

SUPPLEMENTARY MATERIALS

Supplementary Table1. Cases read mammography without/with AI by each radiologist

Supplementary Table 2. Contingency table for comparison of sensitivity from results of a previous study

Supplementary Table 3. Sample size calculation

Supplementary Table 4. Calculation as the sample size decreased how power changes

Supplementary Table 5. Diagnostic Accuracy Assessment According to the Gold Standard

Supplementary Table 6. CONSORT-AI checklist

Supplementary Figure 1. Image acquisition and data management

Supplementary Note 1. Description of the data source and data collection process

Supplementary Note 2. Description of the data monitoring, cleaning, and retention

Supplementary Note 3. Details on the AI-based mammography imaging analysis software

Supplementary Note 4. Sample size and power considerations

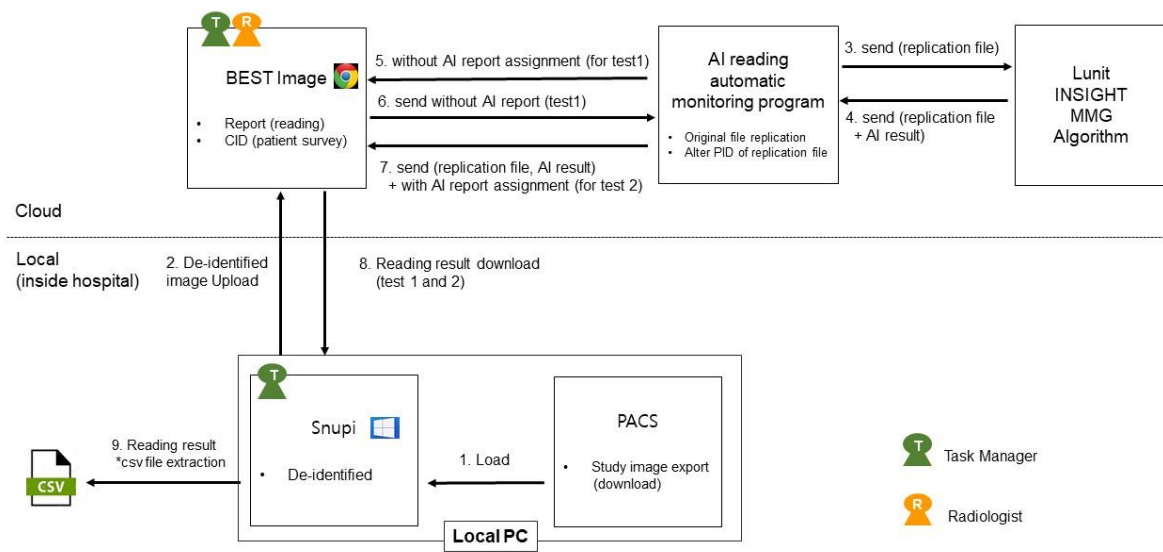
Supplementary Note 5. Detailed statistical analyses

Supplementary Note 6. CONSORT-AI checklist

BR (A-F)	DM screens without/ with AI, N (%) (N=24,545)	GR (G-K)	DM screens without/ with AI, N (%) (N=24,545)
A (2021-2022)	7178 (29.2)	G	7178 (29.2)
B (2021-2022)	5140 (20.9)	H	5804 (23.7)
C (2021-2022)	5414 (22.1)	I	5414 (22.1)
D (2021-2022)	3169 (12.9)	J	3169 (12.9)
E (2021), F (2022)	1707 (7.0), 1937 (7.9)	K	2980 (12.1)

Supplementary Table 1. Cases read mammography without/with AI by each radiologist.

BR; breast radiologist (A-F), GR: general radiologist (G-K)



Supplementary Figure 1. Image acquisition and data management

Supplementary Note 1. Description of the data source and data collection process

Participant-reported demographics and medical history were collected from the Participant Information Sheet and study questionnaire; radiologist-collected mammographic findings were collected from the radiologist data collection forms; standalone AI findings were collected from the image data file; other mammographic and pathologic features were collected from medical records and pathology report (if available); breast cancer onset and survival status were collected from registry databases. Mammograms from participants were exported to the study platform (BEST image) for entry of reading results after Digital Imaging and Communication in Medicine (DICOM) standard de-identification according to the Health Insurance Portability and Accountability Act (HIPPA) ¹. Idea to Reality in Medicine (IRM)'s commercial BEST image is a certified local program included in the platform support for a multicenter study initiative by the Korean Society of Radiology. IRM's commercial SNUPI program was used for de-identification, including participant identification (ID), participant name, study instance unique identifiers (UIDs), participant sex, participant birth date, participant age, study date, series date, private tags, study description, and series description. The study participant ID was assigned instead of the institution participant ID and the relevant mapping information was saved in an independent computer by the Principal Investigator (PI) at the site. If the participant had mammography records at the relevant hospital from within the past four years, all such data were exported to the study platform so that a reader can maintain the same environment as the actual reading at the hospital. Set 1 was part of the standard screening procedure and Set 2 was performed independently after Set 1 and did not affect the actual reading results of the participants.

Supplementary Note 2. Description of the data monitoring, cleaning, and retention

A data cleaning method was employed to correct inconsistencies or errors that were not captured during data entry (e.g., outliers or conflicting data). Data queries were identified on an ongoing basis during data collection. To enable evaluations and/or audits from regulatory authorities or the sponsor, the investigator agreed to retain all records, including the identity of all participants (sufficient information to link records, eg, electronic case report forms and medical charts), source documents, detailed records of participant disposition, and adequate documentation of relevant correspondence (eg, letters, meeting minutes, telephone calls reports). When the study was completed, a clinical study report was submitted. All records and documents were transferred to the document retention staff and retained for five years. If the investigator became unable for any reason (e.g., retirement or relocation) to continue to retain study records for the required period, the study records must be transferred to a designee, such as another investigator or another institution. After the retention period, records and documents will be destroyed to prevent disclosure of the contents of the documents and preserve confidentiality, and documents related to personal information will be destroyed according to the Personal Data Protection Act Enforcement Decree Article 16.

Supplementary Note 3. Details on the AI-based Mammography Imaging Analysis Software

Subsequent advancements in AI technology have improved the accuracy of CAD. In particular, CAD using the convolution neural network, a deep learning technology optimized for image analysis and classification, is providing diagnostic performance that is equivalent or superior to those of radiologists ¹. Lunit INSIGHT MMG, software aids mammogram reading by detecting breast cancer. It was developed by training an AI algorithm using breast cancer mammography data (>50000 cases), and the total training data included more than 200000 cases collected in Korea, the United States, and the United Kingdom ². A clinical study on Lunit INSIGHT MMG was conducted to obtain authorization from the Korean Ministry of Food and Drug Safety. It assessed the improvement in reading performance of 14 radiologists using 160 malignant cases and 160 non-malignant cases (64 positive cases, 96 normal cases) collected at two sites. Reading with AI-CAD significantly improved the accuracy of breast cancer detection and decreased the false positive rate in both the breast radiologist group and the general radiologist group, suggesting the potential of AI-CAD to improve breast cancer detection, and reduce the need for additional tests ². In a recent study using Swedish breast cancer screening cohort data, the AI system showed superior performance compared with two other commercial AI-based software products³. The area under the receiver operating characteristic curve (AUROC) for INSIGHT MMG (0.956) was higher than those for the other products (0.922 and 0.920), and its sensitivity (81.9%) was similar to that of physicians (77.4% and 80.1%); the competitor products exhibited significantly lower sensitivity than physicians³. The superior performance of the AI system compared with other products is a result of using at least four times more breast cancer cases as training data than the other commercial products. Furthermore, over half of the training data were derived from dense breasts, which are difficult

to read; and the lesion locations were marked by breast radiologists for most breast cancer cases in the training data ².

Supplementary Note 4. Sample size and power considerations

1. The sample size was derived based on the following assumptions using R (version 4.3.3.; R Foundation for Statistical Computing, Vienna, Austria)

- Primary endpoint: cancer detection rate (CDR)
- Hypotheses of primary interest:
 - ✓ Null hypothesis: There is no difference in CDR between with and without AI.
 - ✓ Alternative hypothesis: There is a difference in CDR between with and without AI.
- Prevalence:
 - ✓ 3.21 per 1,000 exams; $3,781/1,176,717 \approx 0.00321$ ⁴
- Type I error (α): 0.05 (two-sided)
- Power ($1-\beta$): 80%

Sensitivity results from a previous study are presented in

and were used in the sample size derivation ⁵.

Supplementary Table 2. Contingency Table for Comparison of Sensitivity from Results of a Previous Study ^{2,5}

2 x 2 Table	Without AI (detection O)	Without AI (detection X)
With AI (Detection O)	1,657	242
With AI (Detection X)	29	312

Note: AI, artificial intelligence

The sample size calculation formula is based on the McNemar's test as per the following:

$$n = \left(\frac{z_{1-\alpha/2} \sqrt{p_{disc}} + z_{1-\beta} \sqrt{p_{disc} + p_{diff}^2}}{p_{diff}} \right)^2$$

p_{disc} : proportion of discordant pairs, i.e., 0.12098

p_{diff} : difference in the proportions of detected cases, i.e., 0.095

$z_{1-\alpha/2}$: $1 - \alpha/2$ th percentile of standard normal distribution, i.e., 1.96

$z_{1-\beta}$: $1 - \beta$ th percentile of standard normal distribution, i.e., 0.84

Here, p_{disc} and p_{diff} refer to the proportion of discordant pairs and the difference in the proportion, respectively, between cancer detection in total participants with AI and without AI. As they are not affected by negative data, the proportions derived from data from the table below were multiplied by prevalence and used in the sample size calculation.

Supplementary Table 3. Sample Size Calculation

Prevalence	CDR without AI	CDR with AI	p_{disc}	N
3.21 per 1,000 exams	0.0024	0.0027	0.00038	32,714

Note: AI = artificial intelligence; CDR = cancer detection rate

The target sample size is 32714 participants, corresponding to approximately 16000 participants per year.

2. Results of recalculating the number of subjects

The total number of expected participants changed from the initial study design due to the COVID-19 pandemic, but no effect was observed on the primary study endpoint. Given the same cancer prevalence as 3.21 per 1000 examinations, it was calculated that the sample size could be maintained if approximately 24,000 people were recruited, while still maintaining 80% power and detecting more than 90 cases of cancers. The detailed calculation method was as follows.

2.1. Expected power

In the formula for calculating the number of subjects based on McNemar's test, p_{disc} and p_{diff} are the proportions corresponding to the discrepancy and difference in cancer detection between without AI and with AI among the total subjects, respectively. Since these two are

not affected by negative data, the number of subjects was referenced by multiplying the prevalence in the table based on the Lancet digital health 2020 data².

Assuming the same prevalence, CDR, and discordant proportion as the study plan, we calculated how power changes as the sample size decreases.

Supplementary Table 4. Calculation as the sample size decreased how power changes

Prevalence	CDR without AI	CDR with AI	p_{disc}	power	N
3.21 per 1000 exams	0.0024	0.0027	0.00038	80%	32714
3.21 per 1000 exams	0.0024	0.0027	0.00038	75%	28928
3.21 per 1000 exams	0.0024	0.0027	0.00038	70%	25726
3.21 per 1000 exams	0.0024	0.0027	0.00038	66.9%	24000
3.21 per 1000 exams	0.0024	0.0027	0.00038	65%	22926
3.21 per 1000 exams	0.0024	0.0027	0.00038	60%	20419

2.2. Number of cancer patients needed

Power is set to 80%, the discrepancy rate is set to be the same as at the time of the research plan ($p_{disc} = 0.000388$), and the required CDR difference ($=p_{diff}$) according to the sample size is calculated as follows.

Power	N	P10 : without AI, the rate of detecting, with AI, the rate of no detecting	P01 : without AI no detect, with AI detect rate	P_{diff}	P_{disc} = P10+P01
80%	24,000	0.000372	0.000016	0.000356	0.000388

⇒ The study plan assumed the CDR difference as $0.0027 - 0.0024 = 0.0003$, but in order to

achieve 80% power even when the sample size is reduced to 24,000, the difference value

should be 0.000356, which is 0.00005 larger than the original 0.0003. If the cancer

prevalence is 3.21 per 1,000 exams, and the power is maintained at 80% in the same

situation, the difference in sensitivity should be $0.000356 / 0.00321 = 0.1109$. (In the study

plan, $0.000305/0.00321 = 0.095$)

To maintain the difference in sensitivity as 0.095, which is the same as the study plan, and show 80% power with 24,000 people, the CDR difference should be 0.000356, and to do so, the cancer prevalence should be 0.003744. In other words, if 24,000 people are recruited, $24,000 * 0.003744 =$ approximately 90 cancer patients are needed.

Supplementary Note 5. Detailed statistical analyses

All statistical analyses and reporting took place after the relevant endpoints have been measured from all the study participants. All statistical analyses were performed using R version 4.3.3. (R Foundation for statistical Computing, Vienna, Austria). Data obtained from the mammograms was classified into: (FA set) all mammograms assessed by the AI- based CADe/x software for this study after receiving an identification (ID) code; (per protocol [PP] set)_all mammograms included in the FA set except 1) any mammograms where all study procedures are not completed; 2) any data from the mammograms obtained or handled in violation of the study protocol. Where the analysis results from each set are different, the results of both analyses were presented and the reasons for different results will be described in detail. The interim analyses included diagnostic accuracy in the primary and secondary study outcomes, calculated using pathologic results as the gold standard reference group obtained within one year from screening. Final analyses were performed once the data from all participants had been collected and cleaned, and database lock had occurred. The analysis examined primary study outcomes based on diagnosis accuracies calculated using pathology data as the reference group obtained within one year from screening.

The data structure had the following outline based on the malignancy scale (7-point scale) and the gold standard ³:

Supplementary Table 5. Diagnostic Accuracy Assessment According to the Gold Standard

Screening	Gold Standard		
	(+)	(-)	
(+)	a	b	a + b
(-)	c	d	c + d
	a + c	b + d	N

The following diagnostic accuracy endpoints were derived for each test (eg, with the primary endpoint, breast radiologists in Set 1 without AI-CAD [Test 1] and breast radiologists in Set 1 with AI-CAD [Test 2]):

- Cancer detection rate = $a / N \times 1000$ (i.e., presented as per 1000 participants)
- Recall rate = $(a + b) / N \times 100$ (i.e., presented as a percentage)

For descriptive analyses, continuous data were described by the sample size, mean, median, standard deviation, quartile 1, quartile 3, interquartile range, minimum, and maximum. Unless otherwise specified, the minimum and maximum values were presented with the same number of decimals as the raw data and the other statistics with one more decimal than the raw data except for the standard deviation, where two more decimals than the raw data was presented. Categorical variables were described by frequency and percentages (n, %). Percentages were calculated using the specified denominator in the table and presented with one decimal point. Missing data were included as a separate category in some cases, depending on the nature of the variable.

In the interim analysis, the diagnostic accuracy endpoints (i.e., cancer detection rate, recall rate) were derived, with the corresponding two-sided 95% CI as estimated using the Clopper-Pearson exact method. Unless otherwise specified, the endpoint estimates and two-sided 95% CIs were be presented with two decimal points. When inferential tests were performed for differences between diagnosis tests, a two-sided $p < 0.05$ indicated statistical significance.

Correction for multiple comparisons was not performed in the interim analysis. Therefore, each comparative analysis was presented with an unadjusted two-sided 95% CI.

If the sample size allows, the primary and secondary endpoint comparisons of the diagnostic accuracy derived variables were presented stratified by subgroups of patients. The subgroups of interest included:

- Breast density: fatty breast (density A and B), dense breast (density C and D)
- Participant age (age 40-49, 50-59, 60-69, ≥ 70 years)
- Pathologic features of breast cancer that are detected or missed by AI and reader
 - ✧ Tumor size (< 20 mm, ≥ 20 mm)
 - ✧ Breast cancer type (carcinoma in situ, infiltrating breast cancer)
 - ✧ Histological grade of invasive cancers (grade 1, grade 2, grade 3)
 - ✧ Lymph node statuses of cancers (positive or negative)
 - ✧ Molecular subtype (luminal A, Non-luminal A; luminal B, HER2 enrich, triple negative)

Supplementary Table 6. CONSORT-AI checklist

Section		CONSORT 2010 item ^a	CONSORT-AI item	Pg ^b
Title and abstract				
Title and Abstract	1a	Identification as a randomized trial in the title	CONSORT-AI 1a,b Elaboration	(i) Indicate that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specify the type of model. 1
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstract)		(ii) State the intended use of the AI intervention within the trial in the title and/or abstract. 2
Introduction				
Background and objectives	2a	Scientific background and explanation of rationale	CONSORT-AI 2a (i) Extension	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (for example, healthcare professionals, patients, public). 3
	2b	Specific objectives or hypotheses		4
Methods				
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio		18,19
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons		18,19
Participants	4a	Eligibility criteria for participants	CONSORT-AI 4a (i) Elaboration	State the inclusion and exclusion criteria at the level of participants. 19
			CONSORT-AI 4a (ii) Extension	State the inclusion and exclusion criteria at the level of the input data. 19
	4b	Settings and locations where the data were collected	CONSORT-AI 4b Extension	Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements. 19-21
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	CONSORT-AI 5 (i) Extension	State which version of the AI algorithm was used. 21
			CONSORT-AI 5 (ii) Extension	Describe how the input data were acquired and selected for the AI intervention. 19-21
			CONSORT-AI 5 (iii) Extension	Describe how poor quality or unavailable input data were assessed and handled. 19-21
			CONSORT-AI 5 (iv) Extension	Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users. 19-21
			CONSORT-AI 5 (v) Extension	Specify the output of the AI intervention 19-21
			CONSORT-AI 5 (vi) Extension	Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice. 19-21
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed		21
	6b	Any changes to trial outcomes after the trial commenced, with reasons		NA
Sample size	7a	How sample size was determined		22

	7b	When applicable, explanation of any interim analyses and stopping guidelines			22
Randomization					
Sequence generation	8a	Method used to generate the random allocation sequence			NA
	8b	Type of randomization; details of any restriction (such as blocking and block size)			NA
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned			NA
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions			NA
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how			NA
	11b	If relevant, description of the similarity of interventions			NA
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes			22,23
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses			22,23
Results					
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analyzed for the primary outcome			4,24
	13b	For each group, losses and exclusions after randomization, together with reasons			4,24
Recruitment	14a	Dates defining the periods of recruitment and follow-up			4
	14b	Why the trial ended or was stopped			NA
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group			15
Numbers analyzed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups			5
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)			5,16
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended			5,16
Ancillary analyses	18	Results of any other analyses performed, including subgroup			5,17

		analyses and adjusted analyses, distinguishing pre-specified from exploratory			
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	CONSORT-AI Extension 19	Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, justify why not.	18-23
Discussion					
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses			9,10
Generalizability	21	Generalizability (external validity, applicability) of the trial findings			10
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence			6-9
Other information					
Registration	23	Registration number and name of trial registry			2
Protocol	24	Where the full trial protocol can be accessed, if available			2
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders			10

^aWe strongly recommend reading this statement in conjunction with the CONSORT 2010 Explanation and Elaboration for important clarifications on all the items.

^bIndicates page numbers to be completed by authors during protocol development.

SUPPLEMENTARY REFERENCES

1. Wu, N., *et al.* Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Trans Med Imaging* **39**, 1184-1194 (2020).
2. Kim, H.E., *et al.* Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* **2**, e138-e148 (2020).
3. Salim, M., *et al.* External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. *JAMA Oncol* **6**, 1581-1588 (2020).
4. Hong, S., *et al.* Effect of Digital Mammography for Breast Cancer Screening: A Comparative Study of More than 8 Million Korean Women. *Radiology* **294**, 247-255 (2020).
5. Jo, H.M., *et al.* Prevalence of Women with Dense Breasts in Korea: Results from a Nationwide Cross-sectional Study. *Cancer Res Treat* **51**, 1295-1301 (2019).