 80 years). No exclusion criteria were applied. Information about the study was included in the standard screening invitation letters and in SMS text message reminders before scheduled appointments, with a link to a website containing detailed study information in Swedish and English. Women eligible for screening who did not wish to participate in the trial were asked to opt out at the time of the screening visit and received standard of care. The study was approved by the Swedish Ethical Review Authority (2020-04936, 2023-026848-02), which also waived the need for written informed consent. The protocol was updated to improve clarity; there were no changes in the trial procedures or analyses in the statistical analysis plan from those described in the first and updated protocol versions.

Randomisation and masking

Randomisation was based on a single sequence of random assignments (1:1). After screening mammograms were acquired, examinations were automatically randomised in the Picture Archive and Communications System (PACS; Sectra, Linköping, Sweden) to AI-supported screening (intervention group) or standard double reading without AI (control group) with a pseudo-random number generator. Study participants and the radiographers acquiring the screening examination were masked, and the radiologists doing the screen reading were not masked, to study group allocation.

Procedures

The standard screening examination included two mammographic views per breast (ie, craniocaudal and mediolateral oblique views) with the addition of implant-displacement views for those with breast implants. Mammograms were acquired by use of a single-vendor mammography system (Senographe Pristina, GE Healthcare, Freiburg, Germany). The examinations randomised to the intervention group were analysed by use of the AI system Transpara version 1.7.0 (ScreenPoint Medical, Nijmegen, Netherlands).[5–8,10] Transpara provided an examination-based malignancy risk score on a scale from 1 to 10, and was pre-configured and calibrated to place approximately a tenth of the screening examinations in each risk score group. These scores were categorised into low risk (1–7), intermediate risk (8–9), and high risk (10). Examinations with the highest 1% risk (risk score threshold 9·8) were flagged in the PACS worklist as extra-high risk and labelled as 10H. For examinations with intermediate and high risk, the AI system also provided marks highlighting suspicious findings in the mammogram together with regional risk scores on a scale from 1 to 98 (figure 1). Examinations with low and intermediate risk underwent single reading and those with high risk underwent double reading. The readers had access to information about the AI risk scores both in the PACS worklists and on the image monitor. Image marks and regional scores were initially masked to the readers, who were instructed to turn these on at the end of the case reading. The readers were also instructed to recall cases with extra-high risk (10H), except for obvious false-positive findings. Screening examinations in the control group were not analysed with AI and underwent standard double reading. The final decision on the screen reading was either no suspicion of malignancy or recall. Before the decision was made, the readers had the opportunity to refer challenging cases to a consensus meeting or request a technical recall owing to inadequate image quality or positioning. At the consensus meeting, two radiologists reassessed the case for a joint decision of recall or not. Women could be recalled on the basis of abnormal mammographic findings or self-reported symptoms. The images acquired at technical recall were by default

For the **study protocol (versions 1.1 and 1.2) and the statistical analysis plan** see https://portal.research.lu.se/en/projects/mammography-screening-with-artificial-intelligence

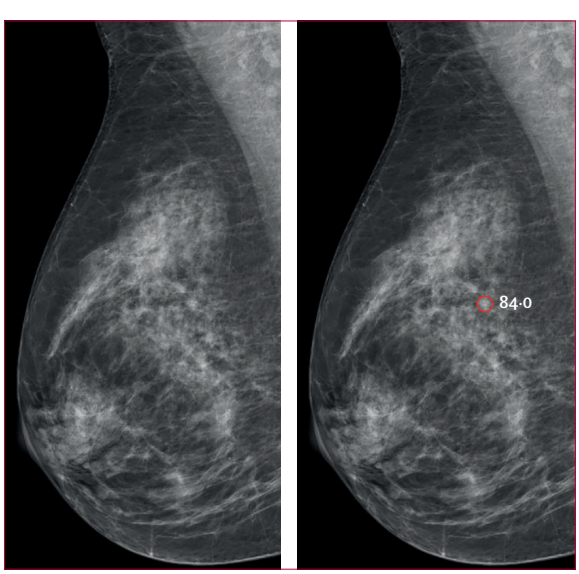

Figure 1: **Mediolateral oblique mammographic view of a screening examination with risk score 10**
The radiologists first read the examination without artificial intelligence (AI) marks followed by AI marks, which in this case highlights a small spiculated mass circled in red. The woman was recalled for investigation and diagnosed with a 5-mm lymph-node negative invasive cancer.

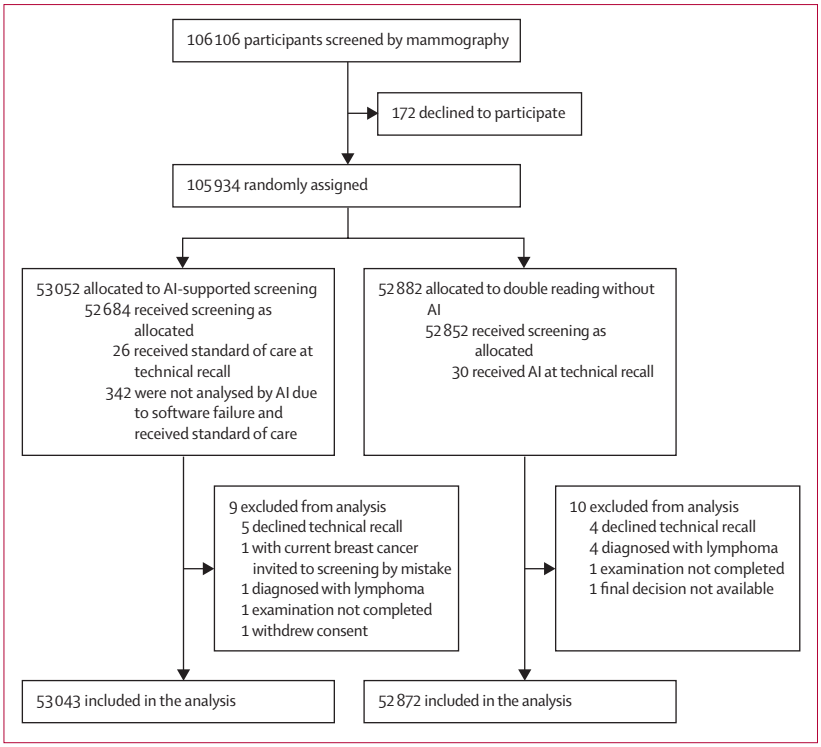

Figure 2: **Trial profile**
AI=artificial intelligence. Due to the technical set-up the technical recalls were randomised de novo but were assessed according to their original allocation (intention-to-treat policy).

randomised de novo owing to the technical setup; however, participants were assessed according to their originally assigned group. Screening examinations allocated to the intervention group that were not processed by AI underwent standard-of-care reading.

In total, sixteen breast radiologists participated in the screening reading, of whom 11 had an annual screen-reading volume of over 5000 examinations, three read more than 3500 examinations annually, and two read less than 3500 examinations annually. All but two radiologists had more than 5 years of experience in screen reading. KL and IA participated in the screen reading as part of their clinical work. The radiologists rolled a dice before each screen-reading session to randomly allocate themselves to one of the two study groups: numbers 1–3 allocated them to the control group and 4–6 to the intervention group. Further details on the screen-reading process are described elsewhere.[12]

Ground truth was based on pathology reports on surgical specimens or core-needle biopsies. True positives were identified through linkage with the Regional Cancer Registry together with manual assessment of the patient records, including pathology reports, of all recalled participants. Cancers were categorised as invasive, in situ, histological type, Nottingham histological grade (invasive), and nuclear grade (in situ). Molecular subtype was based on surrogate classification groups by use of immunohistochemical biomarkers (oestrogen receptor, progesterone receptor, human epidermal growth factor receptor 2 [HER2, also known as ERBB2], and Ki67) according to national guidelines.[13] On the basis of local laboratory routine, a Ki67 hotspot (low 0–20, intermediate 21–30, and high >30) was used until early 2023, and thereafter Ki67 global. TNM staging of detected cancers was determined according to pathological size, lymph-node involvement, and anatomical–prognostic group based on the Union for International Cancer Control's classification system, 8th edition.[14] The clinical tumour size and lymph-node involvement was used for staging when participants received neoadjuvant chemotherapy and for those with conservative treatment (who did not wish to undergo surgical treatment). In case of multifocal or bilateral involvement, the most clinically relevant tumour was selected on the basis of the order of invasiveness, TNM stage, or molecular subtype, followed by histological grade.

Participants could withdraw from the study at any time, at which point all personal data would be removed and they would be excluded from analysis.

Outcomes

The primary outcome measure of the MASAI trial is the interval cancer rate, which will be assessed after all participants of the study have had at least a 2-year follow-up (estimated December, 2024 plus 6 months to ensure all events are registered in the cancer registry). All outcome measures are described in detail elsewhere.[12] In this analysis, the following secondary outcome measures are reported; early screening performance (ie, cancer-detection rate, recall rate, false-positive rate, and positive predictive value [PPV] of recall), screen-reading workload (ie, number of screen readings and consensus meetings), and detection in relation to tumour type and stage.

Statistical analysis

The intention-to-treat population comprised all participants who did not opt out of the study and who underwent breast cancer screening. The modified intention-to-treat population comprised participants with a complete screening examination, excluding those who did not attend technical recall and those who were recalled owing to suspicion and diagnosis of lymphoma. Participants were analysed in their allocated group regardless of the actual reading procedure (treatment policy strategy). If participants were screened twice (applicable to those undergoing annual screening), only the first screening visit was analysed. The sample size of the trial was calculated for the primary variable of interval cancer rate. In order to prove non-inferiority with a non-inferiority margin of 20%, it was determined that 100 000 participants needed to be recruited. Descriptive statistics were used to summarise baseline population characteristics. Frequencies and percentages were calculated for categorical data. The Clopper–Pearson method was used to calculate 95% CIs. The cancer-detection rate, recall

	Intervention group (n=53 043)	Control group (n=52 872)
Age, years		
Mean (SD)	55·1 (10·2)	55·1 (10·2)
Range	40–80	40–80
<45	10 316 (19·4%)	10 286 (19·5%)
45–49	9689 (18·3%)	9739 (18·4%)
50–54	8702 (16·4%)	8710 (16·5%)
55–59	6898 (13·0%)	6650 (12·6%)
60–64	6100 (11·5%)	6281 (11·9%)
65–69	5586 (10·5%)	5454 (10·3%)
≥70	5752 (10·8%)	5752 (10·9%)
Screening indication		
General screening	51 921 (97·9%)	51 708 (97·8%)
History of breast cancer	1071 (2·0%)	1102 (2·1%)
Moderate hereditary risk	51 (0·1%)	62 (0·1%)

Data are mean (SD), range, or n (%).

Table 1: **Baseline population characteristics (modified intention-to-treat population)**

	Intervention group (n=53 043)	Control group (n=52 872)	Proportion ratio	p value
Early screening performance				
Number of recalls	1110	1027	..	..
Recall rate	2·1% (2·0–2·2)	1·9% (1·8–2·1)	1·08% (0·99–1·17)	0·084
Number of detected cancers	338	262	..	..
Cancer-detection rate, per 1000	6·4 (5·7–7·1)	5·0 (4·4–5·6)	1·29 (1·09–1·51)	0·0021
Number of false positives	772	765	..	..
False positive rate	1·5% (1·4–1·6)	1·4% (1·3–1·6)	1·01% (0·91–1·11)	0·92
Positive predictive value of recall	30·5% (27·8–33·3)	25·5% (22·9–28·3)	1·19% (1·04–1·37)	0·012
Workload				
Number of screen readings	61 248	109 692	..	..
Number of consensus meetings	2023	1958	..	..
Consensus meeting rate	3·8%	3·7%	..	..

Data are n or point estimate (95% CI).

Table 2: **Early screening performance and workload measures (modified intention-to-treat population)**

rate, false-positive rate, and PPV of recall were calculated separately for the intervention and control groups. These were compared by use of Fisher's exact test and 95% CIs were computed by use of the log-normal method.[15] A two-sided p value under 0·05 was considered to indicate significance. 95% CIs from the log-normal method were also used to compare the frequency of different subcategories of detected cancers, calculated both on aggregated and single subgroups, of which the single subgroups included only those with at least five cases. The difference in workload was calculated by comparing the number of readings in each group in relation to group size. Statistical analyses were done by use of R version 4.3.0.[16] The trial is registered with ClinicalTrials.gov, NCT04838756. Further details of the statistical analysis are presented elsewhere.[12]

Role of the funding source

The funders of the study had no part in the study design, data collection, data analysis, data interpretation, writing of the report, or decision to submit.

Results

Between April 12, 2021, and Dec 7, 2022, 106 106 women presented for screening, and 172 (0·2%) opted out of the trial. 105 934 participants were randomly assigned: 53 052 to the intervention group, undergoing AI-supported screening, and 52 882 to the control group, undergoing standard of care (ie, double reading without AI). A total of 19 women were excluded from the final analysis, resulting in a modified intention-to-treat population of 53 043 participants in the intervention group and 52 872 participants in the control group (figure 2). The median age of all participants was 53·7 years (IQR 46·5–63·2). The age distribution and indication to screen were similar between groups (table 1). In the intervention group, 3800 (7·2%) examinations were classified as high risk (risk score 10) and underwent double reading, and 655 of these were flagged as extra-high risk (risk score 10H). The AI system did not provide a malignancy risk score for 368 (0·7%) of the examinations in the intervention group. There were 44 (0·1%) technical recalls in the intervention group and 52 (0·1%) in the control group.

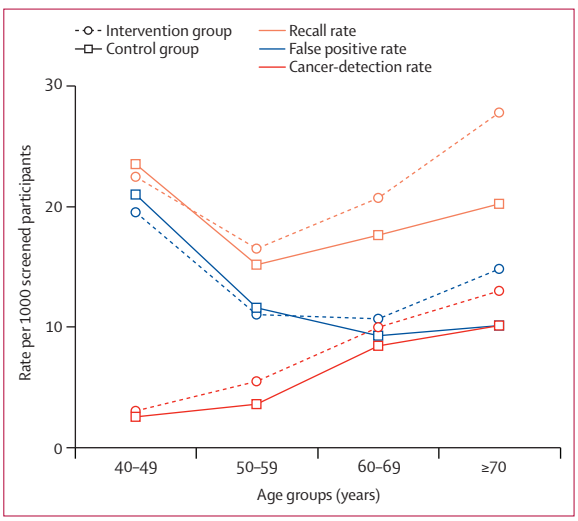

Figure 3: **Early screening performance measures per 1000 screened participants for 10-year age groups (modified intention-to-treat population)**

Early screening performance and workload measures are presented in table 2. There was a significant 29% increase in cancer detection when AI-supported screening was used compared with standard of care. There were non-significant increases in the recall rate (8%) and

	Intervention group (n=53 043)	Control group (n=52 872)	Proportion ratio (95% CI)
Number of cancers, total	338 (6·37)	262 (4·96)	1·29 (1·09–1·51)
Number of invasive cancers	270 (5·09)	217 (4·10)	1·24 (1·04–1·48)
Number of in situ cancers	68 (1·28)	45 (0·85)	1·51 (1·03–2·19)
Histological type			
No special type	204 (3·85)	155 (2·93)	1·31 (1·06–1·62)
Invasive lobular cancer	31 (0·58)	40 (0·76)	0·77 (0·48–1·23)
Other Invasive	35 (0·66)	22 (0·42)	1·59 (0·93–2·70)
Mucinous	8 (0·15)	4 (0·08)	..
Tubular	10 (0·19)	5 (0·09)	1·99 (0·68–5·83)
Mixed type	6 (0·11)	5 (0·09)	1·20 (0·37–3·92)
Miscellaneous*	11 (0·21)	8 (0·15)	1·37 (0·55–3·41)
Ductal carcinoma in situ	64 (1·21)	43 (0·81)	1·48 (1·01–2·18)
Other in situ†	4 (0·08)	2 (0·04)	..
Histological grade, invasive			
I	100 (1·89)	67 (1·27)	1·49 (1·09–2·03)
II	128 (2·41)	118 (2·23)	1·08 (0·84–1·39)
III	35 (0·66)	29 (0·55)	1·20 (0·74–1·97)
Not applicable‡	7 (0·13)	3 (0·06)	..
Nuclear grade, ductal carcinoma in situ			
I	3 (0·06)	3 (0·06)	..
II	29 (0·55)	20 (0·38)	1·45 (0·82–2·55)
III	32 (0·60)	20 (0·38)	1·59 (0·91–2·79)
Molecular subtype, invasive			
Luminal A	166 (3·13)	137 (2·59)	1·21 (0·96–1·51)
Non-luminal A	98 (1·85)	77 (1·46)	1·27 (0·94–1·71)
Luminal B	65 (1·23)	58 (1·10)	1·12 (0·78–1·59)
Triple negative	16 (0·30)	6 (0·11)	2·66 (1·04–6·79)
HER2 positive–ER positive	13 (0·25)	6 (0·11)	2·16 (0·82–5·68)
HER2 positive–ER negative	4 (0·08)	7 (0·13)	..
Not assessable‡	6 (0·11)	3 (0·06)	..
T stage			
Tis	68 (1·28)	45 (0·85)	1·51 (1·03–2·19)
T1	226 (4·26)	168 (3·18)	1·34 (1·10–1·64)
T1mi	7 (0·13)	2 (0·04)	..
T1a	20 (0·38)	18 (0·34)	1·11 (0·59–2·09)
T1b	75 (1·41)	69 (1·31)	1·08 (0·78–1·50)
T1c	124 (2·34)	79 (1·49)	1·56 (1·18–2·07)
T2+	44 (0·83)	48 (0·91)	0·91 (0·61–1·38)
T2	38 (0·72)	39 (0·74)	0·97 (0·62–1·52)
T3	4 (0·08)	8 (0·15)	..
T4	2 (0·04)	1 (0·02)	..
Not applicable§	0	1 (0·02)	..
(Continued from previous column)			
N stage, invasive			
N0	206 (3·88)	160 (3·03)	1·28 (1·04–1·58)
N1+	60 (1·13)	55 (1·04)	1·09 (0·75–1·57)
N1	55 (1·04)	43 (0·81)	1·27 (0·86–1·90)
N2	1 (0·02)	8 (0·15)	..
N3	4 (0·08)	4 (0·08)	..
Nx	3 (0·06)	1 (0·02)	..
Not applicable§	0	1 (0·02)	..
Missing¶	1 (0·02)	0	..
TNM stage			
0	68 (1·28)	45 (0·85)	1·51 (1·03–2·19)
1	193 (3·64)	139 (2·63)	1·38 (1·11–1·72)
1A	181 (3·41)	135 (2·55)	1·34 (1·07–1·67)
1B	12 (0·23)	4 (0·08)	..
2+	73 (1·38)	76 (1·44)	0·96 (0·69–1·32)
2A	49 (0·92)	44 (0·83)	1·11 (0·74–1·67)
2B	15 (0·28)	17 (0·32)	0·88 (0·44–1·76)
3A	3 (0·06)	10 (0·19)	..
3B	1 (0·02)	1 (0·02)	..
3C	4 (0·08)	4 (0·08)	..
4	1 (0·02)	0	..
Not applicable‖	3 (0·06)	2 (0·04)	..
Missing¶	1 (0·02)	0	..

Data are n (per 1000 screened participants). The rightmost column gives the proportion ratio with 95% CI when the intervention group is compared with the control group. This analysis included only those with at least five cases per single subgroup. HER2 (also known as ERBB2)=human epidermal growth factor receptor 2. ER=oestrogen receptor. *Other invasive miscellaneous: apocrine, papillary, microinvasive, cribriform, and one case of angiosarcoma. †Other in situ: florid lobular, papillary, and mixed papillary and ductal carcinoma in situ. ‡Not applicable due to microinvasive cancer, too small to assess, and one case of angiosarcoma. §Not applicable due to one case of angiosarcoma. ¶One case of unknown N stage due to out-of-county treatment. ‖Not applicable due to Nx stage, and one case of angiosarcoma.

Table 3: **Cancer type and stage; frequency of subcategories of detected cancers**

false-positive rate (1%) in the intervention group compared with the control group, which resulted in 83 more recalls and seven more false positives, and a significant increase in PPV of recall of 19% (table 2). In the intervention group, 941 (84·4%) women were recalled owing to mammographic findings and 169 (15·2%) owing to reported symptoms, compared with 847 (82·5%) and 180 (17·5%), respectively, in the control group. There were 48 444 fewer screen readings and 65 more consensus meetings in the intervention group than the control group, which is a 44·2% reduction in the screen-reading workload at a similar consensus meeting rate of 3·8% (table 2).

There were more detected cancers across 10-year age groups and a higher false-positive rate starting from the age of 60 years in the intervention group than the control group (figure 3). For participants undergoing annual screening and with a history of breast cancer (n=2173), ten cancers were detected with AI-supported screening and eight with standard screening, whereas no cancers were detected in the moderate hereditary risk group (n=113). There were eight bilateral cancers in the intervention group and 12 in the control group.

Type and stage of detected breast cancer (number and frequency per 1000 screened participants) and the proportion ratio comparing the intervention group with the control group are presented in table 3. AI-supported screening led to 76 more detected cancers (338 *vs* 262) than did standard screening, consisting of 53 more invasive cancers (270 *vs* 217, a proportion ratio of 1·24 [95% CI 1·04–1·48]) and 23 more in situ cancers (68 *vs* 45, a proportion ratio of 1·51 [1·03–2·19]). There was increased detection across histological types, except for invasive lobular cancers (31 *vs* 40). The largest increase was of invasive cancers of no special type (204 *vs* 155). There was also increased detection of invasive cancer across histological grades, with the largest increase in grade I cancers (100 *vs* 67). There were 21 more detected ductal carcinomas in situ in the intervention group than the control group, of which 12 were nuclear grade III and none were nuclear grade I. Regarding molecular subtypes, AI-supported screening compared with standard screening resulted in 29 more detected luminal A (166 *vs* 137) and 21 more detected non-luminal A (98 *vs* 77) invasive cancers, with the latter consisting of cancers of the triple negative (16 *vs* 6), luminal B (65 *vs* 58), and HER2-positive (17 *vs* 13) subtypes. There were similar numbers of detected T2+ cancers (>20 mm) in the intervention group compared with the control group (44 *vs* 48), but 58 more detected T1 cancers (≤20 mm, 226 *vs* 168), which mainly consisted of an increase in T1c cancers (11–20 mm, 124 *vs* 79). There were five more lymph-node positive and 46 more lymph-node negative invasive cancers in the intervention group than the control group (60 *vs* 55 and 206 *vs* 160, respectively). There was a similar number of women with TNM stage 2+ in the intervention group compared with the control group (73 *vs* 76), but 77 more detected cancers below stage 2 (261 *vs* 184), comprising 23 more stage 0 (in situ), 46 more stage 1A (T1, N0), and eight more stage 1B (T1, N1mi) detected cancers. Only one participant had distant metastasis at diagnosis (stage 4) in the intervention group. Most of the detected non-luminal A cancers in the intervention group were T1 (72 [73%] of 98) and lymph-node negative (69 [70%] of 98), which are proportions similar to those in the control group (46 [60%] of 77 and 50 [65%] of 77, respectively).

Discussion

This analysis showed that a screen-reading procedure that used AI to triage screening examinations to single or double reading and that used AI as detection support in mammography screening led to a significant 29% increase in cancer detection compared with standard double reading without AI (6·4 *vs* 5·0 per 1000 participants screened), with a similar false-positive rate and with a substantial 44% reduction in the screen-reading workload. The increase in detection mostly concerned small, lymph-node negative, invasive cancers.

Early screening performance measures and type and stage of detected cancers give first indications of the clinical impact of AI-supported screening.[17] The large increase in detected small, lymph-node negative, invasive cancers suggests that downstaging by earlier detection with use of AI is possible, which could be of clinical benefit since stage has a major influence on breast cancer treatment and prognosis.[3,4,18,19] The biological profile of breast cancers also has major prognostic and predictive significance.[3,20] AI-supported screening resulted in more detected invasive cancers of the non-luminal A molecular subtype, including more detected triple-negative cancers, compared with standard double reading without AI. Non-luminal A cancers have a poorer prognosis than do luminal A cancers and are more likely to become interval cancers.[20,21] The increased detection of small, lymph-node negative invasive cancers, in particular non-luminal A cancers, could therefore lead to a subsequent decrease in interval cancers and advanced cancers in the next screening round, which would need to be confirmed in long-term follow-up. The primary outcome of the MASAI trial is interval cancer rate, which will be assessed after a 2-year follow-up (estimated December, 2024 plus 6 months to ensure all events are registered in the cancer registry), shedding further light on the clinical impact.

AI-supported screening also led to a relative increase in the detection of in situ cancers, albeit to a smaller number than that of invasive cancers. Importantly, there was no increase in low-grade ductal carcinoma in situ, which would have added to the overdiagnosis burden of breast cancer screening.[22,23] Roughly half of the extra detected ductal carcinoma in situ were of nuclear grade III, which is considered clinically relevant early detection as the biological profile is more aggressive with a high likelihood of becoming invasive.[23] Still, the other half were of intermediate risk and could therefore potentially add to overdiagnosis, but the numbers are small (nine more detected ductal carcinoma in situ grade II). In the first report on 80 000 enrolled participants in the MASAI trial, a 20% increase in cancer detection, roughly equally divided between invasive and in situ cancers, was observed with AI-supported screening compared with standard screening.[12] The 29% increase reported here indicates a larger increase in cancer detection by use of AI and a larger proportion of detected invasive cancers. This could be the result of a learning curve of the participating radiologists who became accustomed to screen reading with AI and perhaps their increased trust in AI capabilities from reinforcing feedback during the routine clinical work-up of recalled women. Of note, participating radiologists, except for the principal investigator (KL), were not informed of the first results before all trial participants had been enrolled.

To date, there are few prospective data on the screening performance and types and stages of cancers detected by use of AI. In a Spanish study of 11 998 screened women aged 50–69 years, AI-supported double reading, with equal parts tomosynthesis and mammography examinations, was compared with double reading without AI in a propensity-score matched historical control group.[10] Taking

only mammograms into account, the study found an increase in the recall rate from 6·2% to 6·6%, and in the cancer-detection rate from 5·7 to 8·1 per 1000 screened women with use of AI. The increase in the cancer-detection rate was similar to ours for the same age group, but at a higher recall rate and without reducing the screen-reading workload. The types of detected cancers were reported for tomosynthesis and mammography together and showed, similar to our findings, relatively more detected invasive cancers, mostly of no special type, in relation to in situ cancer, and fewer detected invasive lobular cancers with use of AI-supported double reading compared with double reading without AI, but the numbers were small.[10] This could perhaps be because lobular cancers are more likely to receive low AI risk scores, but a retrospective study based on enriched data does not support this.[24] Lobular cancers are in general more prognostically favourable than those of no special type but have a higher risk of becoming interval cancers owing to their sometimes subtle mammographic appearance.[21,25] The future assessment of interval cancers in the MASAI trial will be informative.

Two prospective studies with paired design, from Sweden (n=55 581) and Hungary (n=15 953), have shown a relative increase in cancer detection of 4–5% when AI was added to single reading, in which AI was used to put cases above a certain threshold to a consensus meeting at which two radiologists had access to AI information[11] or in a live use double-reading setting,[9] compared with double reading without AI. The additional detected cancers were roughly equal parts invasive and in situ cancers, but the small number of additional detected cancers (n=11) limits further comparison of type and stage. The relatively larger increase in cancer detection in the MASAI trial could perhaps be explained by the use of different AI software, and by the mode of integrating AI in the screen-reading workflow. The MASAI screen-reading procedure emphasised radiologists having access to AI information at screen reading, in terms of both examination risk scores and regional marks highlighting suspicious findings in the mammogram. The rationale was to introduce a beneficial bias by making radiologists aware of the cancer prevalence when reading low-risk and high-risk examinations in order to influence them to reduce false positives in low cancer prevalence readings and to reduce false negatives in high cancer prevalence readings,[26] and to give access to regional marks by AI to reduce the risk of radiologists overlooking potential findings. Of the examinations analysed by AI, 7·2% were classified as high risk (risk score 10) but only 2·1% of participants in the intervention group were recalled. In addition, only 38·9% of the 1·2% of examinations classified as extra high risk by AI (10H) were recalled. This shows the importance of radiologists having the final decision to recall, as well as the high demand for radiologists to discard potentially inaccurate AI-flagged findings, in order to safeguard a low false-positive rate. The recall rate in the MASAI trial was low relative to international data,[27] but within national benchmarks.[28] We had a non-significant 8% increase in recall rate in the intervention group compared with the control group, but a significantly larger proportion of the recalls in the intervention group were true positives resulting in only seven more false positives compared with the control group. Again, further follow-up will show the net effect of this approach. The importance of how AI is integrated into the screen-reading pathway also highlights the need for randomised trials, in which the true effect of the radiologist working with AI, and its influence on medical decision making, can be studied. The MASAI trial is, to the best of our knowledge, the only randomised trial investigating AI in mammography screening, but there are other randomised trials with different screen-reading strategies in the planning or starting phase (eg, NCT06032390).

The large reduction in screen-reading workload made possible by the AI-supported screen-reading procedure would free up time for breast radiologists to spend on more complex patient-centred tasks. Whether the time saving is cost-effective is related to the cost of the AI system.[29] Since breast cancer treatment and related costs escalate with increasing stage,[3,4] downstaging through earlier detection by use of AI would suggest lower morbidity and treatment costs.[3,30] To address the cost-effectiveness of AI-supported screening, health economic analyses based on the MASAI trial are under way.

Limitations of this study relate to generalisability. The trial was done within the context of a Swedish screening programme, starting from the age of 40 years and with low baseline recall rates, and the use of a single mammography and AI vendor. Race and ethnicity were not registered, as these data are not routinely collected in the clinic owing to privacy considerations, but up to 35% of the targeted population is of immigrant background according to official statistics.

For **statistics on immigrant background** see www.scb.se

In summary, analysis of the screening performance and type and stage of detected cancer in the entire study population of the MASAI trial, in which an AI system was used to triage screening examinations to single or double reading and as detection support, showed that AI-supported screening was associated with a significant 29% increase in cancer detection compared with standard double reading without AI. The use of AI did not negatively influence the rates of recalls, false positives, or consensus meetings and the screen-reading workload was reduced by almost half. AI-supported screening predominantly contributed to the increased detection of small, lymph-node negative invasive cancers, which in addition to luminal A, included triple-negative, HER2 positive, and luminal B invasive cancers, and to a lesser extent to the increased detection of in situ cancers, comprising an increase in detection of high-grade, and no increase in detection of low-grade, ductal carcinoma in situ. Altogether, use of AI has the potential to increase the early detection of clinically relevant breast cancer without unduly increasing the harm of false positives

and overdiagnosis of low-grade in situ cancer. This study offers important insights into the clinical impact of the use of AI in mammography screening, but further follow-up addressing the interval cancer rate and cost-effectiveness is needed.

Contributors

KL and AR conceptualised and designed the trial with input from IA and SH. OH did the statistical analysis. KL, VH, VJ, DS, OH, and HS directly accessed and verified the underlying data reported in the manuscript. All authors were involved in data interpretation. VH, VJ, and KL wrote the first draft of the report. All authors revised the report and provided important intellectual content. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

KL has been an advisory board member for Siemens Healthineers and has received a lecture honorarium from AstraZeneca. All other authors declare no competing interests.

Data sharing

De-identified aggregated participant data will be made available on reasonable request, with investigator support and a signed data access agreement. A proposal should be submitted to the corresponding author to be reviewed by the study steering committee. Individual data are not publicly available owing to data protection regulations. The study protocol and statistical analysis plan are available online at https://portal.research.lu.se/en/projects/mammography-screening-withartificial-intelligence. Transpara by Screenpoint is a commercial algorithm and so the software and source code cannot be shared.

Acknowledgments

We thank the funders of the trial: the Swedish Cancer Society (21 1631Pj, 22 0611FE), the Confederation of Regional Cancer Centres (21/00060), and Swedish governmental funding of clinical research (ALF; 2020-Projekt0079, 2022-Projekt0100). We also thank the staff at the Unilabs Mammography Unit at Skåne University Hospital for making this study possible; Unilabs, Sectra, and ScreenPoint Medical for the technical support; and the participants involved in the trial. We would also like to thank Olle Berglind and Maria Swärd, and Lars and Margareta Söderström for providing KL with a place to write.

References

1 Huisman M, van Ginneken B, Harvey H. The emperor has few clothes: a realistic appraisal of current AI in radiology. *Eur Radiol* 2024; **34:** 5873–75.
2 Schünemann HJ, Lerda D, Quinn C, et al. Breast cancer screening and diagnosis: a synopsis of the European Breast Guidelines. *Ann Intern Med* 2020; **172:** 46–56.
3 Harbeck N, Penault-Llorca F, Cortes J, et al. Breast cancer. *Nat Rev Dis Primers* 2019; **5:** 66.
4 Loibl S, André F, Bachelot T, et al. Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann Oncol* 2024; **35:** 159–82.
5 Koch HW, Larsen M, Bartsch H, et al. How do AI markings on screening mammograms correspond to cancer location? An informed review of 270 breast cancer cases in BreastScreen Norway. *Eur Radiol* 2024; **34:** 6158–67.
6 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.
7 Lång K, Hofvind S, Rodríguez-Ruiz A, Andersson I. Can artificial intelligence reduce the interval cancer rate in mammography screening? *Eur Radiol* 2021; **31:** 5940–47.
8 Lauritzen AD, Rodríguez-Ruiz A, von Euler-Chelpin MC, et al. An artificial intelligence-based mammography screening protocol for breast cancer: outcome and radiologist workload. *Radiology* 2022; **304:** 41–49.
9 Dembrower K, Crippa A, Colón E, Eklund M, Strand F. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *Lancet Digit Health* 2023; **5:** e703–11.
10 Elías-Cabot E, Romero-Martín S, Raya-Povedano JL, Brehl AK, Álvarez-Benito M. Impact of real-life use of artificial intelligence as support for human reading in a population-based breast cancer screening program with mammography and tomosynthesis. *Eur Radiol* 2024; **34:** 3958–66.
11 Ng AY, Oberije CJG, Ambrózay É, et al. Prospective implementation of AI-assisted screen reading to improve early detection of breast cancer. *Nat Med* 2023; **29:** 3044–49.
12 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.
13 Bergh J. Nationellt vårdprogram bröstcancer, 2024. https://kunskapsbanken.cancercentrum.se/diagnoser/brostcancer/vardprogram (accessed March 13, 2024).
14 JD Brierley, Gospodarowicz, M, Wittekind C, eds. TNM classification of malignant tumours, 8th edn. Wiley-Blackwell, 2017.
15 Koopman PAR. Confidence intervals for the ratio of two binomial proportions. *Biometrics* 1984; **40:** 513–17.
16 The R Project for Statistical Computing. A language and environment for statistical computing. R Foundation, Vienna, Austria, 2023. https://www.R-project.org/ (accessed Sept 20, 2023).
17 Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. *Ann Oncol* 2008; **19:** 614–22.
18 Saadatmand S, Bretveld R, Siesling S, Tilanus-Linthorst MMA. Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173,797 patients. *BMJ* 2015; **351:** h4901.
19 Colzani E, Liljegren A, Johansson AL, et al. Prognosis of patients with breast cancer: causes of death and effects of time since diagnosis, age, and tumor characteristics. *J Clin Oncol* 2011; **29:** 4014–21.
20 Hennigs A, Riedel F, Gondos A, et al. Prognosis of breast cancer molecular subtypes in routine clinical care: a large prospective cohort study. *BMC Cancer* 2016; **16:** 734.
21 Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. *NPJ Breast Cancer* 2017; **3:** 12.
22 Segnan N, Ponti A. Artificial intelligence for breast cancer screening: breathtaking results and a word of caution. *Lancet Oncol* 2023; **24:** 830–32.
23 van Seijen M, Lips EH, Thompson AM, et al. Ductal carcinoma in situ: to treat or not to treat, that is the question. *Br J Cancer* 2019; **121:** 285–92.
24 Koch HW, Larsen M, Bartsch H, Kurz KD, Hofvind S. Artificial intelligence in BreastScreen Norway: a retrospective analysis of a cancer-enriched sample including 1254 breast cancer cases. *Eur Radiol* 2023; **33:** 3735–43.
25 Pestalozzi BC, Zahrieh D, Mallon E, et al. Distinct clinical and prognostic features of infiltrating lobular carcinoma of the breast: combined results of 15 International Breast Cancer Study Group clinical trials. *J Clin Oncol* 2008; **26:** 3006–14.
26 Evans KK, Birdwell RL, Wolfe JM. If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. *PLoS One* 2013; **8:** e64366.
27 Domingo L, Hofvind S, Hubbard RA, et al. Cross-national comparison of screening mammography accuracy measures in U.S., Norway, and Spain. *Eur Radiol* 2016; **26:** 2520–28.
28 Socialstyrelsen. Nationell utvärdering – bröstcancerscreening med mammografi. 2022. https://www.socialstyrelsen.se/globalassets/sharepoint-dokument/artikelkatalog/nationella-riktlinjer/2022-6-7958.pdf (accessed April 5, 2024).
29 Vargas-Palacios A, Sharma N, Sagoo GS. Cost-effectiveness requirements for implementing artificial intelligence technology in the Women's UK Breast Cancer Screening service. *Nat Commun* 2023; **14:** 6110.
30 Khan SA, Hernandez-Villafuerte K, Hernandez D, Schlander M. Estimation of the stage-wise costs of breast cancer in Germany using a modeling approach. *Front Public Health* 2023; **10:** 946544.