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

Kristina Lång[1,2], Viktoria Josefsson[1,2], Anna-Maria Larsson[3], Stefan Larsson[4], Charlotte Högberg[4], Hanna Sartor[1,2], Solveig Hofvind[5,6], Ingvar Andersson[1,2], Aldana Rosso[1]

[1] Division of Diagnostic Radiology, Department of Translational Medicine, Lund University, Malmö, Sweden

[2] Unilabs Mammography Unit, Skåne University Hospital, Malmö, Sweden

[3] Division of Oncology, Department of Clinical Sciences, Lund University, Lund, Sweden

[4] Department of Technology and Society, Lund University, Lund, Sweden

[5] Section for Breast Cancer Screening, Cancer Registry of Norway, Oslo, Norway

[6] Health and Care Sciences, Faculty of Health Sciences, The Artic University of Norway, Tromsø, Norway

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

Research in context

Evidence before this study

We searched MEDLINE for studies published in English between Jan 1, 2015, and Dec 31, 2020, that included “breast cancer screening” or “mammography screening”, and “artificial intelligence” or “machine learning” in the title or abstract. No prospective trials were identified. There were several retrospective accuracy studies using screening data or enriched datasets. We found no systematic reviews on test accuracy. The retrospective studies, using different artificial intelligence (AI) software and mammography devices, indicated that AI could be used to differentiate between screening examinations with low and high probability of malignancy, which could potentially be used to improve the efficacy of screening and reduce the workload, especially the requirement for double reading.

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 first report, the objective was to assess the safety of an AI-supported screen-reading procedure, involving triage and detection support. AI-supported screening resulted in 20% more cancers being detected and exceeded the lowest acceptable

limit for safety compared with standard double reading without AI, without affecting the false positive rate. The AI supported screen-reading procedure enabled a 44.3% reduction in the screen-reading workload. The results indicate that the proposed screening strategy is safe.

Implications of all the available evidence

The results from this randomised trial support the findings of earlier retrospective studies, indicating a general potential of AI to improve screening efficacy and reduce workload. The clinical safety analysis concludes that the AI-supported screen-reading procedure can be considered safe. Implementation of AI in clinical practice to reduce the screen-reading workload could therefore be considered to help address workforce shortages. The assessment of the primary endpoint of interval cancer rate, together with a characterisation of detected cancers in the entire study population, will provide further insight into the efficacy of screening, possible side-effects such as overdiagnosis, and the prognostic implications of using AI in mammography screening, taking cost-effectiveness into account.

Summary

Background Retrospective studies have shown promising results using artificial intelligence (AI) to improve mammography screening accuracy and reduce screen-reading workload; however, to our knowledge, a randomised trial has not yet been conducted. We aimed to assess the clinical safety of an AI-supported screen-reading protocol compared with standard screen reading by radiologists following mammography.

Methods In this randomised, controlled, population-based trial, women aged 40–80 years eligible for mammography screening (including general screening with 1.5–2-year intervals and annual screening for those with moderate hereditary risk of breast cancer or a history of breast cancer) at four screening sites in Sweden were informed about the study as part of the screening invitation. Those who did not opt out were randomly allocated (1:1) to AI-supported screening (intervention group) or standard double reading without AI (control group). Screening examinations were automatically randomised by the Picture Archive and Communications System with a pseudo-random number generator after image acquisition. The participants and the radiographers acquiring the screening examinations, but not the radiologists reading the screening examinations, were masked to study group allocation. The AI system (Transpara version 1.7.0) provided an examination-based malignancy risk score on a 10-level scale that was used to triage screening examinations to single reading (score 1–9)

or double reading (score 10), with AI risk scores (for all examinations) and computer-aided detection marks (for examinations with risk score 8–10) available to the radiologists doing the screen reading. Here we report the prespecified clinical safety analysis, to be done after 80 000 women were enrolled, to assess the secondary outcome measures of early screening performance (cancer detection rate, recall rate, false positive rate, positive predictive value [PPV] of recall, and type of cancer detected [invasive or in situ]) and screen-reading workload. Analyses were done in the modified intention-to-treat population (ie, all women randomly assigned to a group with one complete screening examination, excluding women recalled due to enlarged lymph nodes diagnosed with lymphoma). The lowest acceptable limit for safety in the intervention group was a cancer detection rate of more than 3 per 1000 participants screened. The trial is registered with ClinicalTrials.gov, NCT04838756, and is closed to accrual; follow-up is ongoing to assess the primary endpoint of the trial, interval cancer rate.

Findings Between April 12, 2021, and July 28, 2022, 80 033 women were randomly assigned to AI-supported screening (n=40 003) or double reading without AI (n=40 030). 13 women were excluded from the analysis. The median age was 54.0 years (IQR 46.7–63.9). Race and ethnicity data were not collected. AI-supported screening among 39 996 participants resulted in 244 screen-detected cancers, 861 recalls, and a total of 46 345 screen readings. Standard screening among 40 024 participants resulted in 203 screen-detected cancers, 817 recalls, and a total of 83 231 screen readings. Cancer detection rates were 6.1 (95% CI 5.4–6.9) per 1000 screened participants in the intervention group, above the lowest acceptable limit for safety, and 5.1 (4.4–5.8) per 1000 in the control group—a ratio of 1.2 (95% CI 1.0–1.5; p=0.052). Recall rates were 2.2% (95% CI 2.0–2.3) in the intervention group and 2.0% (1.9–2.2) in the control group. The false positive rate was 1.5% (95% CI 1.4–1.7) in both groups. The PPV of recall was 28.3% (95% CI 25.3–31.5) in the intervention group and 24.8% (21.9–28.0) in the control group. In the intervention group, 184 (75%) of 244 cancers detected were invasive and 60 (25%) were in situ; in the control group, 165 (81%) of 203 cancers were invasive and 38 (19%) were in situ. The screen-reading workload was reduced by 44.3% using AI.

Interpretation AI-supported mammography screening resulted in a similar cancer detection rate compared with standard double reading, with a substantially lower screen-reading workload, indicating that the use of AI in mammography screening is safe. The trial was thus not halted and the primary endpoint of interval cancer rate will be assessed in 100 000 enrolled participants after 2-years of follow up.

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

Introduction

European guidelines recommend double reading of screening mammograms to ensure high sensitivity.[1] A meta-analysis suggested that double reading resulted in 0.44 more cancers being detected per 1000 people screened than with single reading;[2] however, this comes at the expense of a large screen-reading workload and can potentially increase false positives.[3,4] Double reading can also be difficult to sustain because of a shortage of breast radiologists in many countries.[5] In addition, despite double reading, some cancers might be missed and diagnosed as interval cancers.[6] Interval cancers generally have a worse prognosis than screen-detected cancers, and the interval cancer rate is therefore an important indicator of screening efficacy.[1,6] In retrospective studies, about 20–30% of interval cancers have been shown to display highly suspicious signs of malignancy at the preceding screening mammogram,[6–8] suggesting that mammography alone could have been sufficient for detection—ie, without the need for supplementary imaging methods. Establishing a more efficient and effective mammography screening programme is therefore warranted.

Recently developed image analysis tools based on artificial intelligence (AI) have promising applications in mammography screening, such as facilitating triage of screening examinations according to risk of malignancy or supporting detection with computer-aided detection (CAD) marks highlighting suspicious findings.[9] Retrospective studies suggest that the accuracy of AI is similar to or better than that of breast radiologists.[10–13] AI has also been shown to be able to identify examinations that were normal (ie, true negatives), and, since the vast majority of women who attend screening do not have breast cancer, adapting single and double reading to AI risk scores could allow more efficient screen reading.[14–17] Additionally, AI has been shown to retrospectively classify screening examinations as high risk before a diagnosis of interval cancer, and could, therefore, help radiologists to reduce false negative screening results when used as detection support.[16,18,19] Taken together, the evidence suggests that use of AI could potentially benefit mammography screening by reducing the screen- reading workload and the number of interval cancers, but randomised trials are needed to assess the efficacy of AI-supported screening.[13]

In the randomised, controlled Mammography Screening with Artificial Intelligence trial (MASAI), we investigate an AI-supported screen-reading procedure involving triage of screening examinations to single or double reading, along with detection support. Here we report a prespecified safety analysis, the objective of which was to assess the safety of using AI-supported screening compared with standard double reading by determining the effect on cancer detection, which could be used to inform new trials or programme-based evaluations. In addition, we compared recalls, false positives, positive predictive value of recalls, and screen-reading workload for the two screen-reading procedures.

Methods

Study design and participants

The MASAI trial was designed as a randomised, parallel, non-inferiority, single-blinded, controlled, screening accuracy study to compare AI-supported mammography screening with standard double reading without AI. The study was done within the Swedish national screening programme and participants were recruited at four screening sites in southwest Sweden (Malmö, Lund, Landskrona, and Trelleborg). Screen reading and further assessment of recalled participants were done at a single site, the Unilabs Mammography Unit at Skåne University Hospital (Malmö, Sweden).

The inclusion criterion was women (defined here as people registered with a female Swedish personal identity number indicating female gender, which can include trans women who have changed their legal gender) eligible to participate in population-based mammography screening, which also includes those with moderate hereditary risk of breast cancer and those with a history of breast cancer. No exclusion criteria were applied. The Swedish population-based mammography screening programme invites women aged 40–74 years for screening at intervals of 1.5–2 years. Those younger than 55 years are first screened at 1.5-year intervals, and those aged 55 years or older are screened at 2-year intervals. Annual screening is done for people considered to have a moderate hereditary risk of breast cancer (lifetime risk 18–29%) and for those with a history of breast cancer (for 10 years after surgery, with an upper age limit of 80 years). Information about the study was included in 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. People 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. Information about the race or ethnicity of participants was not collected.

The study was approved by the Swedish Ethical Review Authority (2020-04936), which also waived the need for written informed consent. The study protocol (versions 1.1 and 1.2) and the statistical analysis plan are available at the Lund University website. The protocol was updated to improve clarity; there were no changes in the trial procedures nor 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. The people screened and the radiographers acquiring the screening examinations were masked to study group allocation, since the automatic randomisation was not visible on the radiographer's PACS interface. The screen readers were not masked to the results of the allocation.

Procedures

A single-vendor mammography system was used for the screening examinations (Senographe Pristina, GE Healthcare, Freiburg, Germany). Standard screening examination included two views per breast with the addition of implant-displacement views for people with breast implants. The examinations randomised to the intervention group were analysed using Transpara version 1.7.0 (ScreenPoint Medical, Nijmegen, Netherlands). This system uses deep learning to identify and interpret mammographic regions suspicious for cancer. It was developed with more than 200 000 examinations for training and testing, which were obtained from multiple institutions in more than ten countries covering a range of populations, modality manufacturers, and variations in screening and diagnostic workflows. Annotations of more than 10 000 cancers in the database are based on biopsy results and include regions marked in previous mammograms in which cancers were visible but not detected by radiologists.

The AI system provided an examination-based malignancy risk score on a continuous scale ranging from 1 to 10. The risk scores were also presented on a discrete 10-level scale, calibrated to assign approximately a tenth of screening examinations to each risk score. Examinations were considered to be low risk (risk score 1–7), intermediate risk (risk scores 8 and 9), or high risk (risk score 10). Cancer prevalence increases sharply in the group with a risk score of 10, and retrospective studies using the same AI version as in this trial have reported 87–90% of screen-detected cancers and 45% of interval cancers to be in this group.[16,17] The AI system also provided CAD marks at suspicious regional findings of calcifications and soft-tissue lesions, with a regional risk score on a discrete scale from 1 to 98. To limit the number of CAD marks that could potentially disturb the screen reading or lead to an increase in false positives, the AI system was preconfigured for CAD marks to be available only for examinations with risk scores of 8, 9, and 10, accounting for approximately 30% of all examinations (regional risk score threshold >42). The AI system was also configured to analyse implant-displacement views in screening examinations of people with breast implants. The PACS was customised with separate worklists for single and double reading. Examinations with the highest 1% risk, classified as extra high risk, were flagged in the high-risk worklist as 10H. A risk score threshold of 9.8, which was determined from the observed risk score distribution in the screening population, was used to select this group. Screening examinations in the control group were not analysed with AI at any timepoint.

In the intervention group, examinations with risk scores of 1–9 (low and intermediate risk) underwent single reading and examinations with risk scores of 10 (high risk) underwent double reading (figure 1). Double reading was done by two different breast radiologists. The second reader had access to the first reader's assessment (unblinded double reading), which is the standard of care in the regional screening programme in the Skåne region. Readers were aware of the examination risk score (for all examinations), presented both in the PACS worklists and on the image monitor. Readers first read the examination without CAD marks and then with CAD marks, if available (ie, for examinations with risk scores of 8–10). The readers were instructed to recall cases with the highest 1% risk, except for obvious false positives. In the control group, screening examinations were read with standard unblinded double reading without AI. The outcomes of the screen reading were either no suspicion of malignancy or recall. Participants could be recalled due to mammographic findings or self-

reported symptoms. Current practice in the screening programme is to recall participants with self-reported symptoms, such as a lump, when the mammogram cannot safely be classified as normal. Before the final decision, readers had the option of referral to a consensus meeting or to a technical recall (eg, due to poor image quality), or both. Consensus meetings are common practice in screening programmes with double reading; in these meetings, difficult or equivocal findings are reassessed by two radiologists, with a joint decision made to recall or clear of suspicion of malignancy.[3] The images acquired at technical recall were by default randomised de novo due to the technical setup; however, participants were assessed according to their originally assigned group. Screening examinations allocated to the intervention group that failed to be processed by AI underwent standard-of-care reading.

16 breast radiologists at the Unilabs Mammography Unit at Skåne University Hospital were involved in the screen reading, of whom 15 had more than 2 years of experience in breast imaging and 14 had more than 5 years of experience. 12 of the radiologists had a yearly reading volume of at least 5000 cases. Three radiologists had a yearly reading volume of 1000–3000 cases, and one radiologist read on average 700 cases per year. Based on the group composition, only readers with more than 2 years of experience were allowed to read from the single-reading worklist. Before each screen-reading session, the radiologist rolled a six-sided die to randomly allocate themselves to either of the two groups: numbers 1–3 allocated them to the control group and 4–6 to the intervention group.

Participants could withdraw from the study at any time, at which point all personal data would be removed and they would be excluded from analyses. True positive cases were initially identified through linkage with the Regional Cancer Registry (on Sept 12, 2022); to compensate for a delay in registry reporting, all recalled participants were manually assessed with use of patient records, and true positives were validated by histopathology reports on surgical samples or core-needle biopsies.

Outcomes

The primary outcome measure of the MASAI trial is interval cancer rate, which will be assessed after the full study population of 100 000 screened participants have had at least a 2-year follow-up (estimated December, 2024). Secondary outcome measures are early screening performance (cancer detection rate, recall rate, false positive rate, and positive predictive value [PPV] of

recall), screen-reading workload (number of screen-readings and consensus meetings), detection in relation to tumour type and stage, proportion of interval cancers by cancer type and stage, sensitivity and specificity, and incremental cost-effectiveness ratio. In the current clinical safety analysis, the secondary outcome measures of early screening performance of cancer detection rate (number of cancers detected per 1000 participants screened), recall rate (proportion of screened participants who were recalled), false positive rate, PPV of recall, type of detected cancer (invasive or in situ), and screen-reading workload were assessed. The screen-reading workload was reported as the sum of all screen readings, including those made at consensus meetings. The number and proportion of screenings that resulted in a consensus meetings (consensus meeting rate) were also reported separately.

Statistical analysis

The intention-to-treat population comprised all participants who underwent breast screening. The modified intention- to-treat (mITT) population comprised participants with a complete screening examination, excluding those who were asked to attend a technical recall but did not attend. Participants recalled due to bilateral enlarged lymph nodes and diagnosed with lymphoma were also excluded from the mITT population, since they were not recalled due to suspicion of breast cancer. Participants were analysed in their allocated group regardless of the actual reading procedure (treatment policy strategy).

The hypothesis for the primary analysis was the non- inferiority of AI-supported mammography screening compared with standard double reading, in terms of interval cancer rate, with a secondary hypothesis of superiority. Considering the interplay of screen-reading workload and the number of interval cancers, the non- inferiority margin for the primary endpoint was set at the intervention yielding at most 20% more interval cancers than in the control group. The sample size calculations were done with use of Fisher's exact test to compare the risk ratio based on the observed interval cancer rate in the current screening programme. A total sample size of 100 000 (intention-to-treat population) was expected to have at least 80% statistical power to show that the ratio of the interval cancer rate is at most 1.2 in the intervention group compared with the control group. The mITT population among the 100 000 enrolled participants will be used in the primary analysis. The sample size calculation for the clinical safety analysis was based on a worst case scenario of the intervention yielding a cancer detection rate of 3 per 1000 screened participants (based on the

assumption that single reading could lead to reduced detection), at which rate the study could be halted, compared with a detection rate of 5 per 1000 screened participants in the control group (reflecting the observed rate in the current screening programme). According to Fisher's exact test, a sample size of 80 000 (intention-to-treat population) was needed to show with a power greater than 80% that the proportion of detected cancer did not reach the worst case scenario. The mITT population among the 80 000 enrolled participants was used in the clinical safety analysis. Throughout the study, the overall recall rate was monitored as part of the institutional quality assurance reports to ensure that the recall rate did not drop below what was observed in the clinic 6 months before the start of the trial (average recall rate 2.1%), which could indicate a reduction in cancer detection. The number of enrolled participants was monitored monthly. Trial data were extracted from PACS on Sept 12, 2022, which was 1.5 months after roughly 80 000 participants had been enrolled, to ensure sufficient time for the screen reading and initial investigations of recalled participants. Participants were matched with the Regional Cancer Registry on the same day as data extraction and a rapid preliminary analysis of cancer detection was available 1 week later, which was used to inform the decision to continue the trial.

Descriptive statistics were used to summarise study population characteristics. Frequencies and percentages were calculated for categorical data. 95% CIs were calculated with the Clopper-Pearson method. The cancer detection rate, recall rate, false positive rate, and PPV of recall were calculated separately for the intervention and control groups. The cancer detection rate was compared with Fisher's exact test and the ratio of the proportions with corresponding 95% CIs were reported. A two-sided p value of less than 0.05 was considered to indicate statistical significance. The numbers of screen readings and consensus meetings were calculated separately for the intervention and the control groups.

In a post-hoc analysis, the distribution of AI risk scores by screening examinations, screen-detected cancers, recalls, and PPV of recall were reported with descriptive statistics. This analysis was included to describe AI performance and no inferential statistical analyses were done.

All statistical analyses were done with Stata IC 17.0 software and Python 3.8.5. The trial is registered with ClincialTrials.gov, NCT04838756.

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

Between April 12, 2021, and July 28, 2022, 80 160 women presented for screening and 127 (0.2%) opted out of the trial. 80 033 participants were randomly assigned: 40 003 (50.0%) to undergo AI-supported screening (intervention group) and 40 030 (50.0%) to undergo double reading without AI (control group). 39 996 participants in the intervention group and 40 024 in the control group were included in the clinical safety analysis (mITT population; figure 2). The median age of participants in the mITT population was 54.0 years (IQR 46.7–63.9). The age distribution and indication for screening was similar between groups (table 1). The AI system did not to provide a risk score for 306 (0.8%) of 39 996 participants in the intervention group. There were 38 (0.1%) technical recalls among 40 003 participants in the intervention group and 46 (0.1%) among 40 030 participants in the control group.

Early screening performance and workload measures are presented in table 2. Based on the rapid preliminary analysis of cancer detection internally reported on Sept 20, 2022, among the 39 996 participants screened with AI, 244 cancers were detected and 861 participants were recalled. Among the 40 024 participants in the control group, 203 cancers were detected and 817 participants were recalled. The cancer detection rate was 6.1 (95% CI 5.4–6.9) per 1000 participants for AI-supported screening (ie, above the lower safety limit) and 5.1 (4.4–5.8) per 1000 for double reading without AI, a ratio of 1.2 (95% CI 1.0–1.5; p=0.052). The absolute difference in cancer detection per 1000 screened participants was 1.0 (95% CI 0.0–2.1). The false positive rate was the same in both groups. 36 886 fewer screen readings were done in the intervention group than in the control group, representing in a 44.3% reduction in the screen-reading workload.

Of the 244 cancers detected in the intervention group, 184 (75%) were invasive, among which 152 (83%) were stage T1 (tumour diameter ≤20 mm). In the control group, 165 (81%) of 203

cancers were invasive, of which 129 (78%) were stage T1. In situ cancers constituted 60 (25%) detected cancers in the intervention group and 38 (19%) in the control group.

The distribution of AI risk scores in the intervention group and early screening performance measures per risk score are presented in table 3 (post hoc). The cancer detection rate in the high-risk group (ie, those with a risk score of 10 that underwent double reading) was 72.3 per 1000 participants screened (208 of 2875 participants), a frequency of one cancer per 14 screening examinations. In the high-risk group, 11 (2.6%) of 416 recalls were due to self-reported symptoms. Of the 490 screening examinations flagged as extra high risk by AI (highest 1% risk), 189 (38.6%) were recalled—ie, 22.0% of all 861 recalls in the intervention group. Of the 189 recalled participants classified as being extra high risk, 136 had cancer (PPV of recall 72.0%), resulting in a cancer detection rate of 277.6 per 1000 screening examinations in the extra- high-risk category. Thus, the 1.2% of screening examinations flagged as extra high risk contained 55.7% of all screen-detected cancers in this group. 36 815 (92.0%) of 39 996 screening examinations were those with risk scores of 1–9 (which underwent single reading), among which there were 440 (1.2%) recalls (51.1% of all 861 recalls), including 114 (25.9%) recalls based on self-reported symptoms. 36 cancers were detected by screening in the single-reading with AI group (14.8% of all 244 screen-detected cancers), with an overall cancer detection rate of 1.0 per 1000 participants screened. There was a considerable difference in cancer detection rate between those with risk scores of 1–7 (0.2 per 1000 participants screened; six cancers detected among 30 464 participants) and those with risk scores of 8–9 (4.7 per 1000; 30 cancers detected among 6351 participants), meaning that, to detect one cancer, radiologists had to read 5000 examinations in the group with scores of 1–7 and 212 examinations in the group with scores of 8–9.

Discussion

This clinical safety analysis showed that a screen-reading procedure using an AI tool to triage screening examinations to single or double reading and with use of AI as detection support in mammography screening was safe, because the cancer detection rate (6.1 per 1000 participants screened) was above the prespecified lower limit for safety, and was similar to that of double reading without AI (5.1 per 1000). The use of AI did not influence the rates of recalls, false positives, or consensus meetings, while the screen-reading workload was reduced by almost half.

The MASAI trial aims to answer two key questions on the use of AI in mammography screening. The first question regards whether AI can be safely used to reduce the screen-reading workload with sustained performance, which we addressed in the current report. The second key question regards the effect on screening outcome, with a primary outcome of interval cancer rate, a central indicator of screening efficacy.[1,6] The full study population of 100 000 screened participants and a 2-year follow-up is needed to investigate this endpoint.

The MASAI trial proposes one of several possible strategies of integrating AI in the screen-reading pathway.[13,20] In European screening programmes, in which double reading is standard, AI has been suggested to replace one of the readers, to be used as a standalone reader for examinations with low AI risk scores, to force examinations with high AI risk scores to a consensus meeting or to arbitration, or to automatically recall cases above a specific threshold. Different strategies can also be considered regarding reader access to AI information: having it available at the time of screening or, for example, only at the consensus meeting to limit automation bias. Our strategy was to use AI to triage examinations to single or double reading and to let radiologists have access to AI information in the form of risk scores and CAD marks at the time of screen reading. The rationale underlying this design was to take advantage of the bias introduced by AI. We hypothesised that, in addition to the benefit of CAD marks as detection support, knowledge of disease prevalence would influence radiologists' operator point and thereby reduce false positives when reading low-risk examinations (in addition to single reading itself leading to fewer false positives[4]) and reduce false negatives when reading high-risk examinations.[21] Access to risk scores and CAD marks did not seem to introduce a detrimental automation bias, since the false positive rate remained unchanged. This finding emphasises the importance of radiologists having the final decision to recall, which, besides reducing false positives, constitutes a practical approach to meeting established medicolegal requirements, as opposed to the current ethical and legal uncertainties of using AI as a standalone reader. However, if results from prospective studies show that use of AI in screen reading is safe, it could potentially lead to over- reliance on AI and cause an increased risk of detrimental automation bias over time. In our study, access to AI information also enabled the fast and safe handling of screening examinations with a very high probability of cancer. These examinations were flagged in the PACS worklist and could therefore be prioritised for a timely and scrutinised screen reading. This approach was effective as indicated by the findings that

the 1.2% of screening examinations flagged as extra high risk contained 55.7% of all screen-detected cancers in this group.

AI-supported screening resulted in 20% more cancers (244 vs 203) being detected than with standard screening. 152 stage T1 invasive cancers were detected in the intervention group compared with 129 in the control group, which might indicate an increase in early detection without the need for supplementary imaging methods. The incremental increase was, however, not as large as that observed with digital breast tomosynthesis screening in a previous study.[22] Still, the higher cancer detection with tomosynthesis compared with mammography in screening has not convincingly been shown to translate into a reduction of interval cancers,[22] which could question its clinical importance since it is also a more resource-demanding technique. The clinical significance of the additional detected invasive cancers in our study remains to evaluated. The evolution of AI over time could affect all available tests for breast cancer screening, but the use of AI in tomosynthesis screening has not yet been evaluated in a prospective study.

We also found increased detection of in situ cancers with AI-supported screening compared with standard screening (60 vs 38), which could be concerning in terms of overdiagnosis. The risk of overtreating an in situ cancer is more likely with low-grade cancers, since they might never progress into a clinically relevant event during the patient's lifetime.[23] Hence, the planned characterisation of detected cancers in the full study population will bring some clarity to possible overdiagnosis with AI-supported screening. Fenton and colleagues showed a 34% increase in the detection of in situ cancers (from 1.17 to 1.57 per 1000 screening mammograms, p=0.09) after the implementation of conventional CAD in screening but without a parallel increase in the detection of invasive cancer.[24] Conventional CAD was also shown to increase false positives and related costs, and its use in screening could ultimately not be justified.[2,24–27] AI thus seems to have improved performance compared with that of conventional CAD, but could still have hypersensitivity to calcifications, a typical presentation of in situ cancers.[13] Subsequent screening will show whether the relatively higher detection observed in our trial is a result of screening with a more sensitive technique for the first time (ie, a prevalence effect), causing an initial high incidence that levels out during subsequent screening rounds.[28]

We found that the benefit of AI-supported screening in terms of screen-reading workload

reduction was considerable. The actual time saved was not measured, but, if we assume that a radiologist reads on average 50 screening examinations per hour, it would have taken one radiologist 4.6 months less to read the 46 345 screening examinations in the intervention group compared with the 83 231 in the control group. There was concern about whether AI would lead to an increase in cases referred to consensus meetings, considering the eventual need to discuss CAD findings and the possible reader anxiety arising from single reading. Consensus meetings constitute an important step to increase the specificity, but are resource demanding.[3] Contrary to expectations, the proportion of screenings that led to a consensus meeting was not affected by the use of AI.

These results are promising and can be used to inform new trials and programme-based evaluations to address the radiologist shortage. However, we still need to improve our understanding of what the implications are for patient outcome—most importantly, the effect on interval cancer rates. We also need to investigate whether the higher detection of small invasive cancers will lead to a subsequent reduction of prognostically significant cancers and whether the frequency of in situ cancers detected will be reduced at subsequent screenings. An analysis of the prognostic characteristics of cancers detected in the full study population of the MASAI trial is underway. Furthermore, AI systems come at a financial cost, and, while the market and its business models might develop, the willingness to pay to reduce the workload must be determined. Cost- effectiveness can be determined only when the downstream cost of the intervention has been assessed.

The MASAI trial is, to our knowledge, the first randomised trial investigating AI in mammography screening and can thus provide evidence for clinical implementation. A strength of the study is the close resemblance to a real screening setting, since no exclusion criteria were applied and few people invited for screening opted out. The main limitation of this study is that it was conducted at a single centre in the Swedish screening programme. The study was also limited to the combination of one type of mammography device and one AI system. AI system performance will inevitably vary with technical factors such as AI algorithms and image processing,[12] but will probably be of less importance than the variability of radiologists. Our screening strategy emphasises the central role of the radiologist to make the final decision to recall a patient, and the present results are dependent on the performance of the participating radiologists. The radiologists participating in this trial were, overall, moderately to highly

experienced in breast imaging, which could also limit the generalisability of our findings to some settings (eg, those with predominantly less- experienced screen readers, which have been shown to have a higher rate of false positives[29]), and the qualification for single reading with AI is likely to require revision. Only readers with more than 2 years of experience were allowed to conduct single reading in the intervention group, which could have introduced a bias in reader performance in relation to the control group. Because only one of the 16 participating radiologists had less than 2 years of experience, we do not believe this factor would have had a major influence on the results. Another limitation of the reported results was that the true false positive rate requires a follow-up period in case of later interval cancer diagnosis. Because this is an uncommon event, we do not expect the false positive rate to change substantially.

Results from this study must be considered in the context of AI being a constantly evolving field with continuously updated algorithms and machine learning models, and the related challenge of the transparency of these updates.[30] To mitigate the implications of these updates and estimate the effect on screening performance, mammograms from the MASAI trial will be stored in a data warehouse to enable a reanalysis with updated algorithms using the trial outcome data as a reference. More importantly, implementing AI systems would require having a monitoring system of algorithm performance in place.

In summary, this clinical safety analysis 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 mammography screening can be considered safe, since it resulted in a similar rate of screen-detected cancer—exceeding the lowest acceptable limit for safety—without increasing rates of recalls, false positives, or consensus meetings, and while substantially reducing the screen-reading workload compared with screening by means of standard double reading.

Contributors

KL and AR conceptualised and designed the trial with input from IA and SH. AR did the statistical analysis. KL, AR, VJ, and HS directly accessed and verified the underlying data reported in the manuscript. All authors were involved in data interpretation. KL wrote the first draft of the report with input from AR. 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 lecture honorarium from AstraZeneca. SH is head of BreastScreen Norway at the Cancer Registry of Norway which has a research agreement with Screenpoint Medical. All other authors declare no competing interests.

Data sharing

De-identified data will be made available upon reasonable request, with investigator support and a signed data access agreement. A proposal should be submitted to be reviewed by the study steering committee. The data are not publicly available due to data protection regulations. The study protocol and statistical analysis plan are available online at https://portal.research.lu.se/en/projects/mammography-screening-with- artificial-intelligence.

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.

References

1 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.

2 Taylor P, Potts HWW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer 2008; 44: 798–807.

3 Taylor-Phillips S, Stinton C. Double reading in breast cancer screening: considerations for policy-making. Br J Radiol 2020; 93: 20190610.

4 Posso M, Carles M, Rué M, Puig T, Bonfill X. Cost-effectiveness of double reading versus single reading of mammograms in a breast cancer screening programme. PLoS One 2016; 11: e0159806.

5 Gulland A. Staff shortages are putting UK breast cancer screening "at risk," survey finds. BMJ 2016; 353: i2350.

6 Houssami N, Hunter K. The epidemiology, radiology and biological characteristics of interval breast cancers in population mammography screening. NPJ Breast Cancer 2017; 3: 12.

7 Hovda T, Hoff SR, Larsen M, Romundstad L, Sahlberg KK, Hofvind S. True and missed interval cancer in organized mammographic screening: a retrospective review study of diagnostic and prior screening mammograms. Acad Radiol 2022;29 (suppl 1): S180–91.

8 Hofvind S, Skaane P, Vitak B, et al. Influence of review design on percentages of missed interval breast cancers: retrospective study of interval cancers in a population-based screening program. Radiology 2005; 237: 437–43.

9 Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: state of the art. Semin Cancer Biol 2021; 72: 214–25.

10 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.

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

12 Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol 2020; 6: 1581–88.

13 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.

14 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. Eur Radiol 2019; 29: 4825–32.

15 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.

16 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.

17 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.

18 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.

19 Byng D, Strauch B, Gnas L, et al. AI-based prevention of interval cancers in a national mammography screening program. Eur J Radiol 2022; 152: 110321.

20 Larsen M, Aglen CF, Hoff SR, Lund-Hanssen H, Hofvind S. Possible strategies for use of artificial intelligence in screen-reading of mammograms, based on retrospective data from 122,969 screening examinations. Eur Radiol 2022; 32: 8238–46.

21 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.

22 Houssami N, Hofvind S, Soerensen AL, et al. Interval breast cancer rates for digital breast tomosynthesis versus digital mammography population screening: an individual participant data meta-analysis. eClinicalMedicine 2021; 34: 100804.

23 van Luijt PA, Heijnsdijk EAM, Fracheboud J, et al. The distribution of ductal

carcinoma in situ (DCIS) grade in 4232 women and its impact on overdiagnosis in breast cancer screening. Breast Cancer Res 2016; 18: 47.

24 Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007; 356: 1399–409.

25 Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015; 175: 1828–37.

26 Elmore JG, Lee CI. Artificial Intelligence in medical imaging: learning from past mistakes in mammography. JAMA Health Forum 2022; 3: e215207.

27 Fenton JJ. Is it time to stop paying for computer-aided mammography? JAMA Intern Med 2015; 175: 1837–38.

28 Gur D, Nishikawa RM, Sumkin JH. New screening technologies and practices: a different approach to estimation of performance improvement by using data from the transition period. Radiology 2015; 275: 9–12.

29 Alberdi RZ, Llanes AB, Ortega RA, et al. Effect of radiologist experience on the risk of false-positive results in breast cancer screening programs. Eur Radiol 2011; 21: 2083–90.

30 Larsson SHF, Heintz F. Transparency in artificial intelligence. Internet Policy Rev 2020; 9: 1469.

Table 1. Baseline population characteristics.

Characteristics	**Intervention** (N = 39 996)	**Control** (N = 40 024)
Age		
Age (mean, SD, min-max) – yr	55·3 (10·2) [39·6, 80·1]	55·3 (10·2) [39·5, 79·9]
<45 – no. (%)	7568 (18·9)	7607 (19·0)
45–49 – no. (%)	7 155 (17·9)	7209 (18·0)
50–54 – no. (%)	6505 (16·3)	6559 (16·4)
55–59 – no. (%)	5021 (12·6)	4822 (12·0)
60–64 – no. (%)	5007 (12·5)	5214 (13·0)
65–69 – no. (%)	4345 (10·9)	4265 (10·7)
>70 – no. (%)	4395 (11·0)	4348 (10·9)
Screening indication		
General screening — no. (%)	38 969 (97·4)	38 951 (97·3)
Previous history of breast cancer — no. (%)	984 (2·5)	1017 (2·5)
Moderate hereditary risk – no. (%)	43 (0·1)	56 (0·1)

Table 2. Early screening performance and workload measures.

	Intervention (N = 39 996)	**Control** (N = 40 024)
Early screening performance		
Recalls – no.	861	817
Recall rate – % (95% CI)	2·2 (2·0, 2·3)	2·0 (1·9, 2·2)
Screen-detected cancers – no.	244	203
Cancer-detection rate per 1000 screens – % (95% CI)	6·1 (5·4, 6·9)	5·1 (4·4, 5·8)
False-positive rate – % (95% CI)	1·5 (1·4, 1·7)	1·5 (1·4, 1·7)
PPV-1 – % (95% CI)	28·3 (25·3, 31·5)	24·8 (21·9, 28·0)
Workload		
Screen readings – no.	46 345	83 231
Consensus meetings – no.	1584	1576
Consensus rate – % (95% CI)	4·0 (3·8, 4·2)	3·9 (3·8, 4·1)

Table 3. Distribution of artificial intelligence examination risk scores and early screening performance measures.

Risk scores	**Screened women** (N = 39 996) No. (%)	**Recalls** (N = 861) No. (%)	**Screen-detected cancers** (N = 244) No. (%)	**Positive-predictive value of recalls** %
10	2875 (7·2)	416 (48·3)	208 (85·2)	50·0
9	3212 (8·0)	116 (13·5)	23 (9·4)	19·8
8	3139 (7·8)	65 (7·5)	7 (2·9)	10·8
7	3075 (7·7)	36 (4·2)	1 (0·4)	2·8
6	3193 (8·0)	41 (4·8)	1 (0·4)	2·4
5	3503 (8·8)	52 (6·0)	0 (0·0)	0·0
4	3697 (9·2)	35 (4·1)	1 (0·4)	2·9
3	4247 (10·6)	30 (3·5)	1 (0·4)	3·3
2	4368 (10·9)	31 (3·6)	1 (0·4)	3·2
1	8381 (21·0)	34 (3·9)	1 (0·4)	2·9
Missing	306 (0·8)	5 (0·6)	0 (0·0)	0·0

Figure legend:

Figure 1: Overview of trial intervention

Figure 2: Trial profile

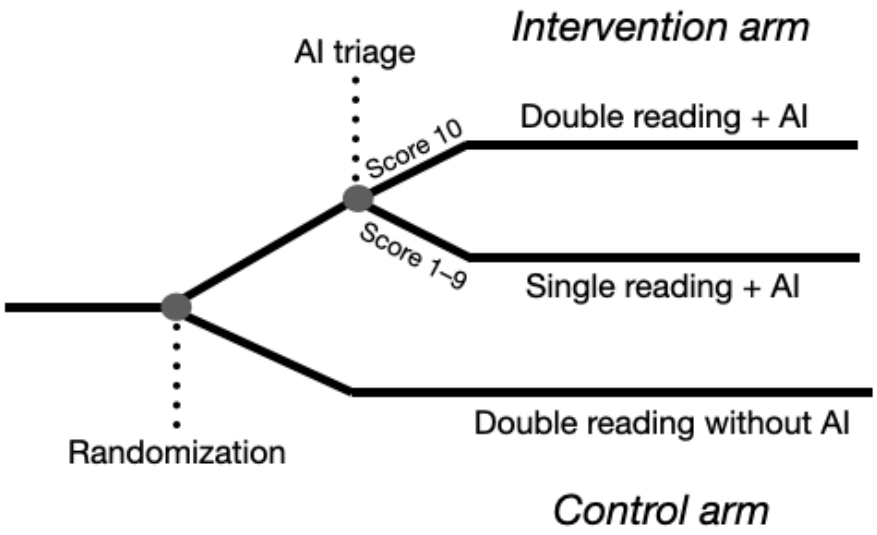
Intervention arm
AI triage
Double reading + AI
Score 10
Score 1–9
Single reading + AI
Double reading without AI
Randomization
Control arm

Figure 1

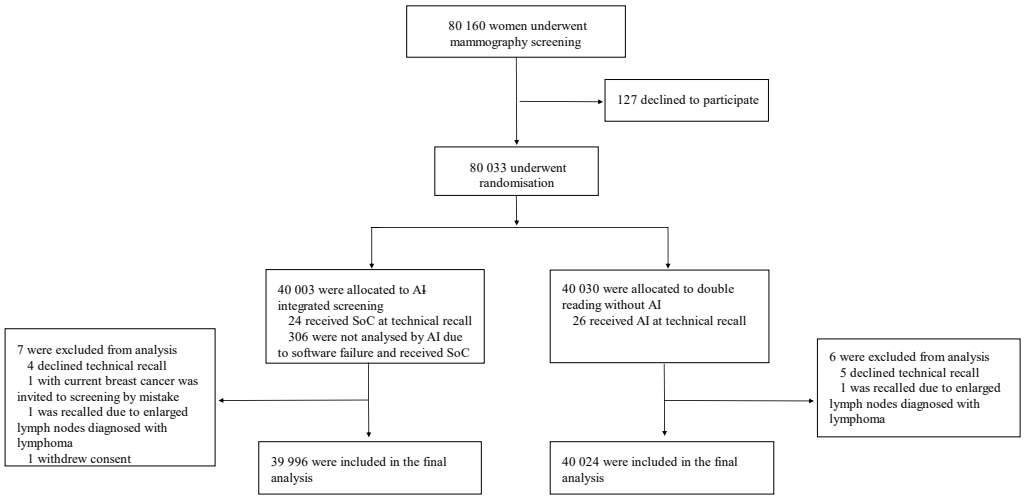
80 160 women underwent mammography screening
127 declined to participate
80 033 underwent randomisation
40 003 were allocated to AI integrated screening
24 received SoC at technical recall
306 were not analysed by AI due to software failure and received SoC
40 030 were allocated to double reading without AI
26 received AI at technical recall
7 were excluded from analysis
4 declined technical recall
1 with current breast cancer was invited to screening by mistake
1 was recalled due to enlarged lymph nodes diagnosed with lymphoma
1 withdrew consent
6 were excluded from analysis
5 declined technical recall
1 was recalled due to enlarged lymph nodes diagnosed with lymphoma
39 996 were included in the final analysis
40 024 were included in the final analysis

Figure 2