

## Peer Review File

---

Artificial intelligence for breast cancer screening in mammography (AI-STREAM): Preliminary analysis of a prospective multicenter cohort study



**Open Access** This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

## REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author): Expert in breast cancer radiology and diagnosis, digital mammography, and AI

NCOMMS-24-38709

### Summary

In this prospective multicenter study conducted in South Korea, the investigators compare the diagnostic accuracy of breast radiologists in interpreting screening 2D mammograms, with and without AI-CAD, in a real-world, single-reading setting. The final cohort consisted of 24,543 patients with 140 screen-detected breast cancers. The CDR was significantly higher for breast radiologists with the use of AI-CAD (5.70 per 1000 versus 5.01 per 1000,  $p < 0.001$ ), with no significant difference in RRs. For general radiologists, CDR and RR were higher with the use of AI-CAD.

### Strengths

- A. This prospective study assesses the utility of AI for mammographic screening in a single-reading setting. Prospective research about AI for breast cancer screening, especially in single-reading settings, is very limited at this time.
- B. The study is multicenter in design, involving six academic hospitals in South Korea.
- C. The impact of AI on radiologists with varying levels of breast imaging experience (breast radiologists versus general radiologists) is evaluated.

### Weaknesses

- 1. The analysis is somewhat limited, focusing only on CDR and RR. Other screening mammography performance metrics to consider are PPV1, sensitivity, specificity, and false-negative rate.
- 2. For standalone AI performance, the authors should clarify whether or not a lesion-specific analysis was done. The use of exam scores alone (without a lesion-specific analysis) would overestimate the sensitivity of AI.
- 3. A summary of the types of cancers detected with AI-CAD support (that is, those cancers that were missed by breast radiologists without AI) should be added.
- 4. The authors can consider including a figure with an example of the AI interface and/or an imaging example of a cancer detected with the support of AI.
- 5. Grammar/syntax errors limit readability.

### Specific Comments

- 6. Abstract: It is stated that "The primary outcome was screen-detected breast cancer diagnosed at a one-year follow-up." This statement is somewhat unclear. Please consider: "The primary outcome was screen-detected breast cancer diagnosed within one year of a screening mammogram interpreted as positive."
- 7. Abstract: Please include recall rates for general radiologists with and without AI-CAD.
- 8. Introduction, lines 57-58: Please consider changing "several countries are currently implementing screening programs..." to "several countries have implemented screening programs..."
- 9. Introduction, line 61: Please consider removing "overdiagnosis" here. The main issue to address is

false-positives (unnecessary imaging and/or biopsies).

10. Introduction, page 66: Please remove “modalities” from “symptom modalities.”

11. Introduction, lines 68-70: Multiple studies have shown that DBT leads to lower false-positive rates.

12. Methods, Study Design, lines 103-106: Please clarify what is meant by “possessing expertise in breast imaging.”

13. Methods, Procedures, line 159: It is stated that five general radiologists participated in the simulation study. How many breast radiologists participated in the first part of the study?

14. Methods, Outcomes: Please clarify how CDR and RR were calculated. (Is CDR based on one-year follow-up data?)

15. Methods: The analysis is focused on CDR and RR. Other metrics to consider are PPV1, sensitivity, and specificity.

16. Methods: Assuming that the authors have one-year follow-up data, false-negative rate/interval cancer rate is an important metric to consider.

17. Methods: What statistical analysis software was used?

18. Results, Participant Selection Criteria and Characteristics, line 215: Please clarify why 24,545 mammograms (rather than 24,543 mammograms) were included.

19. Results, Performance of BRs with and without AI-CAD, starting with line 226: This paragraph is somewhat difficult to follow. Please summarize key points here.

20. Results, Performance of BRs with and without AI-CAD: Please include a summary of the types of cancers detected with AI-CAD support (that is, those cancers that were missed by breast radiologists without AI).

21. Results, Performance of Standalone AI: Please correct: “The CDR of AI standalone was 128...”

22. Results, Performance of Standalone AI: Was a lesion-specific analysis done to determine the performance of standalone AI, or is the analysis based on exam scores only?

23. Discussion, lines 255-256: CDR is typically reported as per 1000 examinations rather than 1000 participants.

24. Discussion, starting with line 258: This paragraph could be divided into shorter paragraphs.

25. Discussion: The limitations paragraph could be re-written in order to improve clarity.

26. Discussion: The authors could consider adding as a limitation that their study focuses on 2D mammography, not DBT.

27. Table 2: Please correct: “AL standalone.”

28. Table 3: This table is not entirely clear. Also, a similar table could also be included for general radiologists.

29. Figure 2: The inclusion of bilateral cancers makes the numbers somewhat confusing. Performance metrics (such as CDR) are typically reported based on number of examinations.

30. Figures: Please consider including an example of the AI interface and/or an example of a cancer detected with the support of AI.

Reviewer #2 (Remarks to the Author): Expert in AI-CAD, machine learning, and breast cancer mammography

This is a timely and interesting paper addressing one of the most important questions regarding AI for the detection of breast cancer from mammograms, namely how well does AI work when implemented clinically. The interim results from this ongoing clinical trial are impressive and the conclusions are sound. This paper has the potential to make a significant contribution to the literature.

There were some areas where the authors could improve the writing to make the paper clearer.

1. In general, please only include information that is pertinent to this study. It is necessary to describe the full study, but not, for example, information for secondary end points of the full study that are not relevant to the study described in the paper.
2. The description of the statistical analysis was not clear. I could not understand what exact statistical methods were used to measure and compare CDR and RR. Were the readings from the 5 radiologists pooled? If the data were pooled, what is the justification of ignoring reader variability? To me, that would be a major weakness.
3. It was not clearly stated how many BRs were used. It seems like 5, one for each site, but I couldn't find it in the text. A table with how many cases each BR and GR read and their CDR and RR with and without AI would be useful.
4. One of the limitations of the study that should be included in the discussion is that unlike the results for the BRs, the results for the GRs were not clinical interpretations. The authors state in the method section that the GRs were a simulation, which is good. In the discussion section, they should state that because the GRs' interpretations did not affect patient care, they may not reflect how the GRs would read clinically. A reference to the "laboratory effect" by Gur and colleagues could be given (Gur D, et al. (2008) The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 249 47-53).
5. It is surprising the study was powered for a cancer prevalence of 3.21, but the measured cancer prevalence was around 5. Could the authors please comment on this.
6. Line 47. Please change "affected" to "a significant change in"
7. Line 102. Did the radiologists also consent to participate in the study?
8. Line 103. Could you please change "from all participants" to "from all screening participants". From my viewpoint the radiologists were also participants.
9. Line 170. I find it surprising that for the AI score ranging from 0-100, the scores are presented to 2 decimal places. Why is that? Also, were there 4 scores (one for each view), or 2 scores (one for each breast)?
10. Line 171. Is it "can" or "was" for this study? If it was "can", each radiologists choose to toggle the heatmap on/off for each case?
11. Line 309. I don't think reference 22 is correct here.
12. Line 317. Please change "which causes" to "which may cause" or add a reference to support your assertion.
13. Lines 283-289. I am not sure why the authors are citing this study as it is completely different than their study. A more appropriate comparison would be: Letter H, et al. (2023) Use of Artificial Intelligence for Digital Breast Tomosynthesis Screening: A Preliminary Real-world Experience. *Journal of Breast Imaging*
14. Lines 319-327. I consider all these assertions to be speculation. I think it is good that the authors try to explain their results in the discussion section, but without proof for their assertions, they should

soften their statements; perhaps using wording like “it is possible” or the “results are consistent with”.

15. Line 319-320. “Although comparison with prior mammograms was possible in this simulation analysis ...” this implies that comparison images were not available to BRs? Is that true?

16. Lines 325-327. I consider this speculation. It could also be that GRs reading in simulation mode are reading more aggressively than they would if they were reading clinically. That is, their diagnostic performance did not change, but they shifted to a higher point on their ROC curve when reading with AI.

17. Did the BRs or GRs receive any training on using the AI tool? If so, please describe it in the paper.

18. I do not see any reason for Appendix 2.

19. Appendix 3 comes across a little too much like a marketing advertisement. Please remove the first three sentences and minimize the number of times Lunit INSIGHT MMG is used (AI system could be used instead). Also please remove “suggesting its potential to contribute to high reading accuracy in small- and medium-sized hospital ...” to the end of that paragraph.

20. Please remove the table in appendix 3 and include the required information in text format. A lot of the information in the table is already in the appendix.

21. Appendix 3 would be more informative if it gave details on how the AI results are presented to the radiologists. I strongly suggest that a picture of the overlay on a screening case be presented in the paper. How is the AI breast score related to the heat map, if there are multiple “hot” locations in the exam?

22. Appendix 4. I don’t understand what is presented here. Most of the paragraph is about needing 32,714 women screened. Then in the last paragraph, it is stated only 24,000 women are needed for the same cancer prevalence, power, and (I assume) alpha.

23. In appendix 5, the authors should only include the statistical methods that pertain to the analyses performed for this paper.

24. I do not understand eFigure 1. From the figure, it appears that the radiologists (I assume BR) read a de-identified image? How do they read a clinical image that has been de-identified?

25. I do not understand the need for eFigure 2.

Reviewer's Comments:

Reviewer #1 (Remarks to the Author)

NCOMMS-24-38709

Strengths

A. This prospective study assesses the utility of AI for mammographic screening in a single-reading setting. Prospective research about AI for breast cancer screening, especially in single-reading settings, is very limited at this time.

B. The study is multicenter in design, involving six academic hospitals in South Korea.

C. The impact of AI on radiologists with varying levels of breast imaging experience (breast radiologists versus general radiologists) is evaluated.

Weaknesses

1. The analysis is somewhat limited, focusing only on CDR and RR. Other screening mammography performance metrics to consider are PPV1, sensitivity, specificity, and false-negative rate.

2. For standalone AI performance, the authors should clarify whether or not a lesion-specific analysis was done. The use of exam scores alone (without a lesion-specific analysis) would overestimate the sensitivity of AI.

3. A summary of the types of cancers detected with AI-CAD support (that is, those cancers that were missed by breast radiologists without AI) should be added.

4. The authors can consider including a figure with an example of the AI interface and/or an imaging example of a cancer detected with the support of AI.

5. Grammar/syntax errors limit readability.

**: The authors thank the review for their thoughtful review of the manuscript and sincerely believe that the manuscript has been strengthened. Please find below our response to the reviewer's specific comments, where the overall manuscript has been proofread by a native speaker.**

Specific Comments

6. Abstract: It is stated that "The primary outcome was screen-detected breast cancer diagnosed at a one-year follow-up." This statement is somewhat unclear. Please consider: "The primary outcome was screen-detected breast cancer diagnosed within one year of a screening mammogram interpreted as positive."

**: It was revised based on reviewer's comments.**

7. Abstract: Please include recall rates for general radiologists with and without AI-CAD.

**: Based on reviewer comments, the following has been added.**

***Similar trends were observed for general radiologists, with a significant higher CDR of 26.4% in those with AI-CAD (n=120 [4.89 %]) compared to those without AI (n=95 [3.87 %]; p <0.001). However, the RRs were significantly difference (p <0.001).***

8. Introduction, lines 57-58: Please consider changing "several countries are currently implementing screening programs..." to "several countries have implemented screening programs..."

**: It was revised based on reviewer's comments.**

9. Introduction, line 61: Please consider removing "overdiagnosis" here. The main issue to address is false-positives (unnecessary imaging and/or biopsies).

**: It was revised based on reviewer's comments.**

10. Introduction, page 66: Please remove "modalities" from "symptom modalities."

**: It was revised based on reviewer's comments.**

11. Introduction, lines 68-70: Multiple studies have shown that DBT leads to lower false-positive rates.

: Based on reviewer comments, the following has been added including 3 references.

*Multiple studies have shown that digital breast tomosynthesis (DBT) has a lower recall rate (RR) and higher cancer detection rate (CDR), leading to lower false-positive rates than digital mammography(DM) <sup>8-10</sup>.*

12. Methods, Study Design, lines 103-106: Please clarify what is meant by “possessing expertise in breast imaging.”

: It was modified as follows:

*having specializing expertise in breast imaging.*

13. Methods, Procedures, line 159: It is stated that five general radiologists participated in the simulation study. How many breast radiologists participated in the first part of the study?

: The participating hospitals were six breast imaging specialists from six university hospitals. Four university hospitals recruited patients and conducted readings in 2021-2022, and one hospital recruited patients and conducted readings in 2021 and the other hospital conducted readings in 2022. To better reflect this, this was modified as follows and added the supplementary eTable 1,

*In this study, mammography interpretation was performed by a breast specialized radiologists at each of the six university hospitals in a real clinical setting (eTable1). A radiologist specializing in breast imaging (i.e. BR) was defined as a radiologist with > 10 years of experience at an university hospital, having specialized expertise in breast imaging.*

| BR<br>(A-F)        | DM screens without/ with AI,<br>N (%) (N=24,545) | GR<br>(G-K) | DM screens without/ with AI,<br>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)                                      |

**eTable 1.** Cases read mammography without/with AI by each radiologist.

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

14. Methods, Outcomes: Please clarify how CDR and RR were calculated. (Is CDR based on one-year follow-up data?)

: CDR was measured by screening-detected cancer confirmed by pathology in 1-year follow up data.

The following information has been added:

*All statistical analyses including CDRs and RR were taken place based on a 1-year follow-up from all the study participants. All statistical analyses were performed using R version 4.3.3 (R Foundation for statistical computing, Vienna, Austria).*

15. Methods: The analysis is focused on CDR and RR. Other metrics to consider are PPV1, sensitivity, and specificity.

: Thank you for your comment. This preliminary outcome of this preliminary analysis was pathologically diagnosed screening-detected breast cancer at the time of 1-year follow-up after completion of the last participant enrollment. For the final analysis, which is planned in the future, a 2-year follow-up is needed to evaluate the true impact of AI-CAD use on interval cancers detected after two years (biennial screening interval in South Korea). The national cancer registry databases will be planned to be reviewed after 2026 for calculation of diagnostic accuracy endpoints that include the metrics the reviewer noted, which are based on participants diagnosed with breast cancer within two years from screening including interval cancer. The final results of the AI-STREAM study including sensitivity, positive predictive value, specificity, and interval cancer rate will be planned to be announced after 2026 on the analysis of data to linked to the National Cancer Register. The PPV1 for screening-detected cancers by breast radiologists, which is currently calculable, has been added.

16. Methods: Assuming that the authors have one-year follow-up data, false-negative rate/interval cancer rate is an important metric to consider.

: Thank you for your comment. For the final analysis, which is planned in the future, a 2-year follow-up is needed to evaluate the true impact of AI-CAD use on interval cancers detected after two years (biennial screening interval in South Korea). The national cancer registry databases will be planned to be reviewed after 2026 for calculation of diagnostic accuracy endpoints that include the metrics the reviewer noted, which are based on participants diagnosed with breast cancer within two years from screening including interval cancer.

17. Methods: What statistical analysis software was used?

: It was added as follows,

*All statistical analyses including CDRs and RR were taken place based on a 1-year follow-up from all the study participants. All statistical analyses were performed using R version 4.3.3. (R Foundation for statistical Computing, Vienna, Austria).*

18. Results, Participant Selection Criteria and Characteristics, line 215: Please clarify why 24,545 mammograms (rather than 24,543 mammograms) were included.

: It was modified as follows;

*Finally, a total of 24,545 mammograms from 24,543 participants including 2 cases of bilateral breast cancer were analyzed in this study.*

19. Results, Performance of BRs with and without AI-CAD, starting with line 226: This paragraph is somewhat difficult to follow. Please summarize key points here.

: It was modified as follows,

**Overall, BR with AI-CAD detected 140 screening-detected cancers and 17 more cancers than BR without AI-CAD, which detected 123 breast cancers.**

20. Results, Performance of BRs with and without AI-CAD: Please include a summary of the types of cancers detected with AI-CAD support (that is, those cancers that were missed by breast radiologists without AI).

: Following the reviewer's comments, the results of the 17 lesions additionally diagnosed using AI-CAD were summarized in Table 3 and the result as follows;

*Of the 17 cases found in BR with AI-CAD compared to BR without AI-CAD, six cases were DCIS ( $p=0.009$ ) and 11 cases were invasive cancers ( $p < 0.001$ ). Among IDCs not otherwise specified, five cases of low-grade IDC were significantly frequently detected using AI-CAD ( $p=0.009$ ).*

*Furthermore, BR with AI-CAD resulted in a significant increase detection of cancer less than 20mm ( $p=0.002$ ) and more than 20 mm ( $p=0.035$ ). BR with the assistance of AI-CAD significantly increased the detection of negative node metastasis ( $p=0.001$ ) and luminal A subtype ( $p=0.002$ ).*

21. Results, Performance of Standalone AI: Please correct: “The CDR of AI standalone was 128...”  
: **It was revised based on reviewer's comments.**

22. Results, Performance of Standalone AI: Was a lesion-specific analysis done to determine the performance of standalone AI, or is the analysis based on exam scores only?  
: **Thank you for your comment. Abnormal scores from 0 to 100, rounded to two decimal places, were derived from the raw data. An abnormal score of  $\geq 10$  is indicated as “high” and is considered a positive test result in this program. For cases with multiple abnormalities, each score is displayed on the image, but the highest AI abnormal score is shown at the bottom of the mammography image. It was planned that the radiologist would record the recalled location as right, left or both whether or not AI assistance was used. The area flagged by the AI abnormality score was then compared to the recalled area, and if confirmed as cancer, lesion-specific analysis was performed. We accordingly added the following to the result section of ‘performance of standalone AI and radiologist with and without AI-CAD’**

*The areas marked with AI abnormal scores were checked to see if they matched the recalled areas and the areas confirmed as cancer, followed by lesion-specific analysis.*

23. Discussion, lines 255-256: CDR is typically reported as per 1000 examinations rather than 1000 participants.  
: **It was modified as ‘examination’.**

24. Discussion, starting with line 258: This paragraph could be divided into shorter paragraphs.  
: **The paragraphs were divided into two shorter paragraphs.**

25. Discussion: The limitations paragraph could be re-written in order to improve clarity.  
: **Limitation was rewritten to improve clarity.**

26. Discussion: The authors could consider adding as a limitation that their study focuses on 2D mammography, not DBT.  
: **It was added as a limitation of the study as follows.**

*Third, this study used DM to interpret the screening effect of AI-CAD. Although DM and DBT are used for routine breast screening, use of DBT has been rapidly disseminated for breast cancer screening. Further study should be needed to assess the performance of AI-CAD in the interpretation of DBT.*

27. Table 2: Please correct: “AL standalone.”  
: **It was corrected as ‘AI’.**

28. Table 3: This table is not entirely clear. Also, a similar table could also be included for general radiologists.  
: **The table was modified to be clearer, and the results were also modified compared to the table. A similar table for general radiologists was included as Table 4 following the reviewer’s comment.**

29. Figure 2: The inclusion of bilateral cancers makes the numbers somewhat confusing. Performance metrics (such as CDR) are typically reported based on number of examinations.  
: **Since bilateral lesions in 2 cases could not be excluded from the analysis, 2 examinations had to be added to the participant. This was added as the limitation, and the CDR analysis was also analyzed by modifying it to examination rather than participant.**

30. Figures: Please consider including an example of the AI interface and/or an example of a cancer

detected with the support of AI.

**:A Figure of cancer cases diagnosed in BR with AI-CAD but not detected in BR without AI-CAD has been added to Figure 3.**

Reviewer #2 (Remarks to the Author): Expert in AI-CAD, machine learning, and breast cancer mammography

This is a timely and interesting paper addressing one of the most important questions regarding AI for the detection of breast cancer from mammograms, namely how well does AI work when implemented clinically. The interim results from this ongoing clinical trial are impressive and the conclusions are sound. This paper has the potential to make a significant contribution to the literature.

**: The authors thank the review for their thoughtful review of the manuscript and sincerely believe that the manuscript has been strengthened. Please find below our response to the reviewer's specific comments.**

There were some areas where the authors could improve the writing to make the paper clearer.

1. In general, please only include information that is pertinent to this study. It is necessary to describe the full study, but not, for example, information for secondary end points of the full study that are not relevant to the study described in the paper.

**: Since this study is a preliminary analysis, a description of the full study design was included. Following the reviewer's comment, the relevant information has been described in previously published studies (Chang, Y. W., et al. (2022). "Artificial Intelligence for Breast Cancer Screening in Mammography (AI-STREAM): A Prospective Multicenter Study Design in Korea Using AI-Based CADe/x." J Breast Cancer 25(1): 57-68), so only brief mentions of study endpoints were included. Supplementary Figure 2 was also excluded.**

2. The description of the statistical analysis was not clear. I could not understand what exact statistical methods were used to measure and compare CDR and RR. Were the readings from the 5 radiologists pooled? If the data were pooled, what is the justification of ignoring reader variability? To me, that would be a major weakness.

**: To consider reader variability, we conducted the analysis using a statistical analysis method called the Generalized estimating equation.**

It was mentioned that reader variability was evaluated using a Generalized estimating equation.

3. It was not clearly stated how many BRs were used. It seems like 5, one for each site, but I couldn't find it in the text. A table with how many cases each BR and GR read and their CDR and RR with and without AI would be useful.

**: The participating hospitals were six breast imaging specialists from six university hospitals. Four university hospitals recruited patients and conducted readings in 2021-2022, and one hospital recruited patients and conducted readings in 2021 and the other hospital conducted readings in 2022.**

*For the general radiologists (GR), five radiologists, who were not breast imaging specialists, independently reviewed 24,545 cases in 2021-2022. All 24,545 cases were first read without AI, followed by a second reading with AI. Additional information for cases read by each radiologist was*

*provided in a supplementary table 1.*

4. One of the limitations of the study that should be included in the discussion is that unlike the results for the BRs, the results for the GRs were not clinical interpretations. The authors state in the method section that the GRs were a simulation, which is good. In the discussion section, they should state that because the GRs' interpretations did not affect patient care, they may not reflect how the GRs would read clinically. A reference to the "laboratory effect" by Gur and colleagues could be given (Gur D, et al. (2008) The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 249 47-53).

**: The following has been added as per the reviewer's comments, and references have also been added.**

*In addition, the results may not accurately reflect how the GRs would interpret mammography in a clinical setting due to the retrospective simulation nature of the readings, despite the use of prospectively enrolled data. It has been reported that retrospective laboratory experiments may not represent expected performance levels or inter-reader variability during clinical interpretations of the same set of mammograms in a real-world clinical environment<sup>25</sup>.*

5. It is surprising the study was powered for a cancer prevalence of 3.21, but the measured cancer prevalence was around 5. Could the authors please comment on this.

**: The reference study, which targeted more than 8 million Korean women who underwent digital mammography from 2011 to 2013, reported a cancer prevalence of 3.21 per 1000 screenings. (Ref: Hong S, et al. Effect of Digital Mammography for Breast Cancer Screening: A Comparative Study of More than 8 Million Korean Women. *Radiology* 2020; 294: 247-55).**

**It is difficult to clearly explain why the prevalence rate in this prospective study is higher. However, we cautiously speculate that the screening group in this study, conducted at an academic hospital, may have had a higher cancer detection rate compared to screening groups in past studies.**

**Generally, while radiologists interpreting screening mammograms have varying levels of experience, the higher cancer detection rate in this study for breast radiologists appear to be because interpreters were breast imaging specialists with over 10 years of experience. In another study, breast imaging specialists had an average cancer detection rate of 5.2 cancers per 1000 mammograms (ranging from 3 to 10 per 1000 for the individual radiologists).**

**The fact that the cancer detection rate was significantly higher with AI compared to without AI, even in a group with a high cancer detection rate like breast radiologists, highlights the effectiveness of AI.**

6. Line 47. Please change "affected" to "a significant change in"

**: It was changed based on the reviewer's comments.**

7. Line 102. Did the radiologists also consent to participate in the study?

**: This study was conducted prospectively, enrolling participants from multi-centers. The breast radiologists who interpreted mammography in the study served as research directors at their respective hospitals, managing research funds, recruiting research participants, and conducting the overall research process. All radiologists consent to the interpretation of mammography.**

8. Line 103. Could you please change “from all participants” to “from all screening participants”. From my viewpoint the radiologists were also participants.

**: It was changed based on the reviewer's comments.**

9. Line 170. I find it surprising that for the AI score ranging from 0-100, the scores are presented to 2 decimal places. Why is that? Also, were there 4 scores (one for each view), or 2 scores (one for each breast)?

**: The AI produces a score per abnormality found on a screening examination, regardless of the breast side, where only abnormalities with a score greater than 10 are displayed to the radiologists. The AI scores originally output a continuous value from 0 to 1, where this score is then multiplied by 100 to have a range of 0 to 100; hence only 2 decimal places of the original score were used (eg, the original score of 0.15 would correspond to 15 and this figure of 15 would be displayed to the radiologists for interpretation). If there are multiple abnormalities with a score greater than 10, then the corresponding number of abnormalities is shown. Moreover, the highest (or maximum) AI score is defined as the AI score for that examination, except for the 2 bilateral cases, where each side was considered separately.**

10. Line 171. Is it “can” or “was” for this study? If it was “can”, each radiologists choose to toggle the heatmap on/off for each case?

**: It was changed to ‘can’’. The option can be selected during the setup process after the initial AI installation. In this study, the radiologist enabled the heatmap feature for use in their interpretation.**

11. Line 309. I don't think reference 22 is correct here.

**: Thank you for checking in detail. There was an error in the reference number. The referenced journal is as follows and the error has been corrected.**

***24. Sickles, E.A., Wolverton, D.E. & Dee, K.E. Performance parameters for screening and diagnostic mammography: specialist and general radiologists. Radiology 224, 861-869 (2002).***

12. Line 317. Please change “which causes” to “which may cause” or add a reference to support your assertion.

**: It was added ‘may’ based on the reviewer's comments.**

13. Lines 283-289. I am not sure why the authors are citing this study as it is completely different than their study. A more appropriate comparison would be: Letter H, et al. (2023) Use of Artificial Intelligence for Digital Breast Tomosynthesis Screening: A Preliminary Real-world Experience. Journal of Breast Imaging:

**: Prospective clinical trials are essential for demonstrating the effectiveness of AI in mammography. However, few prospective studies have evaluated the clinical impact of AI-CAD, and most are limited to the dual-reading strategy in Europe. Two notable examples are the ScreenTrust CAD and MASAI studies, both registered in ClinicalTrials, which have reported interim results. The population-based prospective ScreenTrust CAD study demonstrated that dual reading by a single radiologist and AI was non-inferior to standard dual reading by two radiologists for cancer detection. Another prospective study, the MASAI study used AI scores to triage mammography for single- or dual-reading strategies, despite differences in reading**

methods. Independent reading was conducted when the AI score was low, and cancer detection rates increased compared to the standard dual-reading control group, with a reduced screening workload. While many European countries use a dual-read strategy, some regions outside Europe, as well as certain private practices within Europe, have adopted a single-read strategy. This highlights the need for prospective evidence of AI's effectiveness in real-world single-read settings. Our AI-STREAM study aims to assess the usefulness of AI assistance in such single-reading environments in prospective clinical settings.

Lastly, the DBT study by Letter H et al., which the reviewer mentioned, has been added to the discussion. We thank the reviewer for directing us to this study.

14. Lines 319-327. I consider all these assertions to be speculation. I think it is good that the authors try to explain their results in the discussion section, but without proof for their assertions, they should soften their statements; perhaps using wording like “it is possible” or the “results are consistent with”.

**: It was changed based on the reviewer's comments.**

15. Line 319-320. “Although comparison with prior mammograms was possible in this simulation analysis ...” this implies that comparison images were not available to BRs? Is that true?

**: As with BR, comparison with the prior image was possible in this simulation study. the sentence was modified as follows.**

*as like real clinical setting of BR reading*

16. Lines 325-327. I consider this speculation. It could also be that GRs reading in simulation mode are reading more aggressively than they would if they were reading clinically. That is, their diagnostic performance did not change, but they shifted to a higher point on their ROC curve when reading with AI.

**: I agree that simulation tests cannot fully replicate real clinical settings, and more aggressive readings may occur. We have added this point to the discussion. ROC analysis is challenging at this stage and will be conducted on the final dataset (once we have data on interval cancers as well).**

17. Did the BRs or GRs receive any training on using the AI tool? If so, please describe it in the paper.

**: AI was integrated into the images to provide assistance, so no additional training was necessary. Neither BR nor GR had any previous experience using AI for interpretation.**

18. I do not see any reason for Appendix 2.

**: Since this study was conducted prospectively with participant enrollment, additional information was included regarding the necessary data monitoring, cleaning, and retention.**

19. Appendix 3 comes across a little too much like a marketing advertisement. Please remove the first three sentences and minimize the number of times Lunit INSIGHT MMG is used (AI system could be used instead). Also please remove “suggesting its potential to contribute to high reading accuracy in small- and medium-sized hospital ...” to the end of that paragraph.

**: It was revised according to the reviewer's comments.**

20. Please remove the table in appendix 3 and include the required information in text format. A lot of the information in the table is already in the appendix.

**: It was removed from the table in appendix 3 according to the reviewer's comments.**

21. Appendix 3 would be more informative if it gave details on how the AI results are presented to the radiologists. I strongly suggest that a picture of the overlay on a screening case be presented in the paper. How is the AI breast score related to the heat map, if there are multiple “hot” locations in the exam?

**: AI figure was added. The heat map shows that when there are multiple hot locations, they are displayed from blue for low scores to red for high scores, with the highest AI abnormal score displayed.**

22. Appendix 4. I don't understand what is presented here. Most of the paragraph is about needing 32,714 women screened. Then in the last paragraph, it is stated only 24,000 women are needed for the same cancer prevalence, power, and (I assume) alpha.

**: The target sample size was 32,714 participants, corresponding to approximately 16,000 participants per year. The total number of expected participants was however adjusted from the initial study design due to the COVID-19 pandemic, but no effect on the primary study endpoint was observed. Assuming 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. Additional information on related content is provided in Appendix 4.**

23. In appendix 5, the authors should only include the statistical methods that pertain to the analyses performed for this paper.

**: It was corrected to only include the statistical methods of analyses performed for this paper.**

24. I do not understand eFigure 1. From the figure, it appears that the radiologists (I assume BR) read a de-identified image? How do they read a clinical image that has been de-identified?

**: This study was designed as a multi-center prospective study, utilizing a cloud-based imaging data management platform for data management. Anonymization enabled GR to read the same images as BR replicating the process used in real clinical practice with stored images.**

**This describes the process of anonymizing the images of participants who consented to the study and uploading them to Best Image so that AI-CAD could analyze them. Anonymization of the images was necessary to protect personal information. The Snupi program processed the mammograms, performing various operations (eg, search, inquiry, de-identification, separation) as well as transmission functions on Digital Imaging and Communication in Medicine (DICOM) files. The mammograms were exported to the platform after de-identifying the participant information and assigning identification numbers.**

**As part of the standard single screening procedure, a radiologist reviewed the mammograms, comparing them with previous mammograms when available, and recorded the results. This was followed by the automatic presentation of AI-CAD results (abnormality score and marks) for further review and documentation. The radiologist then reviewed mammograms with the**

**assistance of AI-CAD and the final decision for recall was recorded. The Figure above describes the process of applying AI to images of enrolled participants.**

25. I do not understand the need for eFigure 2.  
**:eFigure 2 was excluded.**

**Thank you again for the reviewers' thoughtful comments on the manuscript.**

## REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

NCOMMS-24-38709A

### Summary

In this prospective multicenter study conducted in South Korea, the investigators compare the diagnostic accuracy of breast radiologists in interpreting screening 2D mammograms, with and without AI-CAD, in a real-world, single-reading setting. The final cohort consisted of 24,543 patients with 140 screen-detected breast cancers. The CDR was significantly higher for breast radiologists with the use of AI-CAD (5.70 per 1000 versus 5.01 per 1000,  $p < 0.001$ ), with no significant difference in RRs. For general radiologists, CDR and RR were higher with the use of AI-CAD.

### Strengths

- A. This prospective study assesses the utility of AI for mammographic screening in a single-reading setting. Prospective research about AI for breast cancer screening, especially in single-reading settings, is very limited at this time.
- B. The study is multicenter in design, involving six academic hospitals in South Korea.
- C. The impact of AI on radiologists with varying levels of breast imaging experience (breast radiologists versus general radiologists) is evaluated.

### Weaknesses

No major weaknesses are noted. The authors have thoroughly addressed my comments and suggestions and have overall improved the manuscript. Some minor suggestions are listed below.

### Specific Comments

1. Abstract: In the third sentence, please add a comma after "with and without an AI-based computer-aided detection..."
2. Abstract: For general radiologists, it is stated that "the RRs were significantly different" – what is the direction of this difference? That is, was RR higher with or without AI-CAD?
3. Introduction, last sentence of first paragraph: Please change to: "Multiple studies have shown that digital breast tomosynthesis (DBT) combined with digital mammography (DM)..."
4. Introduction, last paragraph: Please change to: "...which could support computer-assisted detection (CAD) to mark suspicious findings..."
5. Introduction, last paragraph: Is BR considered to be a standard abbreviation for breast radiologists? I haven't seen it in the literature before.
6. Methods, Participants: The last sentence should be corrected to state that mammography was performed by a technologist and interpreted by a single radiologist.
7. Methods: In the paragraph about the AI-CAD system (Lunit), please clarify that the examination score reflects the highest lesion Abnormality Score and that examinations with scores of 10 or higher are considered positive.
8. Results, Performance of BRs with and without AI-CAD: Please add a p-value for PPV1s of BRs with and without AI-CAD.

9. Results, Performance of BRs with and without AI-CAD: There are some redundancies in the second paragraph (e.g., node-negative metastasis is mentioned twice).
10. Discussion, second paragraph: Please consider mentioning the MASAI trial here (with discussion of ScreenTrustCAD) rather than later on in the Discussion.
11. Discussion, second paragraph: Please remove comma after "Recent advancements in AI."
12. Discussion, third paragraph: Please correct this sentence: "The preliminary analysis of the current AI-STREAM study..."
13. Discussion, third paragraph: May consider mentioning that radiologists can compare to prior mammograms but the AI algorithm does not have a temporal analysis feature.
14. Discussion: The Discussion section is too long and should be condensed. (Also please note that the paragraph beginning with "One prior study of differences in screening mammography..." is too long.)
15. Figure 3: What is the finding on the mammogram?

Reviewer #2 (Remarks to the Author):

In general, the authors were responsive to my previous review. I do have two remaining comments:

1. To my point #4, the authors responded that they added to the text: "In addition, the results may not accurately reflect how the GRs would interpret mammography in a clinical setting due to the retrospective simulation nature of the readings, despite the use of prospectively enrolled data. It has been reported that retrospective laboratory experiments may not represent expected performance levels or inter-reader variability during clinical interpretations of the same set of mammograms in a real-world clinical environment<sup>25</sup>." I do not see this text in the limitation section, nor the citation in the reference list.
2. To my point #5, I am satisfied with the author's explanation on the difference in the prevalence rate between their study and national standard. I would strongly suggest they include this explanation in the paper. The large discrepancy needs to be addressed.

## REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

NCOMMS-24-38709A

### Summary

In this prospective multicenter study conducted in South Korea, the investigators compare the diagnostic accuracy of breast radiologists in interpreting screening 2D mammograms, with and without AI-CAD, in a real-world, single-reading setting. The final cohort consisted of 24,543 patients with 140 screen-detected breast cancers. The CDR was significantly higher for breast radiologists with the use of AI-CAD (5.70 per 1000 versus 5.01 per 1000,  $p < 0.001$ ), with no significant difference in RRs. For general radiologists, CDR and RR were higher with the use of AI-CAD.

### Strengths

- A. This prospective study assesses the utility of AI for mammographic screening in a single-reading setting. Prospective research about AI for breast cancer screening, especially in single-reading settings, is very limited at this time.
- B. The study is multicenter in design, involving six academic hospitals in South Korea.
- C. The impact of AI on radiologists with varying levels of breast imaging experience (breast radiologists versus general radiologists) is evaluated.

### Weaknesses

No major weaknesses are noted. The authors have thoroughly addressed my comments and suggestions and have overall improved the manuscript. Some minor suggestions are listed below.

### Specific Comments

1. Abstract: In the third sentence, please add a comma after "with and without an AI-based computer-aided detection..."

**: A comma was added.**

2. Abstract: For general radiologists, it is stated that "the RRs were significantly different"  
– what is the direction of this difference? That is, was RR higher with or without AI-CAD?

**: It was corrected as follows,**

**the RRs were significantly high in those with AI-CAD ( $p < 0.001$ ).**

3. Introduction, last sentence of first paragraph: Please change to: "Multiple studies have shown that digital breast tomosynthesis (DBT) combined with digital mammography (DM) ..."

**: It was corrected.**

**Multiple studies have shown that digital breast tomosynthesis (DBT) combined with digital mammography (DM) has a lower recall rate (RR) and higher cancer detection rate (CDR), leading to lower false-positive rates than digital mammography (DM).**

4. Introduction, last paragraph: Please change to: "...which could support computer-assisted detection (CAD) to mark suspicious findings..."

**: It was corrected.**

**With substantial promise in this space, artificial intelligence (AI) has been developed and validated with promising applications in mammography screening, which could support computer-assisted detection (CAD) to mark suspicious findings depending on the risk of malignancy.**

5. Introduction, last paragraph: Is BR considered to be a standard abbreviation for breast radiologists? I haven't seen it in the literature before.

**: Although it is not an abbreviation for breast radiologist, it was used repeatedly in this study, so the breast radiologist was named BR, and the general radiologist was named GR.**

6. Methods, Participants: The last sentence should be corrected to state that mammography was performed by a technologist and interpreted by a single radiologist.

**: It was corrected.**

**During breast cancer screening, mammography was performed by technologists and**

**interpreted by a single radiologist (ie, single-reading strategy) using two standard craniocaudal and mediolateral oblique views of each breast using DM, which is the standard procedure for interpreting mammograms in South Korea.**

7. Methods: In the paragraph about the AI-CAD system (Lunit), please clarify that the examination score reflects the highest lesion Abnormality Score and that examinations with scores of 10 or higher are considered positive.

**: It was corrected.**

**The highest lesion abnormality score was reflected and examinations with scores of 10 or higher were considered positive.**

8. Results, Performance of BRs with and without AI-CAD: Please add a p-value for PPV1s of BRs with and without AI-CAD.

**: It was added.**

**The PPV1 of BRs with AI-CAD was higher (12.6) compared to BRs without AI-CAD (11.2) ( $p < 0.001$ ).**

9. Results, Performance of BRs with and without AI-CAD: There are some redundancies in the second paragraph (e.g., node-negative metastasis is mentioned twice).

**: The duplicated parts were deleted and organized as follows.**

**Interpreting mammograms with AI-CAD detected 6 additional cases of DCIS and 11 additional cases of invasive cancer, leading to notable increase in detection for both DCIS ( $p=0.009$ ) and invasive cancers ( $p<0.001$ ). Furthermore, assistance with AI-CAD resulted in a significant increase in the detection of small-sized cancer less than 20 mm ( $p=0.002$ ), node-negative metastasis ( $p=0.001$ ), luminal A subtype ( $p=0.002$ ), and lower grade IDC NOS ( $p =0.009$ ) when compared without AI-CAD (Table 3).**

10. Discussion, second paragraph: Please consider mentioning the MASAI trial here (with discussion of ScreenTrustCAD) rather than later on in the Discussion.

**: The mentioning the MASAI trial was moved to the second paragraph.**

11. Discussion, second paragraph: Please remove comma after "Recent advancements in AI."

**: It was corrected. Comma was removed.**

12. Discussion, third paragraph: Please correct this sentence: "The preliminary analysis of the current AI-STREAM study..."

**: It was corrected.**

**The preliminary analysis of the current AI-STREAM study would be expected to specifically address this knowledge gap, by not only utilizing real-world screening data collected prospectively from multiple centers, but also increased CDRs and unaffected RRs in BRs with AI-CAD.**

13. Discussion, third paragraph: May consider mentioning that radiologists can compare to prior mammograms but the AI algorithm does not have a temporal analysis feature.

**: It was corrected.**

**Additionally, radiologists can compare current mammograms to prior ones, whereas the AI algorithm does not have a temporal analysis feature for previous images. That gives radiologists with AI-CAD interpretation an advantage over compared to stand-alone AI-CAD.**

14. Discussion: The Discussion section is too long and should be condensed. (Also please note that the paragraph beginning with "One prior study of differences in screening mammography..." is too long.)

**: The following discussion section was shortened and was moved some of the content to limitations.**

15. Figure 3: What is the finding on the mammogram?

**: It was corrected as follows.**

**In test 2, BR with AI-CAD interpreted mammography of focal asymmetry in the left upper inner breast as malignant scale 4 and decided on a recall.**

Reviewer #2 (Remarks to the Author):

In general, the authors were responsive to my previous review. I do have two remaining comments:

1. To my point #4, the authors responded that they added to the text: "In addition, the results may not accurately reflect how the GRs would interpret mammography in a clinical setting due to the retrospective simulation nature of the readings, despite the use of prospectively enrolled data. It has been reported that retrospective laboratory experiments may not represent expected performance levels or inter-reader variability during clinical interpretations of the same set of mammograms in a real-world clinical environment<sup>25</sup>." I do not see this text in the limitation section, nor the citation in the reference list.

**: It was added in limited session with reference.**

**Third, the results of GR interpretation may not accurately reflect how the GRs would interpret mammography in a clinical setting due to the retrospective simulation nature of the readings, despite the use of prospectively enrolled data. It has been reported that retrospective laboratory experiments may not represent expected performance levels or inter-reader variability during clinical interpretations of the same set of mammograms in a real-world clinical environment<sup>25</sup>**

2. To my point #5, I am satisfied with the author's explanation on the difference in the prevalence rate between their study and national standard. I would strongly suggest they include this explanation in the paper. The large discrepancy needs to be addressed.

**: It was added the following in discussion.**

**In this prospective study, the prevalence is higher than the expected national standard. We speculate that the elevated prevalence rate in our study may be due to the screening group being conducted at academic hospitals, and screening mammograms being interpreted by breast imaging specialists.**

Thank you very much for consideration of our manuscript.