nature portfolio

Peer Review File

Comparison of AI-integrated pathways with human-AI interaction in population mammographic screening for breast cancer

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author): Expert in artificial intelligence, machine learning, and computer vision;

Overview
The authors introduce a new breast cancer detection model from screening mammograms, BRAIx, trained on a large cohort of patients (623k) from the ADMANI dataset built from participants in the BreastScreen Australia program from 2013-2019. BRAIx achieves impressive performance on a held-out test set of 93k patients (.932 AUC).

The authors then use BRAIx as an AI reader in multiple simulation settings. They replace one reader (reader replacement), selectively request human reading (band pass), and use the model to select 1 or 2 human readers (triage). All three simulations reduce the workload on human readers. The first two also improve performance of the screening protocol as measured by sensitivity and specificity. The authors suggest that these results mean that at least the reader replacement setting should be investigated in clinical trials.

Review Summary
The BRAIx model shows impressive performance on the new ADMANI dataset and the (relatively small) external baselines used. However, because almost no detail is given about the implementation (to protect intellectual property, see Code Availability). As a result, there is limited scientific contribution from the first half of the paper.

The latter half of the paper focuses on simulated reader studies of AI reader integration to the mammogram reading protocol in Australia. The simulations use BRAIx as the AI reader. These approaches all seem plausible and show promising performance in the simulated studies. As such, they are of interest to the scientific community, though they are not clearly positioned against previous simulation studies.

The authors suggest the most likely path forward is the use of the AI-reader in conjunction with a single human reader. The results for this setting are promising. However, in the simulation, they rely on downstream performance of a human third reader holding constant after the introduction of an AI reader in the first round. The authors investigate the effects of AI performance on human readers in the triage setting, but not the reader replacement setting, and then only under the assumption that AI readers will improve human performance. It is unclear why we should assume this positive impact and why we should only consider AI-human interaction effects in the triage setting. Without investigating the potential impact of AI readers in both settings and as potentially negative for human readers, the import of this assumption throws into question the validity of the results (see “second reader as third reader” setting in Appendix 1 - Table 9 for an example where different assumptions lead to ambiguous impacts on performance).

While the paper provides valuable insights, it would benefit from rewriting that 1) clarifies the contributions and methods before introducing and interpreting results; 2) centers the AI-reader

integration rather than BRAIx; and 3) considers potential positive and negative impact of AI readers on downstream human readers in all settings.

Strengths

· The size and diversity of the datasets tested (>400k patients).
· The performance of the BRAIx mammogram reader model is excellent. It generalizes to multiple (small) external datasets.
· The AI-human reader settings demonstrate viability, and are likely of interest for prospective trials, as the authors suggest.

Weaknesses

· Prospective study omits interval cancers because the follow-up data is not available and is limited to a single site.
· The authors suggest the reader replacement setting is the best candidate for prospective trials, but it's unclear how much of the results in the AI-reader replacement setting hinge on the assumption that third reader performance would be held constant. The method for implementing human-reader performance on the simulated third reader is not described.
· Almost no detail on BRAIx implementation is provided. The BRAIx model relies on a new, large dataset for training (ADMANI). As a result, it's unclear if BRAIx's performance is a function of its architecture or training data, and no attempt is made to disentangle the effects.

General Comments

· I cannot find a place where the positive and negative classes are explicitly defined. The task and evaluation are not clearly defined before results are introduced. It might be helpful to include an abbreviated Methods section with key points from the appendix after the Introduction.
· Writing in the Intro and Discussion is difficult to follow. Subsections or breaking out parts (like creating a Related Works instead of embedding it into the Intro) would help.
· A bias towards assuming positive AI-reader human-reader interaction might be reasonable in light of [22], but that evidence is indirect and does not eliminate the possibility of negative interaction, which is what is assumed in the current draft.
· Many of the supplementary sections are not referenced in the body of the paper. Please check that those are mentioned in appropriate places.

Specific Comments

Abstract

· Line 21: "AI reader" is ambiguous, especially given the metrics that are reported later. Please state the task the reader is evaluated against and the classes.
· Line 34: "Workload" is ambiguous
· Line 36: "AI triage performed less strongly" is vague. It's unclear if this is a comparison to the human reader baseline or the improvements from the other settings.

Intro

· Line 48: "The program has successfully contributed to a 41-52% reduction in mortality for screening participants and a 21% reduction in population-level breast cancer mortality in Australia". Please check

these numbers. The closest numbers I can find in the cited work to what you state are not population level (50-69 year olds) and are lower: “The results indicate that for 50–69-year-old women, participation in biennial screening through BreastScreen Australia is associated with an approximate 21–28% reduction in breast cancer mortality at the most recent national screening participation rate of 56%.”
· Line 51: No evidence of “increasing challenges” (emphasis added) is provided aside from suggesting screening compliance fell “due to the COVID 19 pandemic” without citation.
· Line 60: It’s not clear how “and there is little personalisation for individual risk” is related to this research.
· Line 76: What is the important contribution (Line 67) of [16]?
· Line 80: “Finally, some studies have explored AI-integration with human readers in screening pathways with simulation studies”. Please clarify your contribution in light of these studies.
· Line 82 and following: Please clarify that your study does not address the concerns about prospective analysis of interval cancers when introducing that concern about other trials.
· Line 91: Can you be explicit about why the guidance is “specifically targeted to implementation with population breast cancer screening programs in Australia”?
· Line 116: In what manner do you “substantially improve on previous efforts”? This statement would probably also be clearer if it followed the scenarios.
· Line 118: Please report the names of the settings when you introduce them.
· Line 130: Radiologist population has never been specified, but a specific value (91%) is given.
· Line 144: “particularly relevant for application in the Victorian and wider Australian population.” Why is it particularly relevant to these populations?
· Line 147: “An AI reader-replacement system can…” would be better in the context of your specific results.

Results

AI Reader

· Line 159: Weighted mean is introduced without specifying the weighting.
· Line 161: Please report both the AI reader and Human values. Currently it’s ambiguous if AI sensitivity is 75% or human sensitivity is 75%.
· Lines 163-167: It may be helpful to compare these numbers to when human readers have consensus positives and mixed positives, including using the percentage of cases that meet that criteria for thresholds.
· Line 167: A new task of distinguishing types of cancers is referenced but was never defined. Please define the task and classes.
· Lines 176-196:
·· Please quantify all these statements. For instance, please include patient populations and sensitive/specificity and confidence intervals for all these statements.
·· Given that this a summary of a supplemental figure, it may be worthwhile to include an abbreviated version of the figure in the main body with the comparisons you wish to highlight.

AI reader achieves state-of-the-art performance on validation datasets

· Lines 214&215: Which radiologist reader setting are you comparing to?
· Line 216: “Thus, we find no evidence of algorithmic drift or diminished performance of the BRAIx AI-reader”. The statistics quoted provide this relative to the radiologists, but not absolutely.

· Line 230: “AI reader show the relative strength of our model compared to previous state-of-the-art AI readers, as well as the quality and relative difficulty of our dataset.” The external datasets are small, so while the performance is good, this claim should be restricted to the datasets tested. Moreover, I do not believe any evidence of quality and relative difficulty of ADMANI has been provided outside of the BRAIx model performance.

AI Integrated Screening Scenarios
· Line 286: “With a 100% improvement in human performance thanks to AI-human interaction, the AI triage scenario can achieve sensitivity of 91.8% and specificity of 97.9%”.
·· I think this is supposed to provide an upper bound. If so, it would be clearer if it was labeled as such.
·· Reporting these numbers without reporting possible negative impacts has not been justified.

Discussion
· Line 296: No criteria have been established for measuring patient experience.
· Line 342: Suggesting the reader replacement is “conceptually straight-forward and would be the least disruptive” does not take into account that third readers are used to reading 2 human outputs. No discussion of this effect is provided in the paper.
· Line 387: The concluding statement is editorial and again references patient experience.

Figures and Tables
· Figure 1: What is the relationship between the data in B and the data in C? Is the “single point” the same as the values reported elsewhere?
· Figure 2: This figure is very helpful. Consider introducing it earlier to help ground the work.
· Table 3: Please add a total reads row for net effects. Caption mentions “workload and cost”, but no cost information is present.
· Table 4: Please include breakouts by machine vendor.

Methods
Code Availability
· Line 421: “model names and training procedures have been provided in detail in the methods”, but this is not true (also claimed on 424).
· Line 426: “The main conclusions drawn in our work relate to our AI reader simulation experiments”. This is not consistent with other sections, which dedicate considerable time to BRAIx and its performance.

Study Dataset/Figure 4
· Line 587: What kind of missing non-image data leads to exclusion?
· Figure 4: Each exclusion criteria in the first box should be defined.
· Figure 4: What is an “incomplete screening year”
· Figure 4: Please include the number of patients as well as the number of screenings
· Line 602: What does “or when recorded in the screening interval” mean in contrast to histopathology?

AI Reader System
· Line 620: “Optimized for… manufacture and population” provides little insight into the ensembling

process.

AI as a standalone Reader
· Line 633: The methodology for selecting the sensitivity and specificity operating points should be defined here.
· This section would be clearer if the general "compared with" was replaced with what specific comparisons were being made.

AI Integrated Reader Scenarios
· Line 658-660: The simulation of the third reader is not explained in detail, only that it achieves a target sensitivity and specificity. It is also unclear if the stated sensitivity and specificity are targets or actual values.

Evaluation metrics
· Please write out abbreviations when you first introduce them.
· Lines 702-709: AUC is a common metric. Defining it in is prose unnecessary.
· Line 709: "The reader summary ROC curve was fitted using a bivariate model using the individual radiologist sensitivities and specificities as provided by the mada R package". Please clarify. Is this referring to a specific figure?
· Line 716: "because early detection of cancer leads to more effective treatment." should be clarified and have a citation.
· Line 718: Please clearly define the negative class.
· Line 720: "center is costly and induces significant stress on clients and their families" should have a citation.
· Line 723: Please be explicit on the construction of pairings for McNemar's test and the reasoning for using continuity correction.

Supplements
· Supplementary Figure 1: No description in caption. Figures should be interpretable on their own.
· Supplementary Figure 2: These comparisons would be more robust as AUC.
· Supplementary Figure 3: This figure is very helpful. Consider adding false/true positives/negatives after the final step. Also consider including it in the body of the paper.
· Supplementary Figure 5: It's never made clear which outcomes in this flowchart map to the class prediction outcomes for the AI reader.
· Supplementary Table 4,5,6:
·· Needs some description of the economics methodology in the figure caption. For instance, where did the mammogram reading unit costs come from?
·· "Total Cost" should be relabeled "Processing Cost" or the like. The cost estimates do not include cost of licensing/hosting the software nor do they include cost estimates of harms to the patients.
· Supplementary Table 7: Please include negative values for "correction" of equivalent size for all positive values (or provide strong justification for suspecting improvement, see Summary comments).
· Supplementary Table 8: Please include vendors (for comparison with Table 4)
· Supplementary Figure 9: The caveats about reader selection should be in the main body of the paper.

Reviewer #2 (Remarks to the Author): Expert in artificial intelligence, machine learning, and computer vision;
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks to the Author): Expert in mammography, breast cancer imaging, and radiology;

We have reviewed this paper titled, “Integrated AI reader development and evaluation provides clinically relevant guidance for human- AI collaboration in population mammographic screening”, and have the following comments to offer:
This is a well-researched, planned and written paper with sound methodology that explores the potential of AI in providing image analysis assistance to the interpreting reader. The paper outlines current prevalent system of reading mammograms in the Australian continent and offers insights to new approaches, leveraging the power of AI to improve sensitivity and decrease the workload of the radiologist. These are important aspects that merit attention given wide variability in human reads and a shortage of trained mammographers worldwide.
Importantly, the paper states that a Standalone AI performs inferior to the current 2 reader system with arbitration. The paper therefore explores some novel human AI collaboration to improve outcomes.
We however have the following observations, which could be addressed to strengthen the quality of this paper:
Title:
To a reader, from the title it is not very clear whether the stress is on the development of the AI software (BRAIx) or suggested techniques for AI integration, human collaboration and implementation in the real-world setting. Rewording the title to address the research question clearly and concisely would be helpful.
Numerous publications have highlighted the sensitivity of AI in retrospective analysis, to that extent there is nothing novel in this paper about the development and evaluation of BRAIx in a retrospective cohort. We would also suggest that at places the authors refer to the value of this AI software in breast cancer risk assessment. Lesion analysis and breast cancer risk assessment from the mammographic images are 2 distinct and separate entities and not interchangeable.
It will be important for the authors to categorically state in the very beginning whether their study pertains to 2D images or 3D tomosynthesis images and if mixed in what proportion.
The authors describe 3 human AI collaboration methodologies, 2 of which the AI reader replacement and AI band-pass filtering have been documented in this study to show improved sensitivity and considerably reduced workload compared to the current 2 reader and arbitration system. These technologies hold promise and merit further exploration, notwithstanding many current limitations. The

societal challenges to the AI band-pass filtering technique, where many mammograms are not viewed by a human radiologist may not be acceptable in today's 'standard of medical care delivery'. The challenges posed by AI are nicely laid out in a paper by Stacy Carter et al; "The ethical, legal, and social implications of using artificial intelligence systems in breast cancer care". The Breast 49(2020)25-32. We would like to draw the author's attention to that publication and provide possible solutions to counter some of the arguments mentioned in the article.

Additionally, the 2-reader system is followed in the European and Australian continents. In USA, Canada, and many Asian countries it is a single reader. A mention regards global approaches to reading screening mammograms and any potential impact this research may have on the practice in these regions may make for interesting observations.

As a clinical breast imager, myself, I would be interested in seeing some true positive, false positive and false negative mammographic images in the article comparing human readers with AI interpreted images.

Introduction:

With reference to the introduction section some of our observations which we would like addressed are:

Line 42: Breast cancer is not the most cause of cancer related deaths in USA, it is lung cancer. Request authors modify that line.

Line 50: There is mention of little change over the last 30 years and this has served us well. We would prefer if the authors just state how the screening program is run in Australia. Discussing some of the pros and cons regarding national screening programs, at what age to start, what should be the frequency, when to stop, is there a need to personalize, based on risk etc is not the scope of this paper.

Line 60: The authors refer to personalization of individual risk. Though a reference is mentioned here, there needs to be some expansion into what is defined as high or low risk. Some definition needs to be provided to understand how the risk is being assessed, historical or risk models such as Gail, Tyrer-Cuzick, etc. Please also refer table 4.

Line 63: The authors refer to 3 key challenges of accuracy, client experience and efficiency but have not elaborated on them.

Line 66: They refer to limitations of current screening but have not expanded on them.

Line 95-98: Refers to six mammographic machines but does not specify vendors (was it Hologic, GE, Siemens, etc?) Refers to diversity of client ethnicity but the table only specifies country of origin. Breaking the population into Caucasian, natives, blacks, Asians etc and their proportional representation in the final mix would be useful. Please also refer table 4.

Lines 119-123: and elsewhere. The description of bandpass scenario is given without the details of the cutoff defined as "high confidence." What are the criteria used to define this cutoff both on the high and low ends that would lead to straight recall or no-recall. Please elaborate on the thresholds, is it the top and bottom 10% from the AI reader scores?

Lines 124-27: Refers to top deciles, where there lesion annotation and lesion scores?

Lines 163-65: The top 10% decile contained 92% cancers, 99% of screen detected cancers and 84% of interval cancers were covered in the top 50% decile, and that accounts for approx. 50% population (Table 1). Please elaborate on how that affects the volume of the workload.

Results:

In general AI does well in retrospective image analysis in lab settings. This has been established in numerous publications. The authors too have similar results which are not surprising or novel. We draw the authors attention to a paper by Karoline Freeman et al, "Use of artificial intelligence for image

analysis in breast cancer screening programs: systemic review of test accuracy". BMJ 2021;374: n1872 and could the authors explain how their study is different and address some of the concerns raised in the above referenced paper regards the published literature.
There is need for large volume randomized prospective trials in real-world settings to validate and explore AI implementation strategies. The authors reference 2 papers describing non-inferiority in prospective settings. The authors have tried at a small prospective cohort in this study and that is commendable, but this is confined to a single site with limited follow up.
The authors mention that the AI reader performed best on high grade invasive cancers. As a breast imager I would have been more interested in knowing the relative performance of AI reader based on image morphology. Broadly on mammograms, abnormal images fall into 4 broad categories vide the BIRADS Atlas namely masses, calcifications, asymmetries and architectural distortion. Breast imagers would be more interested in knowing the relative strengths and weaknesses of the AI reader in these categories. The paper makes no mention about that.
Line 195 and other places: The mention of ethnicity is only described for the First Nations Australians and other testing sets were examined according to region of birth. Since the datasets is so large and Australia has a diverse population, it would be useful to see the breakdown of patients based on race and ethnicity and evaluate the performance of the AI reader accordingly. This could further help in generalizing the AI reader to other places in the world and not just in Australia.

Discussion:
Line 304: First round screening clients had a higher proportion of breast cancer. Prevalent screening rates of cancer are always higher than incident screening rates, so that only confirms established metrics. Breast cancer rates if also mentioned per thousand rather than percentages would make it easier for international readers to follow.
Line 338: In the AI triage scenario there is mention that 14% improvement would equal or constitute improvement from the current system. Could the authors please explain how this number was derived.
Line 345-347: In discussing the 3 proposed scenarios for human-AI interaction, the band pass scenario while potentially being most beneficial is also likely to face maximum ethical and legal challenges and therefore additional explanation in discussion would help. The authors also conclude that the AI triage scenario is the least beneficial but clinically most relevant. This is important because that is precisely how AI implementation is being used in countries where mammograms are read by a single reader.
AI reader system
Line 617: The BRAIx AI reader provides a score for each breast and the maximum score is taken as the episode score. Does the AI reader also annotate specific regions in the breast and does it classify what type of cancer is being presented?

In conclusion the authors have attempted to tackle the relevant questions pertaining to a population screening program in the province of Victoria in the Australian continent, improving the cancer detection rate, decreasing false positives and decreasing the workload of overburdened staff leveraging the power of AI human collaboration. The study is however mostly retrospective and its impact in real-world clinical setting is currently indeterminate.
In addition to the human AI collaboration, the authors also refer to the development of their BRAIx reader. As mentioned earlier, it is unclear if the scope of this publication is to focus on the human AI aspect or to describe the development of the BRAIx software. Tackling both these topics takes away from

the focus of the paper. Moreover, there are currently other commercially available, FDA approved AI software that are clinically in use. I compliment the authors for comparing their software with others based on the ROC curves from published literature.

Reviewer #4 (Remarks to the Author): Expert in mammography, breast cancer imaging, and radiology;

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewers 1 & 2 (A)

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author): Expert in artificial intelligence, machine learning, and computer vision;

Overview
The authors introduce a new breast cancer detection model from screening mammograms, BRAIx, trained on a large cohort of patients (623k) from the ADMANI dataset built from participants in the BreastScreen Australia program from 2013-2019. BRAIx achieves impressive performance on a held-out test set of 93k patients (.932 AUC).

The authors then use BRAIx as an AI reader in multiple simulation settings. They replace one reader (reader replacement), selectively request human reading (band pass), and use the model to select 1 or 2 human readers (triage). All three simulations reduce the workload on human readers. The first two also improve performance of the screening protocol as measured by sensitivity and specificity. The authors suggest that these results mean that at least the reader replacement setting should be investigated in clinical trials.

We thank the two reviewers for their thoughtful engagement with our work and their joint review.

Review Summary
A01.

The BRAIx model shows impressive performance on the new ADMANI dataset and the (relatively small) external baselines used. However, because almost no detail is given about the implementation (to protect intellectual property, see Code Availability). As a result, there is limited scientific contribution from the first half of the paper.

We thank the reviewers for this comment (also raised by Reviewers #3 and #4) and agree and acknowledge the issue. Indeed, we had grappled with it internally in preparing the original manuscript; noting that the scientific contribution of the BRAIx AI reader and its implementation was limited due to the unavoidable necessity, in this project, to protect potentially commercially valuable project intellectual property as part of our multi-institution agreement and grant obligations. We acknowledge this issue and have comprehensively implemented the reviewers' suggestion and have now completely reworked and restructured the paper to solely focus on the simulation analysis, including extending the simulation studies and not to rely on the BRAIx reader itself as a scientific contribution. We also included the model names and training procedure in response to comment A40. As such, we have completely rewritten the paper and its new title, "Comparison of AI-integrated pathways with human-AI interactions for population mammographic screening," reflects the new focus on more detailed, deeper simulations exploring human-AI interaction effects in five different screening pathway systems. We have put a large amount of work into new simulation results that cover the concerns and vastly improve, in depth and breadth, on previous simulation studies in this area. Details and results pertaining to the BRAIx AI reader have been moved to Supplementary Material to aid understanding and contextualisation of the simulation results but are no longer presented as "Results" per se for

this manuscript.

A02.

The latter half of the paper focuses on simulated reader studies of AI reader integration to the mammogram reading protocol in Australia. The simulations use BRAIx as the AI reader. These approaches all seem plausible and show promising performance in the simulated studies. As such, they are of interest to the scientific community, though they are not clearly positioned against previous simulation studies.

We thank the reviewers for their feedback. Following their suggestion, we have now repositioned our paper, centring around expanded simulation studies. The second and their paragraphs of the new Introduction position our updated work against previous simulation studies.

A03.

The authors suggest the most likely path forward is the use of the AI-reader in conjunction with a single human reader. The results for this setting are promising. However, in the simulation, they rely on downstream performance of a human third reader holding constant after the introduction of an AI reader in the first round. The authors investigate the effects of AI performance on human readers in the triage setting, but not the reader replacement setting, and then only under the assumption that AI readers will improve human performance. It is unclear why we should assume this positive impact and why we should only consider AI-human interaction effects in the triage setting. Without investigating the potential impact of AI readers in both settings and as potentially negative for human readers, the import of this assumption throws into question the validity of the results (see "second reader as third reader" setting in Appendix 1 - Table 9 for an example where different assumptions lead to ambiguous impacts on performance).

Thank you for the insightful question. We took the implicit suggestions in this comment as motivation to drastically re-engineer our approaches to the simulations. We extended the simulations to include positive, neutral, and negative human-AI interaction effects in all the different scenarios that we study. As such, we are confident that our new, more comprehensive results resolve any questions about the validity of the results.

Further, more specific points:

- We hold the human third reader performance constant to respect the alignment of the simulations to the data / empirical performance, as there is no existing evidence to suggest AI would improve or impact the third reader performance under the replacement scenario. Definitive evidence regarding the performance of the third reader in AI integrated systems would require prospective studies, which we note are vital next steps but beyond the scope of this work.

- We consider the human-AI interaction for the triage setting because the interim result of the MASAI trial [22] has shown positive results, where the AI integrated system can match the original system. But under our analysis, the AI triage system tested in that trial can only achieve performance equivalent to the original system if there is positive interaction between human and AI. Hence, we conducted the original simulation only to the triage scenario because there is preliminary evidence that positive interaction exists. However, we have taken the opportunity presented by the reviewers' feedback to comprehensively extend our simulation studies. In the

revised manuscript, we explore positive, neutral, and negative human-AI interaction effects for all five AI-integrated pathways (including, of course AI triage).

- The ambiguous impacts of "second reader as third reader" ([18, 45]) was exactly why we advised against doing it - as explained in the caption of the table – because it overlooks the conditional dependence between the human readers and underestimates the sensitivity. This is a confusion of independence and conditional dependence. In simple terms, two readers are independent when they read two randomly sampled episodes respectively, but they are dependent (in the statistical sense) if they see the same episode. Because if one reader misdiagnoses the episode, it is more likely that the episode is a difficult case and the chance of the second reader also misdiagnosing is higher. The implication of this is that reader 2 and reader 3 are not suitable substitutes for each other in the simulation setting, because reader 2 tends to make the same mistakes as reader 1, while reader 3 does not (as it has access to more information and operates at different sensitivity and specificity).

We simulated the "second reader as third reader" to compare with the existing literature, but our position is that it is inappropriate. We hope this clarifies the issue.

- For completeness's sake, we have now included multiple interaction effects for all the scenarios, incorporating positive, neutral, and negative human-AI interaction effects.

A04.

While the paper provides valuable insights, it would benefit from rewriting that 1) clarifies the contributions and methods before introducing and interpreting results; 2) centers the AI-reader integration rather than BRAIx; and 3) considers potential positive and negative impact of AI readers on downstream human readers in all settings.

We thank the reviewers for this constructive feedback, and have adopted all three of these suggestions in our complete rewriting of the revised paper:

- The contributions have been clarified (the Introduction has been completely rewritten with this and other suggestions in mid), and methods are added before introducing results.

- The paper has been repositioned to centre the AI-reader integration with potential positive and negative (and neutral) impacts included in the simulations in all settings.

Strengths

· The size and diversity of the datasets tested (>400k patients).
· The performance of the BRAIx mammogram reader model is excellent. It generalizes to multiple (small) external datasets.
· The AI-human reader settings demonstrate viability, and are likely of interest for prospective trials, as the authors suggest.

Weaknesses

· Prospective study omits interval cancers because the follow-up data is not available and is limited to a single site.
· The authors suggest the reader replacement setting is the best candidate for prospective

trials, but it's unclear how much of the results in the AI-reader replacement setting hinge on the assumption that third reader performance would be held constant. The method for implementing human-reader performance on the simulated third reader is not described.
· Almost no detail on BRAIx implementation is provided. The BRAIx model relies on a new, large dataset for training (ADMANI). As a result, it's unclear if BRAIx's performance is a function of its architecture or training data, and no attempt is made to disentangle the effects.

We thank the reviewers for their characterisation of the strengths of our study, and we believe we have addressed all of the stated weaknesses in our revision, as described in detail in response to specific comments above and below. We believe the paper is much stronger now as a result of incorporating this feedback.

General Comments

A05
· AI cannot find a place where the positive and negative classes are explicitly defined. The task and evaluation are not clearly defined before results are introduced. It might be helpful to include an abbreviated Methods section with key points from the appendix after the Introduction.

We thank the reviewers for this helpful suggestion. An abbreviated Methods section has been added after the Introduction and at the beginning of the Results section. The positive and negative classes are now defined under the "Study design" subsection under the Results section.

A06
· Writing in the Intro and Discussion is difficult to follow. Subsections or breaking out parts (like creating a Related Works instead of embedding it into the Intro) would help.

Thanks for these suggestions. The Intro and Discussion have both been rewritten. The Related Works are now embedded to follow the journal style.

A07
· A bias towards assuming positive AI-reader human-reader interaction might be reasonable in light of [22], but that evidence is indirect and does not eliminate the possibility of negative interaction, which is what is assumed in the current draft.

We thank the reviewers for this insightful observation and we agree. We now have expanded our AI-reader simulations to include very detailed study of human-AI interaction effects and included positive, neutral and negative interactions to explore the full range of upsides and downsides to human-AI interaction in the different screening pathways.

A08
· Many of the supplementary sections are not referenced in the body of the paper. Please check that those are mentioned in appropriate places.

Thank you for catching these issues. We have checked that Supplementary Figures and Tables are referenced in appropriate places and those that are not referenced in the body are removed.

Specific Comments

Abstract

For the sake of conciseness, we have tried to be brief with responses to the specific comments. We are grateful to the reviewers for their thoughtful observations and hope that any brief responses do not come over as terse.

A09

· Line 21: "AI reader" is ambiguous, especially given the metrics that are reported later. Please state the task the reader is evaluated against and the classes.

We have added the context "detecting breast cancer in population screening programs" to clarify the task and the associated classes. Further detail is provided in the main text under the Study Design section.

A10

· Line 34: "Workload" is ambiguous

The abstract has been rewritten, and the word is no longer there.

A11

· Line 36: "AI triage performed less strongly" is vague. It's unclear if this is a comparison to the human reader baseline or the improvements from the other settings.

The abstract has been rewritten, and the phrase is no longer there.

Intro

A12

· Line 48: "The program has successfully contributed to a 41-52% reduction in mortality for screening participants and a 21% reduction in population-level breast cancer mortality in Australia". Please check these numbers. The closest numbers I can find in the cited work to what you state are not population level (50-69 year olds) and are lower: "The results indicate that for 50–69-year-old women, participation in biennial screening through BreastScreen Australia is associated with an approximate 21–28% reduction in breast cancer mortality at the most recent national screening participation rate of 56%."

Values were checked and the reference was updated.

A13

· Line 51: No evidence of "increasing challenges" (emphasis added) is provided aside from suggesting screening compliance fell "due to the COVID 19 pandemic" without citation.

Both "increasing" and "due to the COVID 19 pandemic" have been removed upon rewriting.

A14

· Line 60: It's not clear how "and there is little personalisation for individual risk" is related to this research.

The line has been removed during rewrite.

A15

· Line 76: What is the important contribution (Line 67) of [16]?

The contribution of the paper has been clarified during rewording of the Introduction in the context of simulation of AI-human interactions.

A16

· Line 80: "Finally, some studies have explored AI-integration with human readers in screening pathways with simulation studies". Please clarify your contribution in light of these studies.

A new paragraph has been added to clarify our contribution considering these studies.

A17

· Line 82 and following: Please clarify that your study does not address the concerns about prospective analysis of interval cancers when introducing that concern about other trials.

Following the reviewers' comments, the focus of the paper has been shifted to simulation of AI-human interactions. Mentions to interval cancers and AI have been removed from the Introduction.

A18

· Line 91: Can you be explicit about why the guidance is "specifically targeted to implementation with population breast cancer screening programs in Australia"?

Following the reviewers' comments, the focus of the paper has been shifted to simulation of AI-human interactions. The methodology behind the simulations can be employed independently of the model used. The phrase "specifically targeted to implementation with population breast cancer screening programs in Australia" has been removed from the Introduction as, while the simulations are based on our Australian dataset, the simulations results are appropriately general to be relevant to all screening settings, in Australia and beyond. We thank the reviewers for the suggestions that, taken together, have allowed us to present simulation results that, we believe, are relevant well beyond the specific screening programs in Australia (as we had originally intimated).

A19

· Line 116: In what manner do you "substantially improve on previous efforts"? This statement would probably also be clearer if it followed the scenarios.

In revision, this statement has been expanded and moved after the description of the scenarios, as suggested.

A20

· Line 118: Please report the names of the settings when you introduce them.

Our simulation pathway scenarios are now named when introduced (with emphasis). Simulation scenarios from the literature are described but not specifically named to group similar approaches for conciseness and clarity.

A21

· Line 130: Radiologist population has never been specified, but a specific value (91%) is given.

The radiologist population is specified in the Results section. The specific value has been removed from the Introduction to avoid any confusion arising from the omission of details in the Introduction (for space considerations).

A22

· Line 144: “particularly relevant for application in the Victorian and wider Australian population.” Why is it particularly relevant to these populations?

Although the AI model presented state-of-the-art performance on international datasets, it was trained using an Australian mammogram dataset, which makes it perform distinctly better in the Australian population. The text has been rephrased to clarify that relevance is related to the model performance.

A23

· Line 147: “An AI reader-replacement system can…” would be better in the context of your specific results.

The paragraph has been rephrased to present specific results from our research instead of general AI features.

Results

AI Reader

A24

· Line 159: Weighted mean is introduced without specifying the weighting.

The weighting is now specified when introduced: “… the mean of individual radiologists (weighted by number of reads) …”

A25

· Line 161: Please report both the AI reader and Human values. Currently it’s ambiguous if AI sensitivity is 75% or human sensitivity is 75%.

In the revisions these values are no longer the same line but both values are now reported when making the comparison, avoiding ambiguity.

A26

· Lines 163-167: It may be helpful to compare these numbers to when human readers have consensus positives and mixed positives, including using the percentage of cases that meet that criteria for thresholds.

Agreed, looking at consensus and mixed positives would be interesting, but as the table has been moved to Supplementary Material in the restructure, we opted not to expand on it further.

A27

· Line 167: A new task of distinguishing types of cancers is referenced but was never defined. Please define the task and classes.

This line has been removed due to restructuring. Task and classes have been better defined in the Study Design section of the results.

A28

· Lines 176-196:

· · Please quantify all these statements. For instance, please include patient populations and sensitive/specificity and confidence intervals for all these statements.

· · Given that this a summary of a supplemental figure, it may be worthwhile to include an abbreviated version of the figure in the main body with the comparisons you wish to highlight.

These lines were removed in restructure to focus the paper on simulations.

AI reader achieves state-of-the-art performance on validation datasets

A29

· Lines 214&215: Which radiologist reader setting are you comparing to?

This line has been removed due to restructuring.

A30

· Line 216: "Thus, we find no evidence of algorithmic drift or diminished performance of the BRAIx AI-reader". The statistics quoted provide this relative to the radiologists, but not absolutely.

This line has been removed due to restructuring.

A31

· Line 230: "AI reader show the relative strength of our model compared to previous state-of-the-art AI readers, as well as the quality and relative difficulty of our dataset." The external datasets are small, so while the performance is good, this claim should be restricted to the datasets tested. Moreover, I do not believe any evidence of quality and relative difficulty of ADMANI has been provided outside of the BRAIx model performance.

Thanks for the comment. We have restricted this claim to the external datasets on which the BRAIx AI reader was evaluated on. The text "…as well as the quality and relative difficulty of our dataset." has been removed.

AI Integrated Screening Scenarios

A32

· Line 286: "With a 100% improvement in human performance thanks to AI-human interaction, the AI triage scenario can achieve sensitivity of 91.8% and specificity of 97.9%".
· · I think this is supposed to provide an upper bound. If so, it would be clearer if it was labeled as such.
· · Reporting these numbers without reporting possible negative impacts has not been justified.

We have now reported also the possible negative impacts and clarified the statements of these results.

Discussion

A33

· Line 296: No criteria have been established for measuring patient experience.
This line has been removed due to restructuring.

A34

· Line 342: Suggesting the reader replacement is "conceptually straight-forward and would be the least disruptive" does not take into account that third readers are used to reading 2 human outputs. No discussion of this effect is provided in the paper.

The disruption referenced is related to the fact that only the third reader would view the AI reader outputs in the reader-replacement scenario and not all readers as in the other scenarios (which is more disruptive in our view). The text has been updated to link the statement to disruption of reader structure, cancer prevalence, and the number of readers who can view AI reader outputs.

A35

· Line 387: The concluding statement is editorial and again references patient experience.
The discussion has been rewritten, and there is no reference to patient experience anymore.

Figures and Tables

A36

· Figure 1: What is the relationship between the data in B and the data in C? Is the "single point" the same as the values reported elsewhere?

(B) shows where the AI reader could outperform human reader and (C) where the AI reader could not outperform the human reader – the caption has been updated to make this clearer. The single point is not the same as other operating points, the figure is effectively an ROC comparison for each individual reader but using a single point for visibility and the panels to aid interpretability. In the caption we changed "a single point" to "an optimal point" for clarity. For now, Figure 1 is now Figure 2 as per A37.

A37

· Figure 2: This figure is very helpful. Consider introducing it earlier to help ground the work.

It has been moved to an earlier section, and it is now Figure 1.

A38

· Table 3: Please add a total reads row for net effects. Caption mentions "workload and cost", but no cost information is present.

We have updated the wording, so now the "human reads" are the total reads. The word "cost" has been removed.

A39

· Table 4: Please include breakouts by machine vendor.

Table has been updated.

Methods

Code Availability

A40

· Line 421: "model names and training procedures have been provided in detail in the methods", but this is not true (also claimed on 424).
Apologies for this oversight in the original manuscript. We have updated the AI Reader System section to include the model names and training procedures.

A41

· Line 426: "The main conclusions drawn in our work relate to our AI reader simulation

experiments". This is not consistent with other sections, which dedicate considerable time to BRAIx and its performance.
We have restructured the paper to focus on the simulation.

Study Dataset/Figure 4

A42
· Line 587: What kind of missing non-image data leads to exclusion?

Missing reader, screening or assessment records leads to exclusion. This detail has been added to the caption.

A43
· Figure 4: Each exclusion criteria in the first box should be defined.

Added definitions to caption.

A44
· Figure 4: What is an "incomplete screening year"

This refers to years that were cancer enriched and not full year samples (2013-2015). Added text to the Study Datasets section.

A45
· Figure 4: Please include the number of patients as well as the number of screenings

Added final breakdown of clients; exclusions were made by episode.

A46
· Line 602: What does "or when recorded in the screening interval" mean in contrast to histopathology?

Thank you for the comment, this was not clear. Line has been updated to reference sourcing interval cancer data from cancer registries and screen-detected cancer from histopathology.

AI Reader System

A47
· Line 620: "Optimized for… manufacture and population" provides little insight into the ensembling process.

This line has been removed due to restructuring.

AI as a standalone Reader

A48
· Line 633: The methodology for selecting the sensitivity and specificity operating points should be defined here.

The section has been merged into the "AI integrated screening scenarios", and the operating points are defined under the "AI operating points" section.

A49
· This section would be clearer if the general "compared with" was replaced with what specific comparisons were being made.

The section has been merged into the "AI integrated screening scenarios", and the operating points are defined under the "AI operating points" section.

AI Integrated Reader Scenarios

A50
· Line 658-660: The simulation of the third reader is not explained in detail, only that it achieves a target sensitivity and specificity. It is also unclear if the stated sensitivity and specificity are targets or actual values.
The section has been expanded to include more detail how the simulation is performed. The stated sensitivity and sensitivity are both targets and actual values. The stated sensitivity and specificity are actual values, and the simulation are performed to match that, now clarified in the text.

Evaluation metrics

A51
· Please write out abbreviations when you first introduce them.

The edit has been made, except for AUC / ROC as commented in A52.

A52
· Lines 702-709: AUC is a common metric. Defining it in is prose unnecessary.
AUC definition has been removed from the text.

A53
· Line 709: "The reader summary ROC curve was fitted using a bivariate model using the individual radiologist sensitivities and specificities as provided by the mada R package". Please clarify. Is this referring to a specific figure?

This line is now removed. It referred to a figure that was deprecated and was not included in the *submis*sion.

A54
· Line 716: "because early detection of cancer leads to more effective treatment." should be clarified and have a citation.
Improvement in effectiveness of the treatment due to early cancer detection has been clarified and a citation has been added to the text.

A55
· Line 718: Please clearly define the negative class.

The negative class has been defined in the earlier sections "Methods - Screening program" and "Results - Study design".

A56
· Line 720: "center is costly and induces significant stress on clients and their families" should have a citation.
Citation has been added to the text.

A57
· Line 723: Please be explicit on the construction of pairings for McNemar's test and the reasoning for using continuity correction.

The samples are paired by episode ID and then compared based on the episode outcome and prediction. Continuity correction is used to account for the difference between the statistical test developed based on continuous distribution (the chi-squared distribution) and the actual data, which follow a discrete distribution (the binomial distribution). These edits have been made to Line 723.

Supplements

A58
· Supplementary Figure 1: No description in caption. Figures should be interpretable on their own.

Description has been added to the caption of the figure.

A59
· Supplementary Figure 2: These comparisons would be more robust as AUC.

In our view, framing the comparisons relative to the reader and standard of care add more context than only the classifier with AUC, especially as some of the comparisons (round, age) have difference cancer prevalence, which confounds results.

A60
· Supplementary Figure 3: This figure is very helpful. Consider adding false/true positives/negatives after the final step. Also consider including it in the body of the paper.

The TP/FP/TN/FN numbers have been added to the flowchart. Reference to the flowcharts has been added to the body.

A61
· Supplementary Figure 5: It's never made clear which outcomes in this flowchart map to the class prediction outcomes for the AI reader.

Thanks for the feedback – we have added positive and negative class to the diagram (which aligns with the mini methods as suggested in other feedback).

A62
· Supplementary Table 4,5,6:
· · Needs some description of the economics methodology in the figure caption. For instance, where did the mammogram reading unit costs come from?
· · "Total Cost" should be relabeled "Processing Cost" or the like. The cost estimates do not include cost of licensing/hosting the software nor do they include cost estimates of harms to the patients.

The Supplementary Table 4,5,6 have been removed as they were not referenced in the text (as requested by comment A08) and has been superseded by the priorities of the restructure.

A63

· Supplementary Table 7: Please include negative values for "correction" of equivalent size for all positive values (or provide strong justification for suspecting improvement, see Summary comments).

The Supplementary Table 7 has been removed as it was not referenced in the text (as requested by comment A08).

A64

· Supplementary Table 8: Please include vendors (for comparison with Table 4)

Table has been updated to include vendor comparison.

A65

· Supplementary Figure 9: The caveats about reader selection should be in the main body of the paper.

It is now included under the "Simulation design – AI integrated screening scenarios".

Reviewers 3 & 4 (B)

Reviewer #3 (Remarks to the Author): Expert in mammography, breast cancer imaging, and radiology;

We have reviewed this paper titled, "Integrated AI reader development and evaluation provides clinically relevant guidance for human- AI collaboration in population mammographic screening", and have the following comments to offer:

This is a well-researched, planned and written paper with sound methodology that explores the potential of AI in providing image analysis assistance to the interpreting reader. The paper outlines current prevalent system of reading mammograms in the Australian continent and offers insights to new approaches, leveraging the power of AI to improve sensitivity and decrease the workload of the radiologist. These are important aspects that merit attention given wide variability in human reads and a shortage of trained mammographers worldwide. Importantly, the paper states that a Standalone AI performs inferior to the current 2 reader system with arbitration. The paper therefore explores some novel human AI collaboration to improve outcomes.
We however have the following observations, which could be addressed to strengthen the quality of this paper:

We thank Reviewers #3 and #4 for their thoughtful engagement with our work and recognition of the merits of our paper. We have further strengthened the paper in the major revisions we have completed following the reviewers' suggestions.

Title

B01
To a reader, from the title it is not very clear whether the stress is on the development of the AI software (BRAIx) or suggested techniques for AI integration, human collaboration and implementation in the real-world setting. Rewording the title to address the research question clearly and concisely would be helpful.

We thank the reviewers for this astute observation, and we fully agree. We have updated the title to "Comparison of AI-integrated pathways with human-AI interactions for population mammographic screening".

General

B02
Numerous publications have highlighted the sensitivity of AI in retrospective analysis, to that extent there is nothing novel in this paper about the development and evaluation of BRAIx in a retrospective cohort. We would also suggest that at places the authors refer to the value of this AI software in breast cancer risk assessment. Lesion analysis and breast cancer risk assessment from the mammographic images are 2 distinct and separate entities and not interchangeable.

We thank the reviewers for this comment and acknowledge the consensus advice of all four reviewers to reposition and rewrite the paper to focus on our novel simulation studies rather than the development of another AI reader. As such, we have repositioned the paper around comprehensive comparison of AI-integrated pathways with human-AI interaction and completely rewritten the paper in this revision.

B03
It will be important for the authors to categorically state in the very beginning whether their study pertains to 2D images or 3D tomosynthesis images and if mixed in what proportion.

Thanks for catching this issue. The updated "Study Design" section now clarifies such questions (2D X-ray images are used here).

B04
The authors describe 3 human AI collaboration methodologies, 2 of which the AI reader replacement and AI band-pass filtering have been documented in this study to show improved sensitivity and considerably reduced workload compared to the current 2 reader and arbitration system. These technologies hold promise and merit further exploration, notwithstanding many current limitations. The societal challenges to the AI band-pass filtering technique, where many mammograms are not viewed by a human radiologist may not be acceptable in today's 'standard of medical care delivery'. The challenges posed by AI are nicely laid out in a paper by Stacy Carter et al; "The ethical, legal, and social implications of using artificial intelligence systems in breast cancer care". The Breast 49(2020)25-32. We would like to draw the author's attention to that publication and provide possible solutions to counter some of the arguments mentioned in the article.

We thank the reviewers for their attention to the Carter et al paper. The lead author of this manuscript (H Frazer) is an author on the Carter et al paper, so it is indeed well known to us, and

we have addressed our oversight by explicitly citing it in our revised paper. We take on board the reviewers' comments here in the totality of our substantial revisions to the manuscript. Specifically, we have completely rewritten the Introduction and paragraphs 2 and 3 more explicitly position our work against the foundation of prior work in this field, including the many important points raised in the Carter et al paper.

B05
Additionally, the 2-reader system is followed in the European and Australian continents. In USA, Canada, and many Asian countries it is a single reader. A mention regards global approaches to reading screening mammograms and any potential impact this research may have on the practice in these regions may make for interesting observations.

Thanks for this important suggestion - a mention has been added to the Discussion.

B06
As a clinical breast imager, myself, I would be interested in seeing some true positive, false positive and false negative mammographic images in the article comparing human readers with AI interpreted images.

Thanks for this suggestion - Images have been added to the Supplementary Figures.

For the sake of conciseness, we have tried to be brief with responses to the more specific comments below. We are grateful to the reviewers for their thoughtful observations and hope that any brief responses do not come over as terse.

Introduction

With reference to the introduction section some of our observations which we would like addressed are:

B07
Line 42: Breast cancer is not the most cause of cancer related deaths in USA, it is lung cancer. Request authors modify that line.
Text has been modified appropriately.

B08
Line 50: There is mention of little change over the last 30 years and this has served us well. We would prefer if the authors just state how the screening program is run in Australia. Discussing some of the pros and cons regarding national screening programs, at what age to start, what should be the frequency, when to stop, is there a need to personalize, based on risk etc is not the scope of this paper.

The introduction has been rewritten, and the line is not there anymore.

B09
Line 60: The authors refer to personalization of individual risk. Though a reference is mentioned here, there needs to be some expansion into what is defined as high or low risk. Some definition needs to be provided to understand how the risk is being assessed, historical or risk models such as Gail, Tyrer-Cuzick, etc. Please also refer table 4.

The introduction has been rewritten, and the line is not there anymore.

B10
Line 63: The authors refer to 3 key challenges of accuracy, client experience and efficiency but have not elaborated on them.

A new reference has been added, but client experience is overall de-emphasised in our revised manuscript.

B11
Line 66: They refer to limitations of current screening but have not expanded on them.

Please refer to reference [12] for the details of the limitations and how they are addressed.

B12
Line 95-98: Refers to six mammographic machines but does not specify vendors (was it Hologic, GE, Siemens, etc?)

We thank the reviewers for their interest in the specifics of the mammographic machine vendors, but we have chosen not to specify vendors. The focus of the comparison is to show consistency across multiple manufacturers, not specific performance on specific manufacturers. Given that focus, we opt not to disclose the details of the vendors to avoid getting drawn into any debates about which vendor may/may not be better/worse than another.

B13

Refers to diversity of client ethnicity but the table only specifies country of origin. Breaking the population into Caucasian, natives, blacks, Asians etc and their proportional representation in the final mix would be useful. Please also refer table 4.

We appreciate the reviewers' concern for the ways in which AI model performance may differ between people of different genetic ancestry or ethnicity. However, the BreastScreen dataset only contains information on country of origin/birth and does not record self-reported ethnicity information. As such, we feel it would be entirely inappropriate to "break the population into Caucasian, natives, blacks, Asians etc". These labels are not self-reported by BSV clients, and the suggested labels are inappropriate for usage in the Australian context. In the updated manuscript, Table 2 provides a "country of origin" summary of dataset composition at what we believe is an appropriate level of granularity given the established reporting practices in Australia and the client metadata available in the ADMANI datasets. We note that reporting practices and labels for ancestry and ethnic groups are different in Australia than in other countries, particularly the USA, so labels or groupings that may be appropriate and useful in other countries can not necessarily be applied to data from the Australian population. We have taken guidance from the National Academies of Sciences, Engineering, and Medicine's recent report on using population descriptors in research (link following), which among much sound advice cautions against using racialized terms like "Caucasian" (US census bureau uses "white" now): https://nap.nationalacademies.org/catalog/26902/using-population-descriptors-in-genetics-and-genomics-research-a-new

B14
Lines 119-123: and elsewhere. The description of bandpass scenario is given without the details of the cutoff defined as "high confidence." What are the criteria used to define this cutoff both

on the high and low ends that would lead to straight recall or no-recall. Please elaborate on the thresholds, is it the top and bottom 10% from the AI reader scores?

Due to space constraints, the details are provided in the "Methods - AI operating points" section.

Briefly, the high and low thresholds are set by running a two-dimensional grid search on the development set for the best combinations to improve the scenario sensitivity and specificity.

The thresholds correspond to approximately the bottom 79% and top 3% from the AI reader scores, and they can be confirmed in Supplementary Fig. 7C by dividing the number of episodes that went to the low-/high-band by the total number of episodes. The asymmetry stems from the fact that the majority of the screening population is at low risk of cancer.

B15

Lines 124-27: Refers to top deciles, where there lesion annotation and lesion scores?
The text has been removed after rewriting and restructuring. The AI reader does produce annotations (lesions are not scored); further detail on the reader is in the methods and an example annotation provided in the supplementary.

Results

B16

Lines 163-65: The top 10% decile contained 92% cancers, 99% of screen detected cancers and 84% of interval cancers were covered in the top 50% decile, and that accounts for approx. 50% population (Table 1). Please elaborate on how that affects the volume of the workload.

Thanks for the feedback: this line has been altered slightly and hopefully clarified in the revisions (to only report the top decile). It was primarily intended as a measure for the performance of the AI reader and for past and future comparison with other studies which have reported deciles in this manner. There was no intended comment on workload.

B17

In general AI does well in retrospective image analysis in lab settings. This has been established in numerous publications. The authors too have similar results which are not surprising or novel. We draw the authors attention to a paper by Karoline Freeman et al, "Use of artificial intelligence for image analysis in breast cancer screening programs: systemic review of test accuracy". BMJ 2021;374: n1872 and could the authors explain how their study is different and address some of the concerns raised in the above referenced paper regards the published literature.

We thank the reviewers for this important feedback addressing the overall understanding and positioning of our work in the field. We agree that the Freeman et al paper is an important resource and had cited it in the original submission. In our revisions we have expanded the Introduction to better position our study and its contribution beyond previous related work in the literature. An issue with many studies in the review (11/12 of studies reviewed had high or unclear risk of bias in reference standard) was either using small, enriched datasets or lack of interval follow up, which we address by using a large, representative test cohort with biopsy-confirmed screen-detected cancers and a full two-year interval follow-up. None of the studies were using an Australian dataset, and three of the population cohorts were overlapping Swedish studies. We also show our AI reader can outperform a single reader (many reviewed couldn't),

and we also show our AI reader does not alter disease spectrum in a negative way (performs better than reader consensus on high grade invasive and relatively poorly on low grade DCIS). Most importantly, however, we have completely restructured our paper to focus on the extended simulation studies, so the information about the BRAIx AI reader itself has been relegated to Supplementary Material in this revised version.

B18

There is need for large volume randomized prospective trials in real-world settings to validate and explore AI implementation strategies. The authors reference 2 papers describing non-inferiority in prospective settings. The authors have tried at a small prospective cohort in this study and that is commendable, but this is confined to a single site with limited follow up.

Thanks for the comment, and we fully agree that more prospective trials are needed and are planning to undertake a randomised controlled trial that includes multiple sites across multiple states. Our revised Discussion also emphasises the need for prospective trials as the next steps arising from this work.

B19

The authors mention that the AI reader performed best on high grade invasive cancers. As a breast imager I would have been more interested in knowing the relative performance of AI reader based on image morphology. Broadly on mammograms, abnormal images fall into 4 broad categories vide the BIRADS Atlas namely masses, calcifications, asymmetries and architectural distortion. Breast imagers would be more interested in knowing the relative strengths and weaknesses of the AI reader in these categories. The paper makes no mention about that.

Thanks for the comment and suggestions. We have added a comparison on performance of the AI reader broken down by morphology for screen-detected cancers with consistent labels. As we restructured the paper to be more simulation focused it has been added to the supplementary material.

B20

Line 195 and other places: The mention of ethnicity is only described for the First Nations Australians and other testing sets were examined according to region of birth. Since the datasets is so large and Australia has a diverse population, it would be useful to see the breakdown of patients based on race and ethnicity and evaluate the performance of the AI reader accordingly. This could further help in generalizing the AI reader to other places in the world and not just in Australia.

Thanks for the comment and consideration of the importance of ensuring that benefits of AI accrue equitbly to all people. However, as described in detail above (B13), no ethnicity data is collected by the Australian BreastScreen screening services other than for First Nations Australians. Region of birth (derived from country of birth) was the closest data element available to provide any information for the wider cohort, and thus we are limited in the detail in this vein that we can accurately and sensibly report.

Discussion

B21

Line 304: First round screening clients had a higher proportion of breast cancer. Prevalent

screening rates of cancer are always higher than incident screening rates, so that only confirms established metrics. Breast cancer rates if also mentioned per thousand rather than percentages would make it easier for international readers to follow.

This line has been removed in the rewrite; cancer rates per thousand are referenced in the introduction.

B22
Line 338: In the AI triage scenario there is mention that 14% improvement would equal or constitute improvement from the current system. Could the authors please explain how this number was derived.

Yes, indeed. We ran the simulation under different level of positive interaction (from 0% to 100%) and find the level at which the AI triage scenario has both sensitivity and specificity at least as good as the current system. That level is what we reported.

B23
Line 345-347: In discussing the 3 proposed scenarios for human-AI interaction, the band pass scenario while potentially being most beneficial is also likely to face maximum ethical and legal challenges and therefore additional explanation in discussion would help. The authors also conclude that the AI triage scenario is the least beneficial but clinically most relevant. This is important because that is precisely how AI implementation is being used in countries where mammograms are read by a single reader.

We thank the reviewers for this comment. We have added further comment and referenced the Carter et al study regarding the AI band-pass scenario drawbacks. Our expanded simulations, looking further into human-AI interaction, reference now that AI triage has the largest upside from positive human-AI interaction. We have also included a single-reader variant and show that some automation-bias could have a positive effect for single readers, extending our original simulations to offer useful insight for countries where mammograms are read by a single reader.

AI reader system

B24
Line 617: The BRAIx AI reader provides a score for each breast and the maximum score is taken as the episode score. Does the AI reader also annotate specific regions in the breast and does it classify what type of cancer is being presented?

Thanks for the comment, region annotation is provided from the model and now an example is now provided in the Supplementary Material. No cancer type classification is available.

B25
In conclusion the authors have attempted to tackle the relevant questions pertaining to a population screening program in the province of Victoria in the Australian continent, improving the cancer detection rate, decreasing false positives and decreasing the workload of overburdened staff leveraging the power of AI human collaboration. The study is however mostly retrospective and its impact in real-world clinical setting is currently indeterminate.

We thank the reviewers for the comment. We agree that further prospective trials are required. We aim to evaluate the BRAIx AI reader in a randomised controlled trial commencing later in

2024 and in the revised Discussion we emphasise the importance of prospective trials as the crucial next step to follow up on work like that presented in this paper.

B26

In addition to the human AI collaboration, the authors also refer to the development of their BRAIx reader. As mentioned earlier, it is unclear if the scope of this publication is to focus on the human AI aspect or to describe the development of the BRAIx software. Tackling both these topics takes away from the focus of the paper. Moreover, there are currently other commercially available, FDA approved AI software that are clinically in use. I compliment the authors for comparing their software with others based on the ROC curves from published literature.

Thanks for the feedback - the publication has been restructured to focus on the simulation of different screening pathways that use AI Readers to assist or replace human readers. Description of the development of the BRAIx software has been abridged or removed as required, the ROC curves and results from public benchmarks have been retained to enable future comparison.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

Thank you to the authors for engaging carefully with the reviews from the first round. This revision addresses all of the major concerns raised in the first round of reviews.

First, the authors refocused the paper on the reader simulation studies. The abstract and introduction both clearly articulate the contribution of the simulation studies, the methods section provides a precise and detailed discussion of the settings, making the results of the studies easier to understand. Figure 1 is especially helpful.

Second, the authors provided appropriate detail about the BRAIx AI Algorithm. This was addressed in two ways: the paper was refocused on the simulation study so that BRAIx itself is not highlighted as a key contribution of this study (since it is protected intellectual property). Then, since BRAIx is used in the simulation studies, sufficient detail about the implementation is provided in the appendices for this purpose.

Finally, the inclusion of new simulations with positive and negative AI-human interactions addresses the key limitation of the first revision. I think the authors choice to present the results without the effects first then add these effects helped make the results easier to digest. The methodology the authors chose (modeling changes in human behavior as a percent of changes on cases with disagreement) is a sensible modeling choice and makes the results easy to understand. The readers acknowledge that their method is limited to choosing a specific operating point for the AI Reader in the limitations. Most importantly, these results address the key concern we had upon first reading: did the improvements in the reader-replacement scenario depend on assumptions about human-AI interactions? The readers give a complete treatment to this question for all of their settings, which yields new insights about how the different settings have different robustness to AI-human interaction effects. Figure 3 is excellent. This treatment is fair and strengthens the author's original conclusion that the reader replacement setting has clinical potential.

Together, these result in a first-of-its-kind simulation study of possible paths to integrate AI into mammography reading accounting for potential human-AI interaction.

Other Notes
The missing data in the tables has been addressed.
The authors do an excellent job in this revision in describing how the differences in performance of the human and AI readers propagate through the settings.
There are dead references throughout the paper resulting in broken links.
The Figure 3 caption would be clearer if it explicitly stated A did not include interaction effects. Same with Appendix 1—figures 6 and 7. A version of Figure 3 B with each setting split out would be a nice supplement.
The authors addressed confusion on the class labels, including in Appendix 1 figure 8.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks to the Author):

Thank you for the revisions.
Most of the comments raised by us earlier have been addressed by the authors.
The current title is more reflective of the scope of this paper. It stresses the human -AI collaborative pathways.
While the role of AI in image interpretation has seen a significant spurt in recent years, its value prospectively is still untested.
The authors in this manuscript based on retrospective data explore potential human AI collaboration in the future and have suggested some pathways, which may impact the way we interpret images in the future.
In the discussion the authors also stress on the “junior partner role” for the AI. While in retrospective cohort studies the AI bandpass filter shows promise, in my view in the current practice environment, that may be a “disruptive approach”, with implications for insurance and radiology charges.
AI reader replacement (in the 1st or 2nd reader positions)and triage options both have the potential to reduce workload and are easier to implement in the current medical environment.

A few comments for the authors:
1. In the result section and elsewhere they refer to supplementary figures but have not numbered them and instead there have placed ??. Request, please correct the same.
2. Also, I was a little surprised that the stand-alone AI had better specificity than the human reader. In my personal experience, the human reader has access to prior images, medical reports and records and is therefore able to better evaluate prior biopsy/surgical sites or stable findings on a mammogram and to call them benign as opposed to an AI reader who does not benefit from these features. In my personal experience, I have found most AI softwares without recourse to prior images and reports actually call more false positives. Their increased sensitivity compared to human readers is not surprising.
3. In the methods section, if the authors could please add a line on the randomization process for the 20% cases that were selected for this study and simulation.

Reviewer #4 (Remarks to the Author):

"I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of

the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts."

Reviewer Feedback

We thank the reviewers overall for their constructive and thoughtful feedback throughout the review process.

Reviewer #1 (Remarks to the Author):

A1
There are dead references throughout the paper resulting in broken links.

Apologies, and thanks for catching this issue – the references have been corrected.

A2
The Figure 3 caption would be clearer if it explicitly stated A did not include interaction effects. Same with Appendix 1—figures 6 and 7.

We have added text to state without interaction effects to each caption.

A3

A version of Figure 3 B with each setting split out would be a nice supplement.

Thanks for this suggestion – we have added a figure of each setting separately in the supplement.

Reviewer #3 (Remarks to the Author):

B1
1. In the result section and elsewhere they refer to supplementary figures but have not numbered them and instead there have placed ??. Request, please correct the same.

Apologies, and thanks for catching this issue – the references have been corrected.

B2
2. Also, I was a little surprised that the stand-alone AI had better specificity than the human reader. In my personal experience, the human reader has access to prior images, medical reports and records and is therefore able to better evaluate prior biopsy/surgical sites or stable findings on a mammogram and to call them benign as opposed to an AI reader who does not benefit from these features. In my personal experience, I have found most AI softwares without recourse to prior images and reports actually call more false positives. Their increased sensitivity compared to human readers is not surprising.

We thank you for the comment. That may be related to the specific operating point of the AI software; we have noted from the literature that manufacturers often use a top 10% setting that if we had used, would have had flagged more false positives. We agree that effectively using prior images should improve AI readers in future and Supplementary Figure 3 (first round vs. second and subsequent) illustrates this current gap in performance.

B3
3. In the methods section, if the authors could please add a line on the randomization process for the 20% cases that were selected for this study and simulation.

We have added a line to further clarify the process.