CONTENTS

DOCUMENT UPLOADED FOR STUDY REGISTRATION

Official Title: Artificial Intelligence in a Population-based Breast Cancer Screening - the Prospective Clinical Trial ScreenTrust CAD

Unique Protocol ID: STGKS001

NCT ID: 04778670

Document Date: May 16, 2022

Comprehensive Study Protocol

for

Image analysis with artificial intelligence to increase precision in breast cancer screening, the ScreenTrust CAD substudy, a prospective trial of AI as an independent reader of screening mammograms

Contents

World Health Organization Trial Registration Dataset

Primary Registry
The trial will be registered at clinicaltrials.gov before the first study person is included in the study.

Trial Identifying Number
STGKS001 (internal). NCT 04778670

Date of registration in Primay Registry
February 22, 2021

Secondary Identifying Numbers
- Protocol number in agreement between Capio St Göran Hospital and Lunit Inc: STGKS001
- Ethical Review Authority, Sweden, EPM 2020-00487
- Karolinska University Hospital, Sweden, K 2020-0807

Source(s) of Monetary or Material Support
- Lunit Inc., Seoul, South Korea, company
- Capio S:t Göran Hospital, Stockholm, Sweden, hospital

Primary Sponsor
- Capio S:t Göran Hospital

Secondary Sponsor
- Karolinska University Hospital

Contact for Public Queries
- screentruststudy@gmail.com, +46 517 700 00

Contact for Scientific Queries
- Principal Investigator: Fredrik Strand, MD, fredrik.strand@sll.se, +46 517 700 00, Bröstradiologi, NB1:03, Gävlegatan 55, 171 64 Solna, Sweden

Public Title
- Artificial intelligence in large-scale breast cancer screening (ScreenTrust CAD)

Scientific Title
- Artificial intelligence in population-based breast cancer screening – the prospective clinical trial ScreenTrust CAD

Countries of Recruitment
- Sweden

Health Condition or Problem Studied
- Breast cancer

Intervention
- Intervention Name: Artificial intelligence-based cancer detector for mammography images (AI CAD)
- Intervention Description: The AI CAD software has the product name Lunit INSIGHT MMG, which is CE certified. The version of the product that will be initially installed for this study is 1.1.6.1 (updated with future official releases). In the regular mammography screening process at the study hospital two radiologist readers review the image. During the study, AI CAD will be used as an independent third reader of screening mammograms and make a binary decision as do the other two readers (i.e., "flag" if anything suspicious is deemed present in the image, otherwise "healthy"). The AI CAD generates a continuous "score", for which a threshold value has been defined, above which the mammogram should be considered flagged, and below which it should be considered healthy. The threshold value was determined in a historic calibration dataset, in which the cut-off point resulted in a joint sensitivity of AI plus first reader radiologist which was 2% higher than the joint sensitivity of first reader and second reader radiologist. By this method the threshold value was preset to 0.534. The threshold value would not be changed during the course of the study.

Key Inclusion and Exclusion Criteria

All women attending screening mammography at Capio S:t Göran hospital will be considered for inclusion. The ethical review authority has waived the need to obtain individual written informed consent.

- Inclusion criterion: age 40-74 years, a standard four-view mammography examination was acquired (right MLO, right CC, left MLO, left CC)

- Exclusion criteria: having a breast implant (excluded by presence of either CCID or MLOID view position; ID = implant displaced) or having undergone mastectomy (excluded by not having a complete four-view mammography).

- For detailed information on planned statistical analysis – see separate Statistical Analysis Plan (SAP).

Study Type

Type of Study: Interventional study

Study design: All examinations will be reviewed by both human readers and AI CAD. During the course of the study, all mammograms that have been flagged by any of the three (human, human, AI) readers will go to a consensus discussion. The consensus discussion consists of an oral discussion of each flagged case, is held between at least two radiologists and results in a decision for each case to "recall" or to define as "healthy" (and not recall). During the course of this study, the radiologists in the consensus discussion will have additional access to any information contained in the radiology system that has been generated by the AI CAD.

Date of First Enrollment

Study start was planned for March 2021. The first individual was included on April 1, 2021.

Sample Size

The study should enroll study persons with a total of 55,000 screening examinations.

Recruitment Status

Recruitment is on-going since April 1, 2021, and is expected to finish in May or June 2022.

Primary Outcome(s)

Diagnosed pathology-verified breast cancer; to be followed up at three time-points: (i) screen-detected - diagnosed after being recalled at the screening examination, (ii) adding non-screen-detected cancer diagnosed within 12 months after the screening examination and (iii) adding non-screen-detected cancer diagnosed within 23 months after the screening examination, not including the following screening examination (such that all cancers diagnosed over a full screening cycle are captured).

Secondary Outcome(s)

1. Reader flagging [YES|NO], i.e., each reader (radiologist or AI) makes an assessment that there is an abnormal finding warranting the examination to continue to the consensus discussion. Radiologist id codes will be recorded.

2. Consensus recall [YES|NO], i.e., a decision by the consensus discussion to recall the woman for further work-up. Radiologist id codes will be recorded.

3. Tissue sampling [YES|NO], i.e., decision by the radiologist to perform tissue sampling after additional diagnostic imaging of a recalled woman.

4. Process failures in generating the AI score or in transferring the AI score to the appropriate datapoint in the radiological RIS/PACS system.

In addition, several additional exploratory objectives and endpoints exist; see the detailed Statistical Analysis Plan.

Ethics Review

Approved on April 28, 2020, with registration id EPM 2020-00487 by the Ethical Review Authority of Sweden (email: registrator@etikprovning.se, phone: +46 10 475 08 00).

Completion date

Estimated May to June, 2022.

Summary Results

N/A

IPD sharing statement

We plan to share individual participant-level data to the extent that the data can be considered anonymous by the responsible research body. A transfer agreement for academic research purposes will be required.

SPIRIT-AI additional items

Protocol Version

1.1, April 29, 2022.

Funding

See WHO above.

Roles and responsibilities

Protocol contributors

Fredrik Strand, MD PhD, Karolinska University Hospital: Principal Investigator

Martin Eklund, PhD, Karolinska Institute: Biostatistician

Karin Dembrower, MD, Capio S:t Göran hospital: Local project leader

Anders Byström, MD, Capio S:t Göran hospital: Head of radiology department

Ki Hwan Kim, MD, Lunit Inc.: Software provision

Trial sponsor:

See WHO above.

Role of study sponsor and funders

The primary sponsor of the study, Capio S:t Göran hospital, is responsible for collection and management. The PI and secondary sponsor of the study, Karolinska University Hospital, is responsible for study design, analysis and interpretation of data, writing of the report, and the decision to submit the report for publication (the PI will have ultimate authority to decide over these activities). The funder of the study, Lunit Inc., will not have authority over the activities and decisions of the primary sponsor, the secondary sponsor or the PI.

Composition, roles and responsibilities of steering committee

The steering committee consists of: Fredrik Strand, PI (Karolinska University Hospital), Anders Byström, head of radiology (Capio S:t Göran hospital), Maria Kedra, IT manager (Capio S:t Göran hospital), Maria Wijk, legal counsel (Capio S:t Göran hospital), and Karin Dembrower, local project leader and breast radiologist (Capio S:t Göran hospital).

Introduction

Background and rationale

Breast cancer is the most common cancer for women. Though the patients have a relatively good probability of survival, around 1500 women die each year in Sweden from the disease. Mammographic screening has been shown to lower mortality by around 30 (1). However, in the screening programs large resources are consumed and around 30 percent of cancers go undetected and the women find them by noticing a lump in the breast (2). One method to increase the accuracy of screening has been through double-reading whereby two radiologists assess all images (3). This increases accuracy, at the cost of requiring more radiologists.

Over the last years, an increasing number of scientific articles have described successful attempts in using deep learning for mammographic tumor detection and lately also for prediction of future breast cancer and workflow management (4-6). Lately, we have evaluated how three different commercially available AI algorithms would perform as independent readers of screening mammograms within a retrospective cohort from Karolinska University Hospital (7). We found that one algorithm had a sensitivity of 81.9% (95%CI: 78.9 to 84.6%) which was markedly better than the second-best algorithm (sensitivity 67.4%; 95%CI: 63.9 to 70.8%). It was also better than the first-reader radiologist (sensitivity 77.4%; 95% CI: 74.2 to 80.4%). The current planned study will further determine how well the AI algorithm performs in a prospective setting in a true screening population.

1. Weedon-Fekjær H, Romundstad PR, Vatten LJ. Modern mammography screening and breast cancer mortality: population study2014 2014-06-17 22:30:49.
2. Törnberg S, Kemetli L, Ascunce N, Hofvind S, Anttila A, Sèradour B, et al. A pooled analysis of interval cancer rates in six European countries. European journal of cancer prevention. 2010;19(2):87-93.
3. Taylor-Phillips S, Jenkinson D, Stinton C, Wallis MG, Dunn J, Clarke A. Double Reading in Breast Cancer Screening: Cohort Evaluation in the CO-OPS Trial. Radiology. 2018:171010.
4. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. Journal of the National Cancer Institute. 2019.
5. Kyono T, Gilbert FJ, van der Schaar M. Improving Workflow Efficiency for Mammography Using Machine Learning. Journal of the American College of Radiology : JACR. 2019.
6. Ribli D, Horvath A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep. 2018;8(1):4165.
7. Dembrower K, Liu Y, Azizpour H, Eklund M, Smith K, Lindholm P, et al. Comparison of a Deep Learning Risk Score and Standard Mammographic Density Score for Breast Cancer Risk Prediction. Radiology. 2019:190872.
8. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. Radiology. 2019:182716.

Objectives

The overall aim of the project is to examine how an AI CAD computer algorithm can independently assess screening mammograms, in a prospective clinical setting, in order to improve the accuracy of breast cancer screening and/or to replace radiologists.

Trial design

See WHO above, "Study Type"

Study setting

The study will be conducted in Capio S:t Göran hospital, one large breast center in Stockholm, in size it is the second of the three that cover the entire Stockholm area. They invite more than 70,000 to breast cancer screening each year. The breast imaging department has around five to seven dedicated breast radiologists. The x-ray equipment for screening mammography is from Philips and Sectra.

Eligibility criteria, Interventions, Outcomes
See WHO above

Participant timeline
Enrollment will start as soon as all IT systems have been integrated and tested, estimated to March 2021. For each participant there will be a follow-up period of 12 and 23 months for collecting information on both screen-detected cancer and interval cancer (i.e., cancer diagnosed after the current screening and before the next planned one).

Sample size
See WHO above

Recruitment
Recruitment will continue until the target of 55,000 examinations is reached

Sequence generation, allocation concealment mechanism, implementation
Not applicable. All participants will receive intervention.

Blinding
In the initial read, the first radiologist will not have access to AI CAD information. The second radiologist will be instructed to first enter his/her own decision, and then toggle on the AI CAD information. During the consensus discussion, the radiologists will have access to any available information from the AI CAD.

Data collection methods
Data from the radiology system and the AI CAD will be stored by each system and collected at the end of the study, or before any interim analysis. Data on cancer characteristics will be collected from the electronic medical records at the study hospital. Data on all cancer diagnoses will subsequently be collected through linking to the regional cancer registry when data is complete for 23-month follow-up of all study participants. All data will be linked based on the unique national personal identity number of each study person.

Data management
Data will remain stored at the study hospital during the course of the study. When required for interim or final analysis, data will be extracted for research purposes under the responsibility of the PI. This data will be stored following the usual practice in the breast imaging research group at the Karolinska University Hospital, including pseudonymization before any statistical analysis. Each data parameter will undergo type and range checks for validity. Further details on data management can be found in the Data Policy document of the breast imaging research group.

Statistical methods
See separate Statistical Analysis Plan (SAP) document.

Data monitoring
There is no data monitoring committee. This has not been deemed necessary since all continuously collected data will be automatically recorded in the radiology system and electronic medical record of the hospital.

Interim analysis can take place regarding acceptance and compliance for the AI system among medical staff at the hospital. Interim analysis can also take place regarding the frequency of initial reads flagged by AI CAD and flagged by radiologists. Interim analysis will not aim to assess the accuracy of the AI CAD (in order not to influence the level of trust that radiologists put in the AI CAD assessments).

Harms

The intervention is a computer-based processing of mammography images that are acquired according to usual practice, which means no additional harm is possible.

Auditing

No auditing is planned.

Research ethics approval

See WHO above.

Protocol Amendments

Any protocol amendment will be decided by the Steering Committee. If the amendment is deemed to require additional permission by the Ethical Review Authority this will be sought before communication. Then, the amendment will be communicated to the Steering Committee and the Funder, as well as included in the updated study protocol.

Consent or ascent

Not applicable. The Ethics Review Authority has waived the need for individual written informed consent.

Confidentiality

All personal information will be handled according to GDPR and other applicable laws. Data will be pseudonymized before statistical analysis is performed. Pseudonymized data may be made available to the Funder for auditing.

Declaration of interests

The Funder (Lunit Inc.) funds the study through cost-based compensation agreed directly with Capio S:t Göran hospital, and by allowing use of the study device (AI CAD) free of charge. Outside this study, the principal investigator receives fees from the funder for conducting presentations of his work. The principal investigator is entitled to regular salary for work hours from the sponsor hospitals.

Access to data

The final trial dataset will be available for the research team of the principal investigator. Pseudonymized data can be made available for external research audit. Anonymous data may be shared with academic researchers.

Ancillary and post-trial care

Not applicable. Patients are always covered by the national Swedish patient insurance.

Dissemination policy

Investigators plan to communicate findings primarily through original research papers and through participation in professional meetings. In addition, the investigators will communicate with the general public through media and through presentations at patient association gatherings. For research papers, the inclusion of co-authors will follow ICMJE recommendations. We do not intend to use professional writers outside the investigator team. While access may be granted to academic researchers, public access to complete participant-level data will not necessarily be granted. The statistical code may be shared publicly.

Appendices

None.

Figure 1. The actual work flow in the study. If any one of the initial readers (First Radiologist, Second Radiologist or AI) decides to flag for a suspicious finding in the image, the exam is routed to the consensus discussion. The consensus discussion consists of an oral discussion between at least two radiologists (could be same or different as the initial readers) while reviewing the images of each case, and deciding to recall the woman for further diagnostics or to determine that she is healthy.

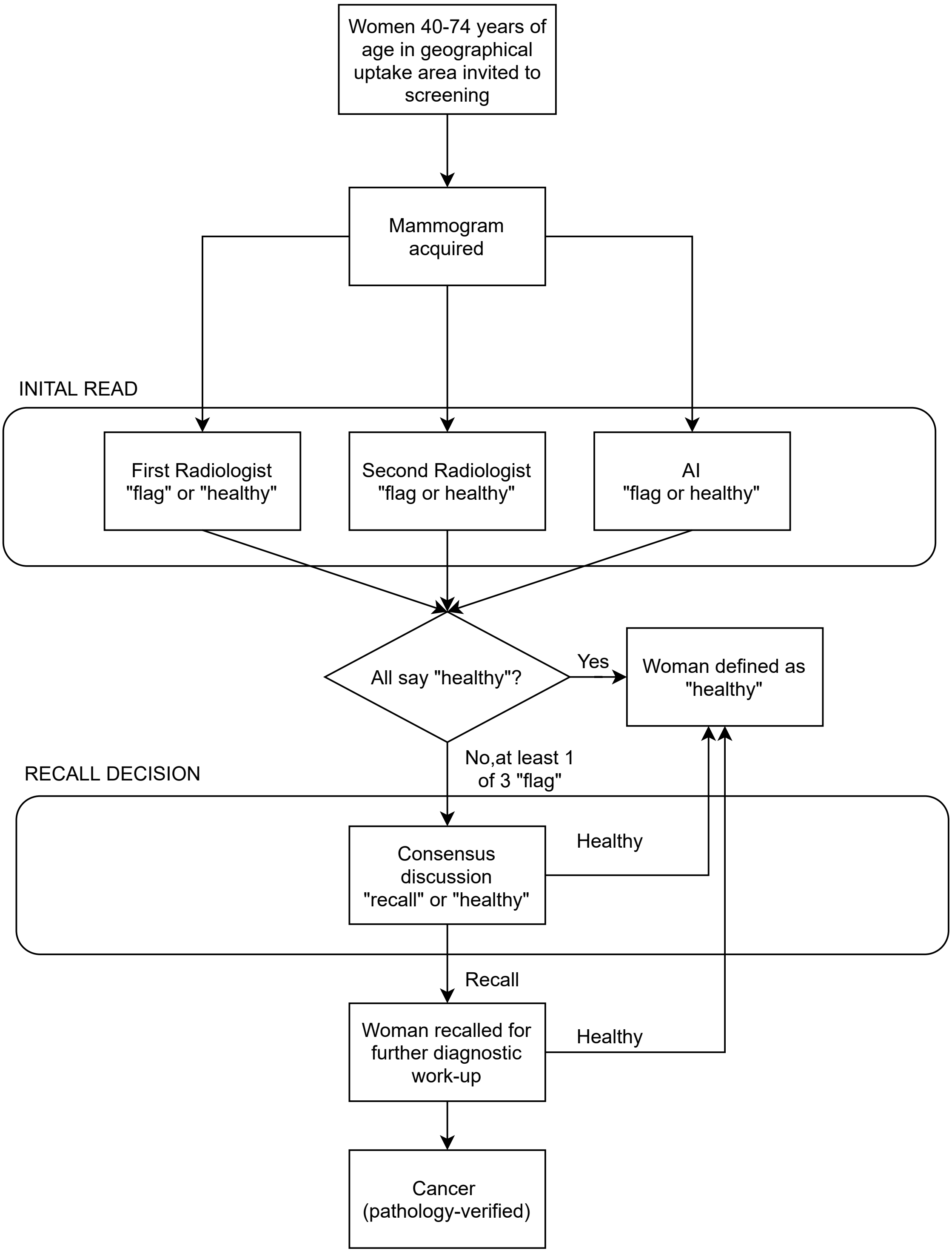
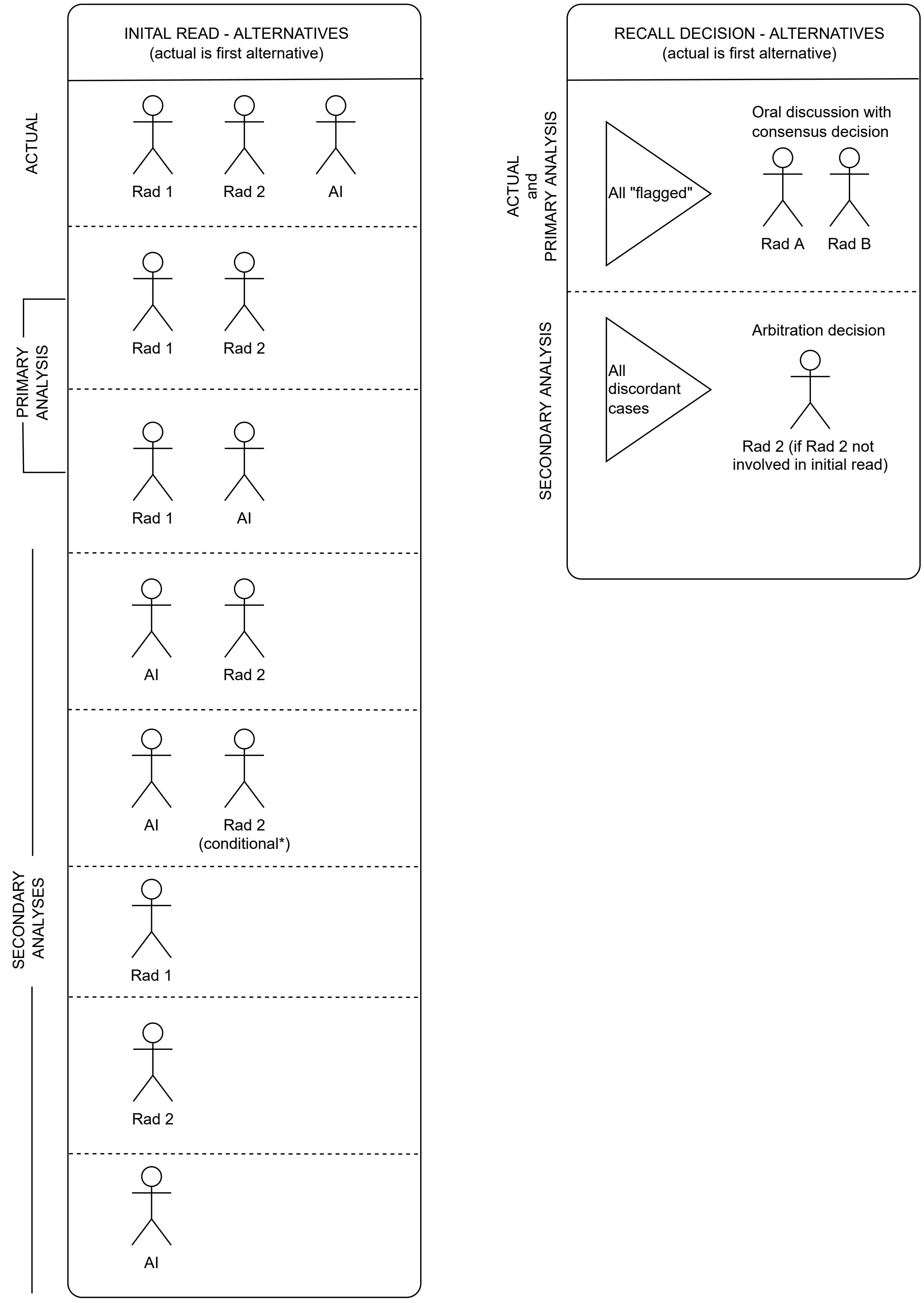

Women 40-74 years of age in geographical uptake area invited to screening
Mammogram acquired
INITAL READ
First Radiologist "flag" or "healthy"
Second Radiologist "flag or healthy"
AI "flag or healthy"
All say "healthy"?
Yes
Woman defined as "healthy"
No,at least 1 of 3 "flag"
RECALL DECISION
Consensus discussion "recall" or "healthy"
Healthy
Recall
Woman recalled for further diagnostic work-up
Healthy
Cancer (pathology-verified)

Figure 2. Alternative scenarios for which outcomes will be analyzed. The alternative at the top of each column is the actual implementation in the study. However, based on the recorded assessments of each radiologist and AI, other alternative scenarios will be analyzed in post-study analysis.


INITAL READ - ALTERNATIVES
(actual is first alternative)
ACTUAL
Rad 1
Rad 2
AI
PRIMARY ANALYSIS
Rad 1
Rad 2
Rad 1
AI
AI
Rad 2
AI
Rad 2 (conditional*)
SECONDARY ANALYSES
Rad 1
Rad 2
AI
RECALL DECISION - ALTERNATIVES
(actual is first alternative)
ACTUAL and PRIMARY ANALYSIS
All "flagged"
Oral discussion with consensus decision
Rad A
Rad B
SECONDARY ANALYSIS
All discordant cases
Arbitration decision
Rad 2 (if Rad 2 not involved in initial read)

STATISTICAL ANALYSIS PLAN (SAP)

DOCUMENT UPLOADED FOR STUDY REGISTRATION

Official Title: Artificial Intelligence in a Population-based Breast Cancer Screening - the Prospective Clinical Trial ScreenTrust CAD

Unique Protocol ID: STGKS001

NCT ID: NCT04778670

Document Date: April 25, 2022

Approvals

ScreenTrust CAD Analysis Plan

Authors	Fredrik Strand, PI
	Martin Eklund, study statistician

Approvals

ScreenTrust CAD Analysis Plan

Authors	Fredrik Strand, PI
	Martin Eklund, study statistician

Revision history

Version	Date	Changes after study start
1.2	May 16, 2022	Added clarification of secondary analyses (4.1 and 5.0) and analyses relying on true and false negatives (5.0, 5.5, and Table 1).
1.1	Apr 29, 2022	Added brief publication plan (1), updated study population definition to be identical to research protocol (3.1), added the numeric value of the abnormality threshold (5.3), added methods to handle and minimize missing data (5.9).
0.9	Feb 22, 2021	Version before study start
0.3	Dec 22, 2020	Draft

1 Preface

This Statistical Analysis Plan (SAP) describes the planned analyses for

Image analysis with artificial intelligence to increase precision in breast cancer screening, the ScreenTrust CAD substudy, a prospective trial of AI as an independent reader of screening mammograms ("ScreenTrust CAD")

Study has been registered at clinicaltrials.gov: 04778670

The planned analyses identified in this SAP will be included in future manuscripts. Exploratory analyses not necessarily identified in this SAP may be performed to support planned analyses. Any post-hoc exploratory or unplanned analyses not specified in this SAP before database lock will be identified as such in manuscripts for publication, and added as amendments to this SAP.

This SAP was written by statistician and principal investigator who were blinded to any assessments already performed by AI or human radiologists, and to outcomes.

The first manuscript for publication after inclusion ends, will focus on the primary endpoint of the study and a screen-positive outcome. Later publication will focus on secondary endpoints and on the primary endpoint at follow-up time 12 months and 23 months after the initial screening.

2 Background

Breast cancer is the most common cancer for women. Though the patients have a relatively good probability of survival, around 1500 women die each year in Sweden from the disease. Mammographic screening has been shown to lower mortality by around 30% (1). However, in breast cancer screening programs large resources are consumed and around 30 percent of cancers go undetected and the women find them by noticing a lump in her breast (2, 3). One method to increase the accuracy of screening has been through double-reading whereby two radiologists assess all images (4). This increases accuracy, at the cost of requiring more radiologists.

Over the last years, an increasing number of scientific articles have described successful attempts in using deep learning for mammographic tumor detection and lately also for prediction of future breast cancer and workflow management (5-7). Lately, we have evaluated how three different commercially available AI algorithms would perform as independent readers of screening mammograms within a retrospective cohort from Karolinska University Hospital (8). We found that one algorithm had a sensitivity of 81.9% (95%CI: 78.9 to 84.6%) which was markedly better than the second-best algorithm (sensitivity 67.4%; 95%CI: 63.9 to 70.8%). It was also better than the first-reader radiologist (sensitivity 77.4%; 95% CI: 74.2 to 80.4%). Positive follow-up findings were determined by pathology-verified diagnosis at screening or within 12 months thereafter. The current planned study will further determine how well the AI algorithm performs in a prospective population-based screening setting.

3 Design

This is a prospective clinical trial following a paired screen-positive design (9), with the aims to assess the performance of an AI algorithm combined with radiologists(s) compared to standard-of-care being two radiologists assessing screening mammograms in a true screening population (Figure 1). The study has paired design. All examinations will receive a flagging decision by: first reader, second reader, and AI. The first reader will not have access to AI. The second reader will first record her/his decision, then review AI to make a joint decision (forced to be flagged if the second reader and/or AI made a flagging decision). In the consensus discussion, two or more radiologists will discuss all flagged cases, having access to AI information. Since all decisions by individual readers will be recorded, it is possible to determine what the outcome would have been had one or two of the readers not been allowed to assess images, and to determine what the outcome would have been had the recall decision been performed by consensus decision (actual) compared to single reader arbitration of discordant cases (Figure 2).

3.1 Study population

All women attending screening mammography at Capio S:t Göran hospital will be considered for inclusion. The ethical review authority has waived the need to obtain individual written informed consent.

- Inclusion criterion: age 40-74 years, a standard four-view mammography examination was acquired (right MLO, right CC, left MLO, left CC)

- Exclusion criteria: having a breast implant (excluded by presence of either CCID or MLOID view position; ID = implant displaced) or having undergone mastectomy (excluded by not having a complete four-view mammography).

3.2 Study period

The study will start as soon as all technical component are in place and integrated, currently predicted to be February 2021. The duration of inclusion of study persons will depend on when 55,000 women has been included in the study, which we assess to be around 1,5 years. UPDATE: The first study person was included on April 1, 2021, and the study is estimated to close inclusions in May or June, 2022. Then, for the final analyses of interval cancer cases there is a need for 23-month follow-up after the last included patient. An overview is shown in Figure 3 below.

4 Objectives and Endpoints

The purpose of the study is to examine the integration of an independent AI computer-based reader into a population-based screening program. First, in the initial read and decision to flag suspicious examinations, we will determine whether AI and one human radiologist is non-inferior to two human radiologists - in terms of detecting breast cancer. Second, we will estimate the operating characteristics of various combinations of AI and radiologists, in the initial read and in the arbitration step, as described in Figure 2.

4.1 Primary endpoint

Diagnosed pathology-verified breast cancer; to be followed up at three time-points: (i) screen-detected - diagnosed after being recalled at the screening examination, (ii) adding non-screen-detected cancer diagnosed within 12 months after the screening examination and (iii) adding non-screen-detected cancer diagnosed within 23 months after the screening examination, not including the following screening examination (such that all cancers diagnosed over a full screening cycle are captured).

4.2 Secondary endpoints

4.2.1 Reader flagging of women evaluated as cancer free [YES|NO], i.e., each reader (radiologist or AI) makes an assessment that there is an abnormal finding warranting the examination to continue to the consensus discussion. Radiologist id codes will be recorded.

4.2.2 Consensus recall of women evaluated as cancer free [YES|NO], i.e., a decision by the consensus discussion to recall the woman for further work-up. Radiologist id codes will be recorded.

4.2.3 Tissue sampling of women evaluated as cancer free [YES|NO], i.e., decision by the radiologist to perform tissue sampling after additional diagnostic imaging of a recalled woman.

4.2.4 Process failures in generating the AI score or in transferring the AI score to the appropriate datapoint in the radiological RIS/PACS system.

4.3 Additional data collected in the study

4.3.1 Patient characteristics

4.3.1.1 All women

- age at diagnosis: median and categorized into (i) 55 years or less and (ii) more than 55 years to approximate pre- and postmenopausal groups.
- mammographic density by automated classification mimicking BI-RADS categories A, B, C, D

4.3.1.2 Diagnosed women

- reduction surgery, lipo-filling, biopsy markers and surgical clips, pacemakers or other non-breast implants.

4.3.2 Cancer characteristics

- invasiveness by categories: in-situ only, invasive
- histology by categories: ductal, lobular, other
- tumor grade by categories: 1, 2, 3
- lymph node metastasis by categories: 0 nodes, 1-3 nodes, 4+ nodes
- tumor size, median size of invasive component and in situ component
- invasive tumor size by binary categories: (i) 0-19 mm, (ii) 20 mm+;
and by five categories: 0-9 mm, 10-19 mm, 20-29 mm, 30-49 mm, 50 mm+
- receptor status ER, PR, HER2 by categories: positive, negative
- proxy molecular subtype defined by:
    Luminal A: ER+/PR+/HER2-; ER+/PR-/HER2-; ER-/PR+/HER2
    Luminal B : ER+/PR+/HER2+; ER+/PR-/HER2+; ER-/PR+/HER2+
    Non-luminal HER2-overexpressing: ER-/PR-/HER2+
    Triple negative: ER-/PR-/HER2-

4.4 Exploratory objectives and endpoints

4.3.1 The first exploratory objective is to examine how different choices of AI prediction score cut-off point affects the operating characteristics. Exploratory end-points: the image-level AI prediction scores for each of the four views in the screening examination; the exam-level AI prediction score defined by the maximum image-level score.

4.3.2 The second exploratory objective is to assess how the cancer and patient characteristics listed above affect the diagnostic performance of the various combinations of radiologists and AI (Figure 2), (i.e. to test whether there are interactions between patient of cancer characteristics and diagnostic performance).

4.3.3 The third explorative objective is to determine how cancer detection and false positive flagging would be affected by combining AI with radiologists with varying sensitivity and specificity (the sensitivity and specificity of each individual radiologist in flagging in the initial read, and of each combination of radiologists in the consensus discussion).

4.3.4 The fourth exploratory objective is to explore how the AI flagging affects the consensus discussion in terms of the number of cases, the time required by radiologists and to what extent the radiologists in the consensus discussion take the AI information into account, how that changes over time and potential determinants (in addition to the end-points above we will issue repeated questionnaires based on technology acceptance models to the staff involved in the study and measure the time duration of a sample of consensus discussions).

4.3.5 The fifth explorative objective is to assess what the operating characteristics of AI and radiologist was for the following excluded groups had not been excluded: prior breast cancer, breast implants, partial mastectomy, reduction surgery, lipo-filling, biopsy markers and surgical clips, pacemakers or other non-breast implants.

4.3.6 In addition, we will perform exploratory reader studies to understand potential image-related causes of AI false negative and false positive initial read decisions.

5 Statistical methods

All analysis will take place after the relevant end-point(s) has been measured for all study persons. To be explicit, the **fourth explorative objective** above does not require any end-points in terms of cancer detection, but concerns the radiologists' behavior in the consensus discussion and specifically their adherence to the AI decisions. Since it will provide no information on whether the AI system performs well or not, the analysis can be performed before the inclusion has ended.

Analysis based on screen-detected cancer as ground truth will be performed at the earliest 1 month after the mammogram of the last included study person has been acquired and the first version of the database has been released (Figure 3). Analyses including clinical cases will be performed according to the definition of the primary end-point following the release of the database in its second and third version respectively. All statistics, including tables, figures and listings, will be performed using Stata version ≥16 or R version ≥ 4.0.

The data structure in the study will have the following outline (where the endpoint=negative 2x2 table will be constructed for each stage of radiological assessments: 1. Reader flagging; 2. Consensus discussion; 3. Tissue sampling (biopsy)):

	Endpoint = positive (yes)		Endpoint = negative (no)	
	SOC positive	SOC negative	SOC positive	SOC negative
Positive using novel reader combination including AI positive	a	b	e	f
Negative using novel reader combination including AI negative	c	[d]	g	[h]

Here, SOC denotes standard of care (defined as two human radiologists). The SOC can then be compared with respect to number of endpoint positive (cancers detected) and endpoint negative (no cancer detected) with novel reader combinations including AI (Figure 2).

It should be noted that since only patients who screen positive on either SOC or SOC together with AI are referred for further work-up, the number of patients reported between brackets are unknown. This is a feature of the screen-positive design and means that absolute sensitivity and specificity cannot be estimated.

Analyses will compare the true positive fraction between different combinations of readers (radiologists or AI). Comparisons will be made on a relative scale. The relative true positive fraction (rTPF) is defined as TPF_r/TPF_{SOC}, where TPF_r is a specific reader combination *r* (Figure 2) being compared with the TPF of using two radiologists (SOC). The analysis will largely follow the methods described by Pepe and Alonzo (10).

The rTPF is estimated as (a+b)/(a+c) (or (a+c)/(a+b), as appropriate) and an approximate 100(1-α)% two-sided confidence interval for rTPF is calculated as

$exp\left(log(\widehat{rTPF}) \pm\ z_{\alpha/2}\sqrt{\frac{b+c}{(a+b)(a+c)}}\right).$

Analogous formulas are used for comparisons within the experimental arm and for the relative False Positive Fraction.

Non-inferiority and superiority tests

For comparisons of the rTPF, the null and the alternative hypothesis for non-inferiority and superiority tests are

$$H_0\colon rTPR\ \leq\ exp(-\delta)$$

$$H_a\colon rTPR\ >\ exp(-\delta)$$

and

$$H_0\colon rTPR\ \leq\ exp(\theta)$$

$$H_a\colon rTPR\ >\ exp(\theta)$$

respectively, with non-inferiority and superiority margins equal to $\delta > 0$ and $\theta\ \geq\ 0$.

One-sided p-values will be calculated based on the test considered (non-inferiority or superiority).

Switching from non-inferiority to superiority: if the two-sided $(2(1-\alpha) \times 100)\%$ confidence interval for rTPF not only lies entirely above the non-inferiority margin, but also above the superiority margin, superiority will be claimed at the same alpha-level set for the non-inferiority test. In this case, we will also calculate the p-value associated with a test for superiority.

We will assess the non-inferiority of novel reader combinations involving AI versus SOC for detecting cancers ($exp(-\delta) = 0.85$). The one-sided α level is set to 0.025. Two-sided 95% confidence intervals will be reported.

For the secondary endpoints, the relative false positive fraction (rFPF, defined as FPF_r/FPF_{SOC} for reader combination r) will be assessed at each stage of radiological assessments: 1. Reader flagging; 2. Consensus discussion; 3. Tissue sampling (biopsy). The rFPF is estimated as (e+f)/(f+g) for each stage of radiological assessments and confidence intervals are computed analogously to the rTPR.

Heterogeneity of treatment effect

Subgroup analyses will be performed for patient and cancer characteristics listed in 4.3 Additional data collected in the study. Statistical tests for effect heterogeneity across subpopulations will be performed by jointly testing the interaction (product) terms in generalised linear models or marginal models(10), as appropriate. No correction for multiple comparisons will be made.

5.1 Populations

The study population consists of all consecutive women attending regular screening mammography at Capio S:t Göran hospital in Stockholm (excluding, by necessity, any time period when the AI system is not functional). The data for each woman consists of: mammographic images, date of examination, age at examination, breast density measure, assessment by first radiologist, assessment by second radiologist, assessment by AI, assessment by consensus discussion (if any), pathology-verified cancer diagnosis (if any) including key cancer characteristics (listed above under 4.1 and 4.2).

5.2 Demographics and baseline data

All data will be presented using descriptive statistics. Continuous variables will be summarized using number of women, mean, standard deviation, median, interquartile range (IQR), minimum and maximum. Categorical variables will be categorized as described in the end-points and summarized using the number and percentage of cases in each category.

5.3 Setting the operating point of the AI system

A historical calibration dataset from the same institution will be used to set the AI operating point so that AI plus the recorded first reader would have a joint sensitivity 2 % higher than the joint sensitivity of the recorded first and second reader. The AI algorithm outputs a continuous score, and the threshold for the binary flagging decision was set to 0.534 (described in the WHO format research protocol). The threshold has so far not changed over the course of the study.

5.4 Primary analyses: Non-inferiority

Non-inferiority analyses in terms of sensitivity (rTPF as described on the prior page) of cancer as defined in the primary end-point (4.1) for initial read flagging performed by (i) AI and the first-reader radiologist compared to (ii) the first- and second-reader radiologists. The study has been powered for the primary analysis of cancer at the first screening examination (Figure 3), and adding cases at the subsequent follow-up time-points will further increase the statistical power.

5.5 Secondary analyses

In several of the secondary analyses we use the term **operating characteristics** referring to the various performance indicators used in evaluation of screening programs. These operating characteristics are based on the number of true and false, positive and negative, assessments by the various actors (radiologists and AI) at the various time-points (initial read, consensus discussion, tissue sampling) and based on the cancer ground truth defined by the primary end-points. Operating characteristics include sensitivity, specificity, recall rate, false negative rate, false positive rate, and the positive and negative predictive value at initial read (by each reader and by readers combined) and at consensus discussion (see Table 1). It should be noted that operating characteristics that relate to negative samples (such as the specificity, negative predictive values, and false negative rate) are not directly estimable under a screen-positive design (see Section 5.0). True and false negatives can however be estimated approximately with follow-up time, under the assumption that no new cancers develop and present clinically after the screen and that all cases are detected within the follow-up period. A longer follow-up period will increase the likelihood that cancers are discovered, but will also increase the chance of new cancers developing and being discovered. We will use a follow-up time of 23 months (the time of one screening cycle in a biennial screening program) for estimating statistics based on estimates of true and false negatives (e.g. specificity, negative predictive values, and false negative rate). For each analysis of operating characteristics, we will perform a sensitivity analysis where women who were recalled due to symptoms (i.e., with radiologists finding no suspicious finding in the mammogram) are either excluded or re-classified as detected by all readers, AI and radiologists alike.

SENSITIVITY ANALYSES RELATED TO PRIMARY ANALYSIS

5.5.1 Repeat the primary analysis after re-classifying women who were recalled in screening due to symptoms (not due to mammographic signs) as detected by both AI and radiologists.

5.5.2 Repeat the primary analysis as a per-protocol-analysis where the examinations for which AI was not used are excluded.

5.5.3 Repeat the primary analysis without cancers with less aggressive characteristics: grade 1 in situ cancer grade 1, all in situ cancer.

5.5.4 Describe and compare the distribution of patient and cancer characteristics (described in 4.2.4) and stage (according to AJCC 7 and 8) between the two reader combinations in the primary analysis.

OTHER SECONDARY ANALYSES

5.5.5 Determine initial read operating characteristics (Table 1) for each combination of (i) initial read and (ii) arbitration as described in Figure 2.

5.5.6 Determine initial read operating characteristics for AI-only read below, and AI plus first-reader radiologist above, various AI score cut-off points.

5.5.7 Determine arbitration operating characteristics for an initial read by AI and one radiologist followed by either (i) the actual consensus discussion or (ii) arbitration decision based on recall of concordant cases and arbitration by second radiologist-decision for discordant cases.

5.5.8 Compare whether the cancer characteristics are different for the initial read cases flagged and not flagged between the AI reader and the radiologist readers.

5.6 Exploratory analyses

5.6.1 Determine the initial read operating characteristics of a binary detection at each possible cut-off point for the AI scores based on the ground truth of cancer defined by end-point 4.2.1 and 4.2.2.

5.6.2 Determine the sensitivity of radiologist and AI depending no mammographic density (divided into four categories) and patient age (divided by 10-year intervals).

5.6.3 Determine whether the joint decision by AI and radiologist is more often false positive or false negative depending on the overall sensitivity and specificity of the radiologist with whom the AI decision is combined.

5.6.4 a) Sample the time and number of cases for a sample of consensus discussions during one or more weeks before use of the AI system, and at various time-points after taking the AI system into use. Compare if there is a difference in the total time of consensus discussion for a standardized number of screening examinations, and if there is a difference in the time per case in the consensus discussions. b) Distribute a query based mainly on the Technology Acceptance Model 3 to all staff, and especially important to the radiologists participating in the consensus discussion. Have radiologists complete a mini-survey at the end of each consensus discussion.

5.6.5. a) Determine the operating characteristics of AI and radiologist for the following excluded groups: prior breast cancer and breast implants. b) To the extent possible, determine the sensitivity of flagging by radiologist and by AI for diagnosed women stratified by each characteristic listed in 4.3.1.2.

5.6.6 Determine differences in visual radiological characteristics between cases (i) not flagged by AI and (ii) not flagged by a radiologist to determine systematic differences between the two groups.

5.6.7 Examine a random selection of healthy women flagged by AI but not by any of the two initial read radiologists, and not recalled, to understand to what extent prior images or clinical information could have contributed to or averted a false positive read.

5.6.8 Examine all cases where at least one reader (AI or any radiologist) made a true positive initial read and at least one reader (AI or any radiologist) made a false negative initial read. Perform a reader study to identify potential image-related findings explaining the false negative assessments.

5.6.9 Examine the tendency of the consensus discussion to make more false-negative or more false-positive assessments (in terms of end-point 4.2.2 including interval cancers) depending on whether the initial read assessments were concordant, discordant with AI flagging and no radiologist, discordant with one radiologist flagging, other discordant flagging.

5.6.10 Determine the sensitivity difference between the combinations of AI and radiologists stratified by women 40-41 years old (initial screen) and women in subsequent screens.

5.7 Sample size

The sample size calculation relates to the primary analysis after women have undergone the baseline mammogram. Power for the additional two follow-up time points (at 12 and 23 months) will be higher, since there will be a larger number of cancers detected in the study at these time points than at the first time points (3 months).

The sample size calculation follows the methods for paired screen positive designs described in Alonzo, Pepe, and Moskowitz (9). The sample size calculation is based on the following assumptions: A prevalence of breast cancer in the screening of 0.5%, 0.70 TPF for a mammogram, and rTPF (AI+radiologist vs. SOC) equal to **1.02**. The non-inferiority margin was set to 0.85 and a one-sided alpha to 0.025. For the sample size calculations, we assumed that every woman who is recommended to undergo a biopsy also does so. Based on these assumption, 55,000 women will be included in the study giving a power of 87%.

5.8 Adjustment for multiplicity

We will not perform any correction for multiple comparisons. Each analysis will be presented with unadjusted 95% confidence interval.

5.9 Handling of missing or contradictory data

Endpoints

Our primary approach is of intention-to-treat-type: for any examination where the AI process failed, the AI assessment will be replaced by the second reader radiologist. In addition, we will perform a per-protocol analysis, excluding all examinations where the AI process failed.

For missing data cases where Reader 2 failed to register her/his own decision, this will be defined as the same as the AI decision if there is less than 1% of examinations with this type of missingness. If there is a large number of this type of missing data, imputation will be used.

For contradictory data cases where Reader 2 registered a negative joint decision (AI and Reader 2), despite AI being positive, the joint decision will be changed to positive. If also Reader 1 registered a negative decision this means that the exam did not go to consensus discussion and the woman would not be recalled. This follows the intention-to-treat methodology that we apply.

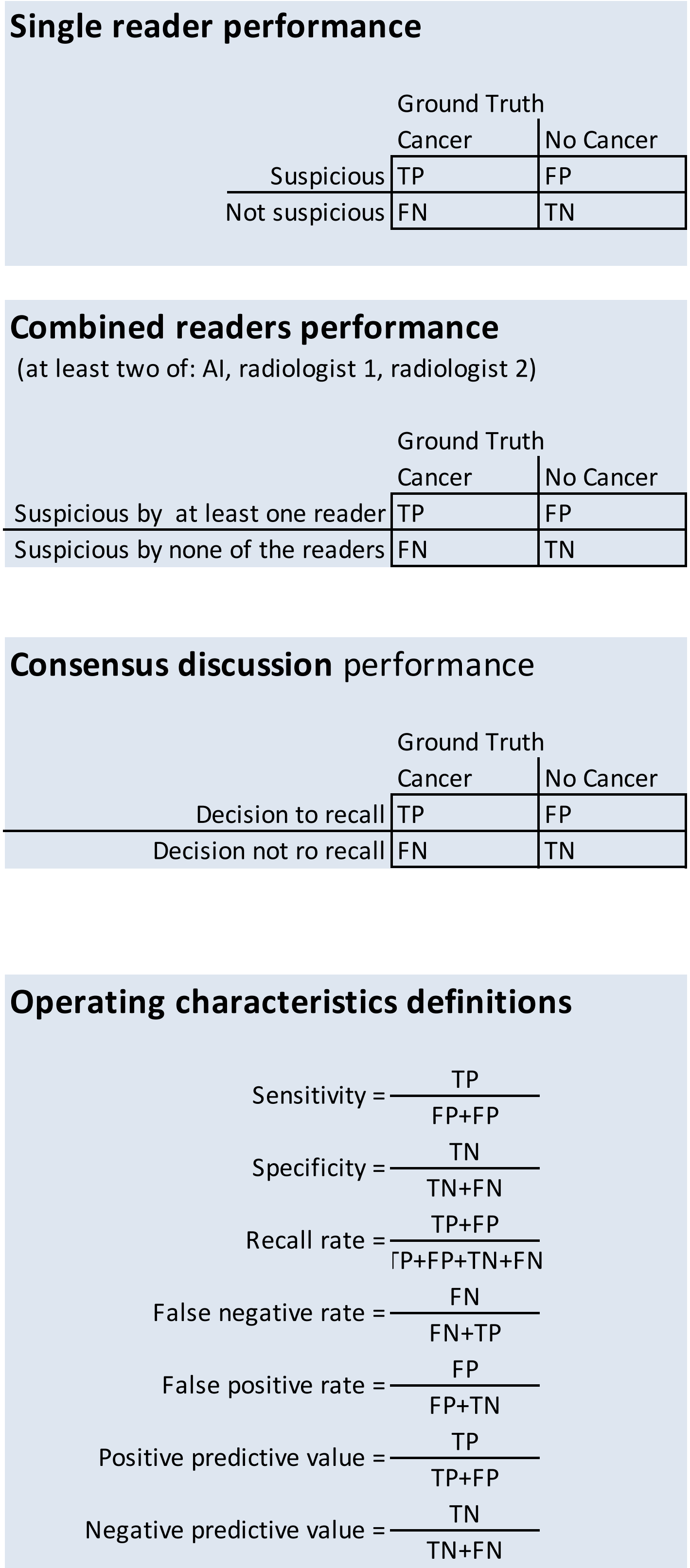
To reduce missing data at the 12-month and 23-month follow-up time points, the follow-up will be based on the national quality register of breast cancer in order to obtain information for women who have died or moved to another region of Sweden.

Patient characteristics

Since data collection is performed through the electronic medical record, missing data is expected to be minimal. For analyses where patient characteristics are used (e.g. analyzing the interaction between a patient characteristics and study outcomes), we will exclude patients with missing data. As a sensitivity analysis, we will impute the missing data using the mean or median, whichever is appropriate.

6 Amendment with description of post hoc analyses

Tables and Figures



Single reader performance

	Ground Truth	
	Cancer	No Cancer
Suspicious	TP	FP
Not suspicious	FN	TN

Combined readers performance

(at least two of: AI, radiologist 1, radiologist 2)

	Ground Truth	
	Cancer	No Cancer
Suspicious by at least one reader	TP	FP
Suspicious by none of the readers	FN	TN

Consensus discussion performance

	Ground Truth	
	Cancer	No Cancer
Decision to recall	TP	FP
Decision not ro recall	FN	TN

Operating characteristics definitions

$$\text{Sensitivity} = \frac{TP}{FP+FP}$$

$$\text{Specificity} = \frac{TN}{TN+FN}$$

$$\text{Recall rate} = \frac{TP+FP}{TP+FP+TN+FN}$$

$$\text{False negative rate} = \frac{FN}{FN+TP}$$

$$\text{False positive rate} = \frac{FP}{FP+TN}$$

$$\text{Positive predictive value} = \frac{TP}{TP+FP}$$

$$\text{Negative predictive value} = \frac{TN}{TN+FN}$$

Table 1. Definition of operating characteristics. Ground truth is defined by pathology-verified breast cancer for the three follow-up time points described in primary end-point (4.1). It should be noted that operating characteristics that relate to negative samples (such as the specificity, negative predictive values, and false negative rate) are not directly estimable under a screen-positive design (see Section 5.0). True and false negatives can however be estimated approximately with follow-up time, under the assumption that no new cancers develop and present clinically after the screen and that all cases are detected within the follow-up period. A longer follow-up period will increase the likelihood that cancers are discovered, but will also increase the chance of new cancers developing and being discovered. We will use a follow-up time of 23 months (the time of one screening cycle in a biennial screening program) for estimating statistics based on estimates of true and false negatives (e.g. specificity, negative predictive values, and false negative rate).

Figure 1. The actual work flow in the study. If any one of the initial readers (First Radiologist, Second Radiologist or AI) decides to flag for a suspicious finding in the image, the exam is routed to the consensus discussion. The consensus discussion consists of an oral discussion between at least two radiologists (could be same or different as the initial readers) while reviewing the images of each case, and deciding to recall the woman for further diagnostics or to determine that she is healthy.

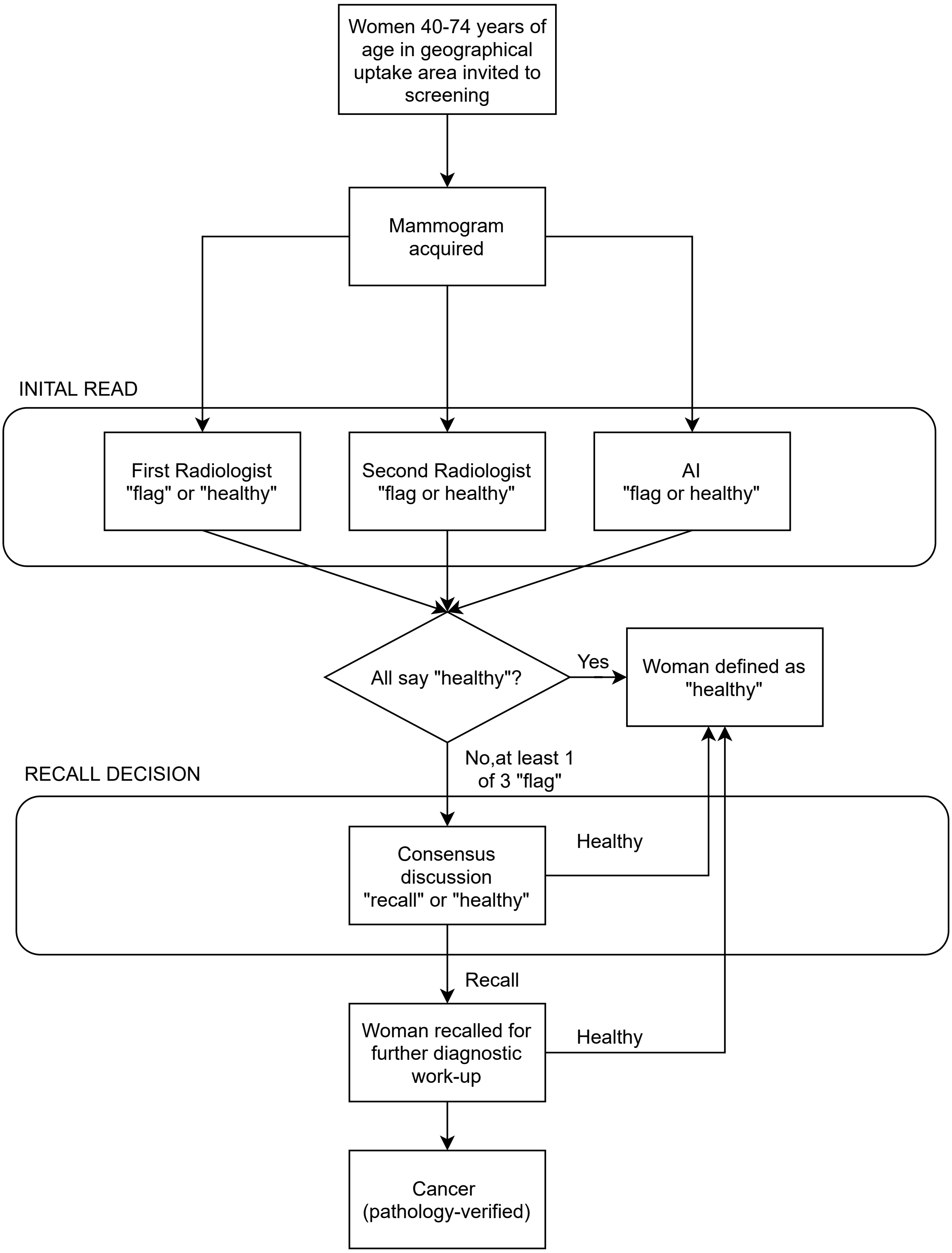

Women 40-74 years of age in geographical uptake area invited to screening
Mammogram acquired
INITAL READ
First Radiologist "flag" or "healthy"
Second Radiologist "flag or healthy"
AI "flag or healthy"
All say "healthy"?
Yes
Woman defined as "healthy"
No,at least 1 of 3 "flag"
RECALL DECISION
Consensus discussion "recall" or "healthy"
Healthy
Recall
Woman recalled for further diagnostic work-up
Healthy
Cancer (pathology-verified)

Figure 2. Alternative scenarios for which outcomes will be analyzed. The alternative at the top of each column is the actual implementation in the study. However, based on the recorded assessments of each radiologist and AI, other alternative scenarios will be analyzed in post-study analysis. To minimize the risk of incomplete cases, for any case of initial where AI is the only reader or combined with one radiologist, and the AI process fails, the AI assessment will be replaced by a radiologist read.

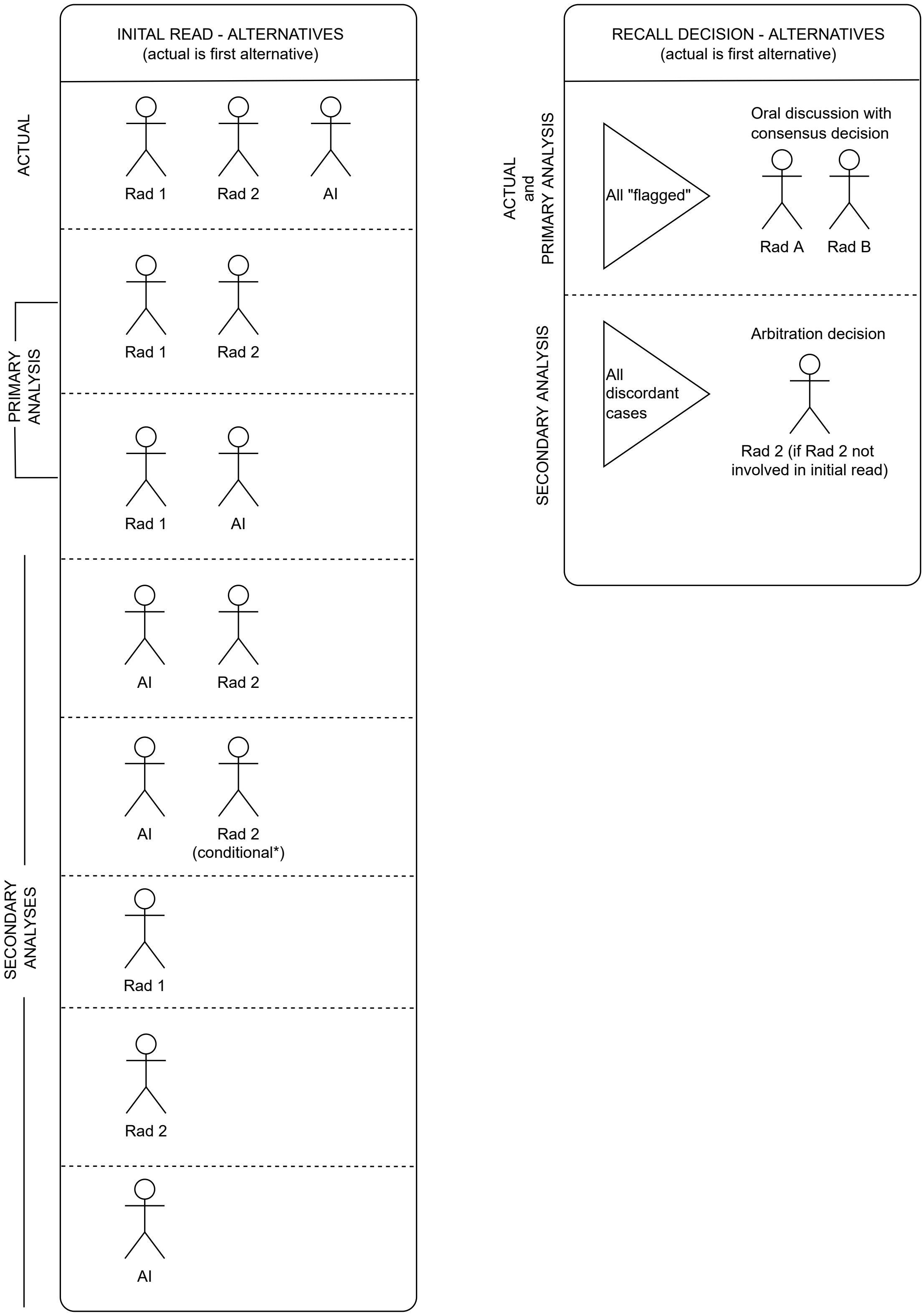
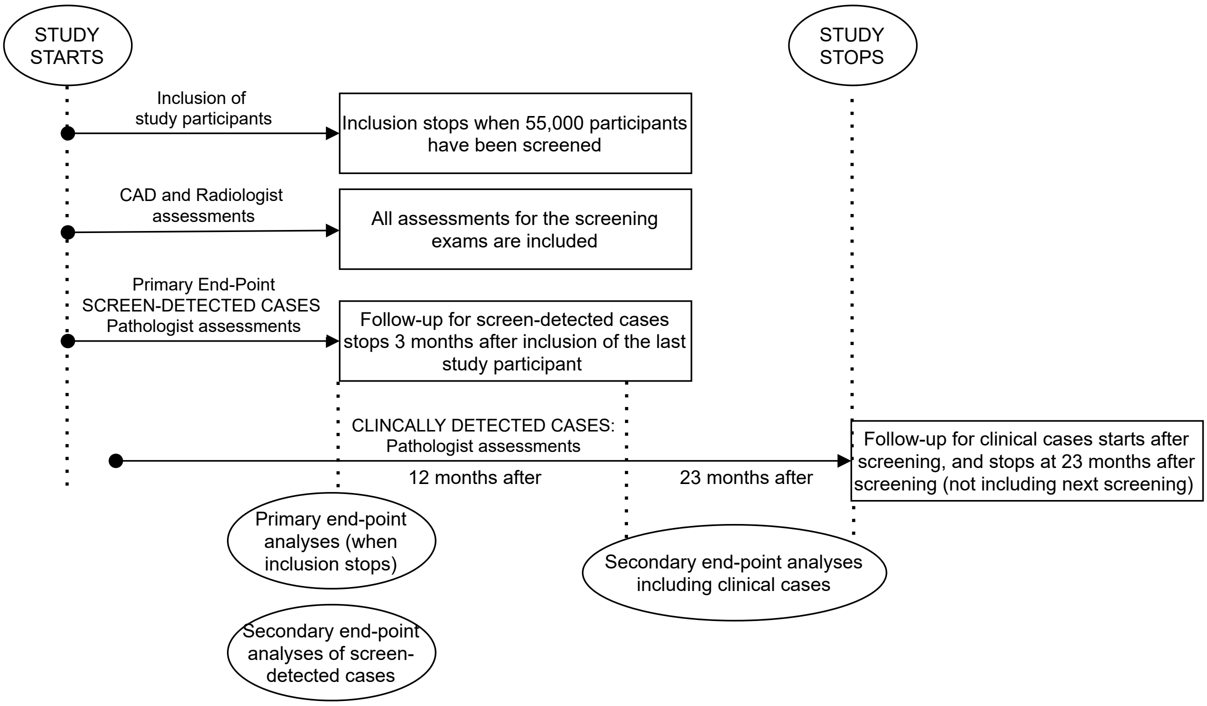

INITAL READ - ALTERNATIVES
(actual is first alternative)
ACTUAL
Rad 1
Rad 2
AI
PRIMARY ANALYSIS
Rad 1
Rad 2
Rad 1
AI
SECONDARY ANALYSES
AI
Rad 2
AI
Rad 2 (conditional*)
Rad 1
Rad 2
AI
RECALL DECISION - ALTERNATIVES
(actual is first alternative)
ACTUAL and PRIMARY ANALYSIS
All "flagged"
Oral discussion with consensus decision
Rad A
Rad B
SECONDARY ANALYSIS
All discordant cases
Arbitration decision
Rad 2 (if Rad 2 not involved in initial read)

Figure 3. Overview of study time line. The primary end-point is screen-detected cancer, which will be defined by CAD and/or radiologist assessment in combination with pathology-verification of cancer. One secondary end-point is all cancer, which is defined by screen-detected cancer plus all other, clinically detected, cancer for the study participants during a follow-up period of 23 months from the included screening exam.


STUDY STARTS
STUDY STOPS
Inclusion of study participants
Inclusion stops when 55,000 participants have been screened
CAD and Radiologist assessments
All assessments for the screening exams are included
Primary End-Point SCREEN-DETECTED CASES Pathologist assessments
Follow-up for screen-detected cases stops 3 months after inclusion of the last study participant
CLINCALLY DETECTED CASES: Pathologist assessments
12 months after
23 months after
Follow-up for clinical cases starts after screening, and stops at 23 months after screening (not including next screening)
Primary end-point analyses (when inclusion stops)
Secondary end-point analyses including clinical cases
Secondary end-point analyses of screen-detected cases

References

1. Weedon-Fekjær H, Romundstad PR, Vatten LJ. Modern mammography screening and breast cancer mortality: population study2014 2014-06-17 22:30:49.

2. Törnberg S, Kemetli L, Ascunce N, Hofvind S, Anttila A, Sèradour B, et al. A pooled analysis of interval cancer rates in six European countries. European journal of cancer prevention. 2010;19(2):87-93.

3. O'Donoghue C, Eklund M, Ozanne EM, Esserman LJ. Aggregate cost of mammography screening in the United States: comparison of current practice and advocated guidelines. Annals of internal medicine. 2014;160(3):145-53.

4. Taylor-Phillips S, Jenkinson D, Stinton C, Wallis MG, Dunn J, Clarke A. Double Reading in Breast Cancer Screening: Cohort Evaluation in the CO-OPS Trial. Radiology. 2018:171010.

5. Rodriguez-Ruiz A, Lang K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. Journal of the National Cancer Institute. 2019.

6. Kyono T, Gilbert FJ, van der Schaar M. Improving Workflow Efficiency for Mammography Using Machine Learning. Journal of the American College of Radiology : JACR. 2019.

7. Ribli D, Horvath A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep. 2018;8(1):4165.

8. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA oncology. 2020.

9. Alonzo TA, Pepe MS, Moskowitz CS. Sample size calculations for comparative studies of medical tests for detecting presence of disease. Statistics in medicine. 2002;21(6):835-52.

10. Pepe MS, Alonzo TA. Comparing disease screening tests when true disease status is ascertained only for screen positives. Biostatistics. 2001;2(3):249-60.