PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

Title (Provisional)

AI Assisted Detection for Chest X-rays (AID-CXR): a Multi-Reader Multi-Case Study Protocol

Authors

Khan, Farhaan; Das, Indrajeet; Kotnik, Marusa; Wing, Louise; Van Beek, Edwin; Murchison, John; Ahn, Jong Seok; Lee, Sang Hyup; Seth, Ambika; Espinosa Morgado, Abdala Trinidad; Fu, Howell; Novak, Alex; Salik, Nabeeha; Campbell, Alan; Shah, Ruchir; Gleeson, Fergus; Ather, Sarim

VERSION 1 - REVIEW

Reviewer	**1**
Name	**Seah, Jarrel**
Affiliation	**Annalise-AI Pty Ltd**
Date	**05-Nov-2023**
COI	**Employee, Harrison.ai**

Ground truthing- the manuscript describes a two plus one method of consensus ground truthing. However as multiple findings are being assessed - will this be performed on a per finding basis or will any discrepancy in any of the findings lead to all findings being adjudicated by the third truther? This is particularly pertinent as some of the findings eg mass vs consolidation vs atelectasis can be confused for one another and a ground truthing process that is performed on a per finding basis may lead to inconsistent ground truths e.g describing a single abnormality as both a mass and consolidation simultaneously.

Apart from this the protocol is well described and sound.

Reviewer	**2**
Name	**Khan, FA**
Affiliation	**McGill University**
Date	**04-Jan-2024**

COI **None.**

**please see word document for a version of this review with proper spacing.

Thanks for the opportunity to review this protocol. The approach is sound with respect to comparing Lunit to the ground truth readers, and will provide important insight into the accuracy of Lunit.

I have the following questions:

- Can the authors please provide more information on the selection process for the images? I am specifically interested to know if the 40 chosen for each type of abnormality are a consecutive or random selection, and if not, what steps might be taken to ensure the cases represent the full spectrum of severity to avoid spectrum bias?

- How will anatomic location of an abnormality be taken into account when comparing the 'ground truth' reading to LUNIT or the other human readers? For example- how to ensure that if the ground truth readers report a nodule in the left lower lobe, that that is the same nodule being reported by LUNIT or other readers?

- The limitations of CXR reading are well known, even in the hands of expert readers. Have the authors considered verifying the accuracy of their 'ground truth' readers on a set of CXR that have CT-scans performed within 1-2 weeks, using the CT scan as the reference? Perhaps even comparing Lunit as well, against the CT scan rather than reading by two human experts?

- The sample size calculation is based on 500 cases-- it would be useful to have a statistical reviewer comment on whether there are concerns about applying this approach when the dataset consists of 10 sets of 40 images (plus 1 set of 100 normal images), with each of the 10 sets having a distinct diagnostic abnormality.

- It would also be useful to have a statistical reviewer comment on estimating PPV and NPV when the prevalence of the abnormalities is set by study design rather than reflecting prevalence in a real-world scenario.

- Is the study powered for subgroup analyses?

- I am concerned that the approach for assessing accuracy of human reading with AI support could over-estimate the benefit of AI because of the study design wherein all humans will first read images without AI and then the same set of images will be read again with AI. The authors will change image presentation order and use a 'washout' period to try to address potential bias-- can they please provide data supporting the effectiveness of this approach in mitigating the presumed improvement in accuracy one might have at a second reading of the same CXR set? Did they consider other approaches to address the potential bias, such as: randomly assigning some people to have the AI support during their first reading of the CXR image, rather than with the second; vs. splitting the dataset such that the same CXRs are not being evaluated?

- Will the study funders have influence on study design, analysis, reporting/decision to publish?

Reviewer	**3**
Name	**Hayashi, Shuto**
Affiliation	**Tokyo Medical and Dental University, Medical Research Institute**
Date	**19-Feb-2024**
COI	**I have no competing interests to declare.**

This study aims to explore the potential of artificial intelligence (AI) to assist physicians in interpreting chest X-rays (CXRs) and enhance the quality and speed of diagnosis. To this end, the performance on 500 CXR images without and with AI assistance will be compared. Overall, the study is well-conceived and thoughtfully designed.

Major Comments:

1. While the authors propose a washout period of four weeks to mitigate recall bias, it is worth questioning whether this interval is sufficient to completely neutralize the influence of the first session, particularly concerning diagnostic speed. I would recommend considering the addition of a control group that does not utilize AI assistance in both the first and second sessions. This approach could more effectively eliminate the influence of recall bias.

VERSION 1 - AUTHOR RESPONSE

Reviewer 1’s Comments (Dr. Jarrel Seah):

- Ground truthing - the manuscript describes a two plus one method of consensus ground truthing. However as multiple findings are being assessed - will this be performed on a per finding basis or will any discrepancy in any of the findings lead to all findings being adjudicated by the third truther? –
  This is a very pertinent observation and one that we needed to consider in our study design. For each case, the ground-truthers and the readers will be asked to select all the possible options that an abnormality could be categorised as. The arbitration will be done at a finding level and the arbitrator will only review the findings where there is a disagreement between the initial ground truthers.

Reviewer 2’s Comments (Dr. FA Khan):

- Can the authors please provide more information on the selection process for the images? I am specifically interested to know if the 40 chosen for each type of abnormality are a consecutive or random selection, and if not, what steps might be taken to ensure the cases represent the full spectrum of severity to avoid spectrum bias? –
  A random sampling approach will be taken when selecting the abnormal cases. The following sentence has been added to the cases selection section to clarify this: *“A random sampling*

approach will be taken to ensure that the cases represent the natural spectrum of disease severity."

- How will anatomic location of an abnormality be taken into account when comparing the 'ground truth' reading to LUNIT or the other human readers? For example - how to ensure that if the ground truth readers report a nodule in the left lower lobe, that that is the same nodule being reported by LUNIT or other readers? –
  The ground truthers will add a region of interest to the image and the readers will mark any abnormality with a click-point. The locations will be matched to ensure that the correct pathology has been identified. The following sentences have been added to the methods section:
  "The ground truthers will be asked to mark the location of the abnormality with a region of interest."
  "Where a case is deemed to have a positive finding, the readers will be asked to click on the image to indicate the abnormality location."

- The limitations of CXR reading are well known, even in the hands of expert readers. Have the authors considered verifying the accuracy of their 'ground truth' readers on a set of CXR that have CT-scans performed within 1-2 weeks, using the CT scan as the reference? Perhaps even comparing Lunit as well, against the CT scan rather than reading by two human experts? –
  Thanks for the suggestion. Where this data is available, we will perform a secondary analysis using the CT results as a reference standard. The following paragraph has been added to the ground-truth section:
  "Where a contemporaneous chest CT scan is available (scan performed within 2 weeks of the CXR), an analysis will be performed using the results of the CT scan as the reference standard."

- The sample size calculation is based on 500 cases - it would be useful to have a statistical reviewer comment on whether there are concerns about applying this approach when the dataset consists of 10 sets of 40 images (plus 1 set of 100 normal images), with each of the 10 sets having a distinct diagnostic abnormality –
  Using the conservative assumptions of a 4:1 normal to abnormal ratio, moderate reader accuracy, high inter-reader variability and a moderate improvement in AUC, 30 readers provide sufficient power for the study.

- It would also be useful to have a statistical reviewer comment on estimating PPV and NPV when the prevalence of the abnormalities is set by study design rather than reflecting prevalence in a real-world scenario –
  We agree that positive and negative predictive value analysis would be misleading in the context of an artificial disease prevalence and this has been removed from our analysis plan.

- Is the study powered for subgroup analyses? -
  The study has been powered to detect an overall change in performance for each of the 10 pathologies.

I am concerned that the approach for assessing accuracy of human reading with AI support could over-estimate the benefit of AI because of the study design wherein all humans will first read images without AI and then the same set of images will be read again with AI. The authors will change image presentation order and use a 'washout' period to try to address potential bias - can they please provide data supporting the effectiveness of this approach in mitigating the presumed improvement in accuracy one might have at a second reading of the same CXR set? Did they consider other approaches to address the potential bias, such as: randomly assigning some people to have the AI support during their first reading of the CXR image, rather than with the second; vs. splitting the dataset such that the same CXRs are not being evaluated? –

We did consider starting some readers with the AI aided reads first but felt that providing AI outputs to readers poses a greater risk of recall bias compared with starting all the readers with unaided reads. The readers will be blinded to the ground truth diagnosis and their interpretation of the same case by themselves or by other readers during phase 1. As per reviewer 3 suggestion, we will add an arm to the study where readers perform the unaided reads twice to assess any change in performance. This will help answer the query regarding recall bias.

A paired reader and paired patient study design has been chosen as this reduces variability by ensuring that like patients are compared and like readers are performing the interpretations. Paired designs also require smaller sample sizes than unpaired, randomized designs (*Zhou XH, Obuchowski NA, McClish DL. Statistical Methods in Diagnostic Medicine. 2nd ed. New York, NY: Wiley & Sons, 2011*).

We acknowledge that the above are not perfect solutions but the best practical options given time and resource constraints. We have acknowledged this as a study weakness.

- Will the study funders have influence on study design, analysis, reporting/decision to publish? –
  The funders will have no input in the study design, analysis, reporting or decision to publish.

Reviewer 3's Comments (Dr. Shuto Hayashi):

- While the authors propose a washout period of four weeks to mitigate recall bias, it is worth questioning whether this interval is sufficient to completely neutralize the influence of the first session, particularly concerning diagnostic speed. I would recommend considering the addition of a control group that does not utilize AI assistance in both the first and second sessions. This approach could more effectively eliminate the influence of recall bias. –

  Thank you for your suggestion of introducing a control group to mitigate recall bias and truly isolate the impact of the AI tool on reader accuracy for our analysis. We have incorporated this into our study design and the following text has been added to the manuscript: *"Five additional readers, one from each clinical specialty group, will be selected as a control group. They will perform unaided reads in both phases and their results will be used to assess for any improvement due to learning effects."*

VERSION 2 - REVIEW

Reviewer	**1**
Name	**Seah, Jarrel**
Affiliation	**Annalise-AI Pty Ltd**

Date	**28-Jun-2024**
COI	**Harrison.ai employee**

Thanks for the answer to my question. The arbitration process as described in the response should be inserted into the manuscript along with a brief discussion on how this might the quality of the ground truth.

VERSION 2 - AUTHOR RESPONSE

Reviewer’s Comments (Dr. Jarrel Seah):

- Thanks for the answer to my question. The arbitration process as described in the response should be inserted into the manuscript along with a brief discussion on how this might the quality of the ground truth – The process of arbitration is now described in the ‘Ground truthing’ and ‘Performance of readers with and without AI assistance’ paragraphs of the ‘METHODS’ section.

VERSION 3 - REVIEW

Reviewer	**1**
Name	**Seah, Jarrel**
Affiliation	**Annalise-AI Pty Ltd**
Date	**17-Oct-2024**
COI	

Nil