MJ - Decision on Manuscript ID BMJ-2019-052733.R1

Body: 26-Dec-2019
BMJ-2019-052733.R1
Artificial intelligence vs. clinicians – a systematic review of the design, reporting standards, and claims of deep learning studies in medical imaging

Dear Dr. Nagendran,

Thank you for sending us your revised paper. We recognise the amount of work that went into the revision, especially since you had to respond to seven reviewers. I asked our statistical consultant, Dr. Richard Riley, to take a look at the revised paper. He has a few remaining concerns that we would like you to address before making a final decision. I am hopeful that this will not be too difficult.

Thank you again for entrusting us with this important paper. We are very pleased to have it.

Sincerely,

Dr. Elizabeth Loder
Head of Research
eloder@bmj.com

To start your revision, please click this link or log in to your account: *** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

https://mc.manuscriptcentral.com/bmj?URL_MASK=3ac27fc55752453389546788e816d20e

** Comments from our statistical consultant **

Reviewer: 1

Comments:
The authors have responded to a wealth of comments from 7 reviewers. This remains an important article for disseminating the current limitations with AI studies, and to push back against the (often not justified) hype that we currently experience. It is a large undertaking. I think the authors need to do more to address some specific points still. Some responses or corrections are too brief, which probably stems from the wealth of comments receive from reviewers, but also feels somewhat rushed at times. For example, some replies to reviewers are clear and adequate, but there is not a corresponding revision to the article.

Major comments
• The authors have not addressed our suggestion to include a concise summary table in the main paper, of included observational studies summarizing their objectives, clinical context, , etc. They do have bullet points of the study characteristics now in the supp material but suggest it would be too large as a table, and so retain in supp material. I recommend that a simpler version could still be included in the main article and encourage the authors to reconsider this.
• I asked the authors to clarify what exactly they mean by deep learning methods in this paper. They only added one short sentence to address this comment "The volume of published research on deep learning in medical imaging, a branch of artificial intelligence

(AI) in which the algorithm learns for itself which features of the image are important for classification is rapidly growing" – however, this is about as vague as the actual phrase "deep learning". What do the authors mean when they say the algorithm learns for itself. How? What methods does the AI actually implement. There is always some programming behind the scenes for the AI to start the process of learning and optimising the algorithm – so what is the approach (method) utilised in the programming? At least give some common examples. The typical BMJ reader, or indeed any reader, needs to know the context far better here.

• I also asked the authors to provide a Box with some examples to illustrate the type of application and, again, methods that are the focus of this review. However, the authors have not done this either, simply saying in their response to me that ""In plain language, this means the algorithm learns for itself the features of an image that are important for classification rather than being told by humans which features to use" – I hope they can reconsider this

• I asked the authors to clarify in the abstract the type of studies and outcomes of interest. They added the following sentence, which I struggle to follow: "There was no limit placed on the aim or specific outcome measures used in these studies (absolute risk prediction [probability of disease] or classification [disease or not])." – I struggle because the authors say there is no limit placed, but then restrict to risk prediction or classification. So the focus is on the latter types of studies? If so, I think the authors should just say they included studies where the aim was to use medical imaging for predicting absolute risk of existing disease or classification into groups (e.g. disease or non-disease). Emphasise the diagnostic setting again for clarity.

• I asked about why the item within TRIPOD for predictor variables were not considered, and the authors respond clearly that "It is true that deep learning algorithms can consider multiple predictors. However, in the cases we assessed, the only predictors (almost exclusively) were the individual pixels of the image. That is to say the algorithm did not also receive information on for example the patient age, gender, medical history etc." – this information has not been clarified in the article however.

• I asked the authors to better define real-time clinical environment. Their response is clear: "We defined a real-world clinical environment as a situation in which the algorithm was embedded into an active clinical pathway. For example, instead of an algorithm being fed thousands of chest x-rays from a database, in a real-world implementation it would exist within the reporting software used by radiologists and be acting or supporting the radiologists in real-time." – again, this information has not been added to the actual revised article.

• The Box of Terms is a useful addition at the end of the methods, thank you, although currently is not a box as such, but a list of points. Also, the explanation for bootstrapping is a bit vague (at least say each sample is the same size as the mode development dataset) and there is a spelling mistake "… but relies on ransom sampling with replacement" – change ransom to random.

• In response to reviewer 4, the authors give adequate responses about queries to Appendix 3 and Appendix 5 – but again not always has the text been changed or clarified in the actual revision.

• The authors criticise the lack of prospective non-RCT studies – why so critical? Why does data collection need to be prospective? If the aim is diagnosis, then a cross-sectional study may be appropriate. Or if an existing dataset is available, of high quality, then why can't it be used to develop and validate a deep learning algorithm may be appropriate. Prospective may be important to inform the reference standard (true disease status) and impact on patient outcomes, and to compare groups in a trial. So, I think it needs to be made clear that prospective studies are required to make an unbiased comparison on predictions or classifications based on a deep learning method or clinical judgement (and not necessarily to actually develop the model in the first place)

• I think the title should make it clear that the setting is diagnostic

Minor comments

• “Three quarters of studies stated in their abstract that the AI performance was at least comparable to (or better than) clinicians” – suggest change to “Three quarters of studies stated in their abstract that the AI performance was comparable to, or better than, clinicians’ performance”
• Conclusions state that the studies ‘demonstrate substantive bias’ – I think we can’t say whether they ARE biased (we don’t know the truth), just that they are at high risk of bias
• “The authors found that accuracy of cataract diagnosis and treatment recommendation were 87% and 71% respectively” – define accuracy here. Why not split into sensitivity and specificity?

I hope these comments are helpful for the authors going forward.
Best wishes, Richard Riley

Additional Questions:
The BMJ uses compulsory open peer review. Your name and institution will be included with your comments when they are sent to the authors. If the manuscript is accepted, your review, name and institution will be published alongside the article.

If this manuscript is rejected from The BMJ, it may be transferred to another BMJ journal along with your reviewer comments. If the article is selected for publication in another BMJ journal, depending on the editorial policy of the journal your review may also be published. You will be contacted for your permission before this happens.

For more information, please see our peer review terms and conditions.

Please confirm that you understand and consent to the above terms and conditions.: I consent to the publication of this review

Please enter your name: Richard Riley

Job Title: Professor of Biostatistics

Institution: Keele University

Reimbursement for attending a symposium?: No

A fee for speaking?: No

A fee for organising education?: No

Funds for research?: No

Funds for a member of staff?: No

Fees for consulting?: No

Have you in the past five years been employed by an organisation that may in any way gain or lose financially from the publication of this paper?: No

Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?: No

If you have any competing interests <A HREF='http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/declaration-competing-interests'target='_new'> (please see BMJ policy) please declare them here: