Peer Review File

Article information: http://dx.doi.org/10.21037/jmai-20-29

Reviewer Comments:

Comment 1: The authors have made a review of how Artificial Intelligence can be used in the diagnosis process of lung nodules. They look into the work done so far for two main applications in which AI has been used so far, which is nodule detection and nodule malignancy characterization. The topic is important and the analysis timely given the exponentially increasing presence of AI in research and the high level of excitement or concern that such technology triggers among clinicians.
The paper is well written and well structured. The authors provide a nicely constructed list of issues associated with the use of AI for evaluation of lung nodules. The main criticism I would draw from the paper is that the authors present the list as a litany of problems without pointing to solutions that either have started to be implemented or should be implemented for AI to be useful and used in practice. This left me as a reader as if AI will never make it in the field, whereas I personally believe it is too early to throw the towel. Instead, the authors could have enriched the article with directions, guidance in order to help the researchers and clinicians developing AI solutions to be in the right path.

Reply 1: We appreciate your overall comments. This review is written from the point of clinicians who are experts in lung nodule evaluation. Therefore we focused the suggestions on the importance to demonstrate clinical utility of AI systems in addition to analytical validation which most studies have demonstrated. This is discussed in "steps to implementation of AI systems in clinical practice".

Comment 2: - Line 194: “False-positives”:
o The authors should clarify that their comment in the paragraph starting at line 194 refers to “false positives in nodule detection” rather than “false positive” (the problem of FPs for nodule characterization raise very different questions).
o For clinical evaluation of nodules, what would be the clinical applications for which acceptable levels could be achieved?

Reply 2: We added the "false positives in nodule detection" for clarification. Regarding the second question of Comment 2, the specific question was not clear to us for this revision. However, when applying any tool for lung nodule evaluation, it is important for clinicians to consider their accuracy as well as their limitation. In general, models to estimate the probability of malignancy of lung nodules perform best when applied in the population from which they were developed. There is no clear or universally accepted "acceptable levels" when it comes to false-positives. However, they are important to be recognized when counseling patients on the use of AI systems in lung nodule evaluation.

Comment 3: - Line 203: Lack of reasoning: this is an insightful analysis, yet it would be good if the authors could clarify what would be required to address this issue. What clinical evidence would be needed to alleviate this concern?

Reply 3: Thank you for your comment. Our understanding is that "lack of reasoning" is intrinsic of how AI systems for lung nodule evaluation work. Although we don`t have specific suggestions to address "lack of reasoning", we provide general suggestions about what evidence is needed for AI to be accepted in clinical practice. As we discuss in the section "Steps to implementation of AI systems in clinical practice" the next steps should focused on demonstration of clinical utility, improvement in the efficiency of care, compliance with care recommendations, and cost-effectiveness.

Comment 4: - Line 221: Liability issues: I am afraid I find this argument a bit too sensationalist. The FDA provide excellent guidance as to what evidence the manufacturers should produce, precisely to clarify that the liability either rests with the manufacturer of the device or the clinician who uses it. AI is not a black box with a mind of its own. The approval of medical devices rely on demonstration of clinical efficiency and safety, whether using AI or not, so the liability question is not more important if the devices use AI.

Reply 4: We made modifications in the text so the argument does not sound sensationalistic.

Comment 5: - Line 229: the scarcity of labelled data: indeed it is an issue for researchers and manufacturers, yet collection of such data is eminently doable as many published research have demonstrated.

Reply 5: Thank you for the comment. We did not intent so say that it is not doable. We just point out that large datasets are necessary. Building large and accurately annotated data is certainly a challenge.

Comment 6: - Line 238: confidentiality of data: the issue of confidentiality comes up if the data is not anonymized properly. Such anonymization is very doable and has been achieved on a very large scale (see the NLST or NELSON datasets and many others), the problem is very solvable.
Reading these, the authors present a list of problems, but they all have solutions. I strongly recommend a rewording of this section to clarify this. It would change the tone and the power of the paper.

Reply 6: Thanks for the comment. We just point out that confidentiality of data is necessary. In the beginning of the section "Challenges with the use of AI" we changed the word "limitation" to "challenge".

We did not feel that the text overall had a negative tone towards AI. Most of the text is spent discussing its accuracy and potential use, and steps for implementation. Only a little more than a one page is spent discussing the challenges related to AI therefore we did not feel that we overemphasized them.

Comment 7: There were additional points which should be addressed as well: they are presented linearly:
- Line 42: “The clinical context which the nodule is found…” -> “The clinical context within which the nodule is found…”

Reply 7: We added the word "within" as suggested.

Comment 8: - Line 90: the authors mention “inter-grader variability”: can the authors clarify what they mean?

Reply 8” We changed the term to "as variability of lung cancer probability estimation between clinicians or radiologists"

Comment 9: - Line 105: the authors refer to the work by “Huang et al” but is not cited in the following sentences nor is listed in the list of references.

Reply 9: thank you for catching that. The reference was added.

Comment 10: - Line 129: “trying to identify small nodules”: I would rather use the word “detect” or “characterise” rather than “identify”, which is a bit ambiguous.

Reply 10: we changed the word "identify" for "detect".

Comment 11: - Line 156-160: the comments made on the research listed in Table 3 is probably correct, however there are important publications I am aware of that are missing and should be added to table 3. They both present results comparing the performance of their method to existing clinical prediction models (line 157) and validated in other populations (line 159):
o https://www.atsjournals.org/doi/full/10.1164/rccm.201903-0505OC
o https://thorax.bmj.com/content/thoraxjnl/early/2020/03/05/thoraxjnl-2019-214104.full.pdf
Therefore, these references should be added and the lines 156-160 should be amended to reflect these findings (or at least not claim that no method has been compared to previous models or validated on independent datasets).
- Line 156-160: I think another comment could be made on methods reporting results using the LIDC data for detection or characterisation of nodules: the LIDC data essentially comprise big nodules, sometimes, masses rather than nodules, not representing lesions typically identified in clinical management of pulmonary

nodules. This makes claim of clinical performance made on such data difficult to port to clinical practice where smaller nodules are seen.

Reply 11: We are happy to include the 2 references suggested to improve our paper to make it useful and relevant. We would like to point out that our review paper was submitted to JMAI before their publication.

Comment 12: - Line 182: The authors refer to the work by “Wang et al” but is not cited in the following sentences nor is listed in the list of references.

Reply 12: thank you for catching that. The reference was added.

Comment 13: - Line 245-246: the authors discuss straight into the question of AI devices adapting their performance in real-time: indeed it is a question to be addressed but I believe the authors should start with AI systems which are designed from a fixed dataset (rather than self-improving ones). The former will make it in clinical practice (some already are available commercially and are used in practice), the latter will not be available in the short to medium term, for the reasons well developed by the authors.

Reply 13: we remove the initial sentence of the paragraph.

Comment 14: - Line 265: “demonstration of clinical utility is not required for FDA”: this is not quite correct. Special controls impose that a device must demonstrate improvement of performance of the clinician using the device versus the clinician not using the device. That could be considered “utility”. But maybe the term “utility” is too ambiguous and maybe the authors should clarify what they refer to. “utility” as a word is not mentioned in FDA regulations for any device, whether AI based or not, but maybe it is because “utility” is not a precisely well defined term?

Reply 14: We expanded the text to clarify clinical utility.

Comment 15: - Line 283-284 (last sentence): this last sentence of the article sets a strange tone for the reader and, as a last parting thought, is probably not the most significant that should be presented. Is liability with the use of AI really the last topic to mention?

Reply 15: we deleted the last sentence.

Comment 16: - Line 429 (table 1): in table 2, the authors have done a nice effort in listing the data used in the validation work, it would be useful to add this in table 1 as well.

Reply 16: The models listed in Table 1 have been validated in several studies as we summarized in " Choi et al. Ann AM Thorac Soc 2018 (10):1117-1126". The List is extensive and we did not feel that it would corroborate to this review paper in particular.