Dear Dr. Nagendran,

Thank you for sending us your paper. We sent it for external peer review and discussed it at our manuscript committee meeting. We are very interested in the paper and would like to publish it provided you respond fully to the comments of the editors and peer reviewers.

Please remember that the author list and order were finalised upon initial submission, and reviewers and editors judged the paper in light of this information, particularly regarding any competing interests. If authors are later added to a paper this process is subverted. In that case, we reserve the right to rescind any previous decision or return the paper to the review process. Please also remember that we reserve the right to require formation of an authorship group when there are a large number of authors.

When you return your revised manuscript, please note that The BMJ requires an ORCID iD for corresponding authors of all research articles. If you do not have an ORCID iD, registration is free and takes a matter of seconds.

Thank you again for considering The BMJ as a home for your important paper. We are grateful to have it.

Dr Elizabeth Loder
Head of Research
eloder@bmj.com

*** 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=291ccf82f74c4cf2a27c55d2f703f37e

Report from The BMJ's manuscript committee meeting

These comments are an attempt to summarise the discussions at the manuscript meeting. They are not an exact transcript.

Report of the manuscript meeting of 24 October 2019
Updated By: Loder Elizabeth - Research Committee on 24-Oct-2019

Present: Elizabeth Loder (chair); Richard Riley (statistician); Wim Weber; Tim Feeney; Helen MacDonald; Tiago Villanueva; David Ludwig

Decision: Request revisions. Professor Riley will review the revision.

* We thought this paper was timely, as partly evidenced by the high number of invited reviewers who agreed to look at the paper.

We are also encouraged to see the interest in this topic. The number of invited reviewers has provided a large volume of valuable suggestions that we have tried to incorporate in order to improve the manuscript.

* We agree with reviewer Van Calster that it would be nice to have a table of included observational studies summarizing their objectives, clinical context, findings, etc.

We have responded to this comment in the response to the reviewer. In essence, the table is reasonably large given the high number of studies and we feel it would be more useful to interested parties as a supplementary online excel file rather than unduly condensing it to fit A4 size.

* Our statistician commented that "There is huge hype in the AI field, and much is not justified in my experience. I think the BMJ should be encouraging a more rigorous, transparent and evidence-based framework for the use of AI in healthcare research. Funders are spending billions on this each year. It is the modern version of genetic epi from the early 2000s, promising the world but ultimately delivering relatively little. So, it is important to have critical reviews like this that push back against the (often huge) claims from the AI world." However, he recommends a revision that addresses the following points:

- I think the term 'deep learning' needs defining in the abstract, as this can mean different things to different people. Is it one particular AI method? Or a suite of methods? I think the latter, but this does not come across in the paper.

We have added an extra sentence to clarify this in the first line of the abstract: "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."

- In the introduction, the authors state deep learning methods are "'computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction" – this will mean nothing to most BMJ readers. May I suggest a box with some clear examples (perhaps taken from the actual papers reviewed) to put this into context?

We have added a sentence to clarify what we mean: *"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."*

- What are the clinical fields of interest? Is it any? What are the aims of these deep learning algorithms? Absolute risk prediction (probability of disease)? Classification (disease or not disease)? The abstract should be clearer on the aims (and outcome measures) of the studies being evaluated.

We have added extra clarification to the second sentence in the abstract section on data sources: "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])."

- Why was PROBAST used for risk of bias assessment and not QUADAS-2 (or both?). The latter is more linked to classification (diagnostic tests), whereas the former is linked to individual risk prediction. So I wonder, should the RoB tool have been tailored to the actual research question of each study? Similarly, TRIPOD may not always have been the correct reporting tool to evaluate for non-randomised studies. I am concerned that there is too much of a 'broadbrush' evaluation of the papers, without tailoring assessment to each study based on their specific design, question and actual deep learning method used.

We feel that the PROBAST and TRIPOD tools can be applied to these topics and that they are informative. As things stand, there is no perfect tool developed specifically for AI studies although various efforts are in progress to create more specific tools and reporting guidelines. We had a debate at the outset of the project on the relative pros and cons of various tools as well as the pros and cons of individualising reporting assessment to various sub-types of study versus losing the ability to compare between studies and create a descriptive overview if studies were being judged against different standards. This debate included consultation with author GC who is an international expert and panel member for TRIPOD and PROBAST.

- Why were items in TRIPOD that discussed 'points relating to predictor variables' not relevant to the deep learning? To me, this is hugely relevant as the deep learning approach may consider multiple predictors (features, pathways, information, etc) to inform the predictions it makes. This would lead to overfitting if not mitigated against.

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.

- I'm surprised sensitivity and specificity are not discussed for the trials, as it seems that there is a comparison of true detection rates.

There were many interesting aspects of both trials that were not directly discussed in the results section owing to space constraints and it was not felt to be as important as the standard of reporting and headline results from the trials.

- "tested in a realworld clinical environment" – as defined by what?

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.

- The language in the results quickly escalates to include terms that were not really explained in the methods, like external validation, bootstrapping, split-sample, etc. This mirrors the problem of a lack of context and clear objective of what these studies are aiming to achieve.

We have added these and other terms to a box at the end of the methods section.

- Why do this set of authors represent an unbiased opinion on deep learning, to be able to judge risk of bias etc correctly? Probably needs some discussion, as personally speaking, I might be more critical on a deep learning study given prior experiences of this field.

The authors that conducted the data extraction and risk of bias assessments (MN, YC, CAL) have no overt conflicts of interest in the form of pre-existing research into deep learning. The impetus for the paper from the first and senior authors (MN, MM) was a general feeling that deep learning may be over-hyped. However, the systematic nature of the search and data-extraction process will have mitigated against most bias. Nevertheless, we have added to the limitations section in the discussion to acknowledge that *"risk of bias entails some subjective judgement and people with different prior experiences on AI performance may vary in their perceptions"***.**

* We wonder if some conclusions are too general. One editor comments that you "are correct to caution against unfounded optimism about DL, but one RCT shows superior performance of the polyp-detecting algorithm, and the eye algorithm did a relatively good job in a quarter of the time needed by the consultants." Please comment.

We agree with the editor that there are promising signals from the two RCTs. However, that we found only two published RCTs was an indicator of how young this field is. This is why we tried not to sound too dismissive of the potential for deep learning in medicine which is massive (as opposed to what has been

delivered thus far which is not as extensive as has been hyped). We therefore stated in the discussion that: *"There is enthusiasm to speed up the process by which medical devices featuring AI are approved for marketing. Better design and more transparent reporting should be seen eventually as a facilitator of the innovation, validation, and translation process and may help avoid hype."*

* One editor asked if you might consider giving some examples of the types of health relevant AI headlines that there are/which fields/type etc

We have included two references in the introduction already but have now spelled out the headlines in full as an example of health relevant AI headlines (from the news sources the Telegraph and Fortune).

-I wonder if the authors could work harder at a box/plain English sum up of the problems they find in the literature to date and also potentially a summary of how things could be better.

The problems are mostly summarised by our main findings. We have added some suggestions for how things could be better to the summary box by creating a third sub-section titled: "Suggestions for improvement".

* One editor commented that "if I do a logistic regression and instead of that call it an AI algorithm I can almost guarantee publication."

Book about weapons of mass destruction - putting people at disadvantage

We agree with the editor that sometimes the badging of a method can affect how it is perceived by the wider research community.

* Our patient editor commented that this is @An interesting read and of relevance to the public who read the headlines and are least equipped to critically appraise them. In light of this, what dissemination are authors prepared to do to see that the knowledge gets into the public domain to counteract the hype they speak about? Please have authors share their plans to disseminate to members of the public

We have a multi-pronged dissemination plan as we feel that this paper can help stimulate and direct a very important conversation as it relates to acknowledging the potential and promise of deep learning without being blinded by the hype. The patient editor rightly points out that patients may be in a more difficult position when it comes to appraising the primary research studies that are released and interpreted by the media/press.

In addition to the publication of the paper, we plan to leverage the social media profiles of our authorship team. However, this can lead to an echo chamber in which it is predominantly clinicians and researchers that hear about the news. Therefore, getting the word out to patient groups will also be important. As well as aiming for a national press release and/or local interviews, we would be keen to engage with patient groups to see how they can help us spread the findings from this work.

Reviewer: 1

Comments:
I found this very timely - the authors do well to point out the mis-match between the FDA's willingness to approve the marketing of AI methods and the quality and quantity of evidence supporting them. One expects this sort of unfounded enthusiasm in the popular press (where one also finds, too often, its opposite in the form of, for example, scare-mongering associated with well-proven vaccination programmes), but to find unjustified encouragement at an authoritative level is worrying. I think this point alone makes publication obligatory, but there were a few other points that occurred to me while reading the paper that are mentioned below.

We thank the reviewer for their kind comments and agree with the importance of this issue.

On p 8, the use of the word "cohort" was confusing since in RCTs it usually applies to subjects, but here it seems to mean a medical reference group of which, however, only one member needs to be "expert" - all very puzzling and I think it needs a bit more description.

We have tried to avoid confusion by changing the word 'cohort' to 'group'.

The comments on patient satisfaction on p 11 were opaque to me because it was not clear what the patients had been asked, nor whether they were blinded as to the AI/human variable. This point, and some others, might have been improved had there been lay involvement in the study management, even if limited to the reporting stage (see p 10).

We agree that patient satisfaction is a more nuanced outcome and is heavily dependent on the framing of the question and the context in which the information has been sought. We opted not to delve further owing to word constraints in the manuscript but would be happy to add more detail from the original paper if the editors so wish.

I do not think that the use of "deep learning" software is of great significance here. We are still aiming to compare machine performance with human, no matter how sophisticated the automated method is by its own lights. There will certainly be future developments - "even deeper learning" - but the question remains the same: is it better than having humans do it?

We agree with the author that the core question is probably "*is it better than having humans do it*". This may be on account of superior performance, increased speed, reduced susceptibility to tiredness and lethargy or cognitive biases or even increased cost efficiency which allows funds to be diverted to other areas of the healthcare system. There is also an issue of access to specialist advice in remote regions which traditionally suffer from access compared to urban areas. We chose to focus in this paper on deep learning for two main reasons.

Firstly, covering the entirety of machine learning would have been a prohibitively large task. Secondly, deep learning has received a great deal of attention in the lay press as its major focus on imaging lends itself well to the 'low hanging fruit" of clinical problems (e.g. does this x-ray show cancer or not).

Please enter your name: John Walsh

Job Title: Lay reviewer

Institution: none

Reviewer: 2

Comments:
A systematic review (https://link.springer.com/article/10.1007%2Fs00259-019-04372-x) on a very similar topic was very recently published. That review gave a more low-level overview on image processing itself, had broader inclusion criteria and used a different risk of bias tool.
The review which I am reviewing here is different because the authors provide a more thorough look on RCTs and clearly focus on implications for clinical practice, which ultimately leads to a higher-level review that is adequate to inform clinicians or policy makers on the state of AI in image processing.
This paper provides an important overview, given that mainstream media (and reporting of studies) can often be over-enthusiastic about AI. With respect to this question, both systematic reviews agree.
Especially in the field of image processing it is important to put emphasis on new standardised methods of how to evaluate AI and neural networks.
I think that the methods used for assessing the studies were a good fit, and the reporting of the results is both clear and relevant.

We thank the reviewer for their kind comments and positive feedback.

The nature of these studies led to deviations in reporting and assessment (adaption of TRIPOD tool, exclusion of a PROBAST domain 2 and some single items), but authors disclosed and discussed this. These adaptions are interesting not only in terms of evaluating AI in image processing, but also for AI in text/language processing or any medical application. CONSORT was applied, but not reported in detail (see last comment below).

We agree with the reviewer. Given that we were assessing transparency of reporting we felt it would have been remiss of us not to acknowledge our own protocol deviations and provide justifications for them. The need for adaptations to the existing reporting guidelines also highlights the importance of developing AI/ML/DL-specific extensions to the guidance to facilitate best practice in reporting.

Line 36 Page 9
"The TRIPOD statement consists of a 37-item checklist"
As per the cited literature: "The TRIPOD Statement comprises a 22-item checklist." I figured out that you did not add any items, and that the difference in numbers comes from counting the sub-items in full. Initially, this formulation in line 36 was confusing because it seemed like you added additional items.

We have added a clarification to explain that the difference is the sub-items.

Appendix 2, alteration of first TRIPOD element
'Had to […] mention deep learning or appropriate synonym in title'
'LSTM' or 'CNN', 'RNN' .. are common abbreviations that are often used in titles, and I was wondering if they were included in any of the search strategies?

Abbreviations were not included in the search strategy as our preliminary searches indicated that the abbreviation was almost exclusively spelled out in full in the abstract. Furthermore, the CENTRAL search also looked for the unabbreviated version in the keyword field.

With respect to the last question, I did not find information on how Arxiv was searched (although it was mentioned in the appendix. Did you search all axives?). When doing a quick test, the abbreviations from above did return some results in arxiv title-only searches.
With respect to overall inclusion of studies, could a title like this: 'Longitudinal detection of radiological

abnormalities with time-modulated LSTM' for example have led to the automatic exclusion of this paper, due to any filtering processes within the titles?

We did not search arXiv (nor was this planned). This is because we were looking predominantly for papers published in clinical journals as opposed to pre-prints where there is less guidance available for reporting standards. With respect to the second question, it is difficult to say if such a title would be included without seeing the abstract. All of our search strategies made specific use of the abstract field as well to ensure that there was also reference to a clinician of some type (i.e. the comparator group) so that we did not unnecessarily pickup thousands of papers looking only at the technical development of an algorithm without performance assessment.

Line 44, page 13
The median event rate for development, validation and test sets was 42%, 44% and 44%
I assume that by median event rate, you mean the proportion of images labelled as a 'positive' diagnosis? Does this reflect real-world circumstances (roughly) for any of the diagnosed diseases? It is hard to say, but that might constitute a bias/ affect prediction rates when using these systems in the real world. By that I mean that the human expert would maybe label only 2% as positive, if the proportion of true positives were that low. The network might continue to label around 40% as positive because that is what it learned.

Yes, this refers to the proportion of images being labelled as "positive" for a certain diagnosis. Some studies commented on the difference between the prevalence of disease in their sample versus in real life. In most cases, the prevalence in the learning set was massively increased to provide enough training volume for the algorithm to learn from. We agree with the reviewer that this could constitute a bias when the system then goes in to be tested in a 'real-world' environment and this is why we feel that this is such a crucial step in the evaluation of such algorithms.

Study flow diagram
Exclusion of non-randomized trial registrations (n=76)
I did not notice the mentioning of this exclusion criterium before/ in protocol. Did these not add any valuable information on proposed methods for the upcoming RCTs? For upcoming evaluations on how to review these AI studies, do you think these would be interesting? (although they provide no results yet...).

This was an oversight on our part for which we apologise. The reviewer quite rightly points out that this was not an exclusion criterion in the original protocol. We have appendix 3 which details our protocol deviations and the rationale for them. Non-randomized trial registrations may provide some interesting information but given the sparse data within the registries, it is usually more difficult to parse out the exact details of proposed observational studies compared to randomized trials.

Line 60, page 15
'unintended adverse effects may emerge that are not apparent from an in silico evaluation'
Sorry for my confusion here, but does this mean adverse effects not discovered by the network (due to lack of human instinct etc.), or an actual formation of adverse events due to the use of the technology? Could you possibly give an example for this in order to make it clearer?

We predominantly referred to the latter and have given a brief example to hopefully make this clearer: *"…a higher area under the curve may not necessarily lead to clinical benefit and may even have unintended adverse effects, such as an unacceptably high false positive rate, that is not apparent from an in silico evaluation."*

Landscape format might increase readability.

We have made this landscape as requested.

Figure 3

Similar as above, text on graph is not readable at all in this version.

We wonder if a numbered key would improve readability and make the image less crowded and would be very happy to work with the BMJ editorial and production staff on their style preferences.

Page 13. Line 24
'compared to development only (9/81, 11%) or validation only (9/81, 11%)'
In the appendix (page 20 line 11), 'development only' has 0% associated with it, although wording is exactly the same as in main text.

We agree that the terminology is not always clear. In the appendix, we have stuck to the official phrasing used in the TRIPOD Explanation and Elaboration document. TRIPOD study types 1a and 1b are both technically 'development'. We have changed the wording in the manuscript to clarify this.

Page 11, Line 56
Page 12, Line 21
'the CONSORT checklist (which was included with the manuscript).'
VS.
'the CONSORT checklist (though the CONSORT checklist itself was not included or referenced by the manuscript).'
Possibly include checklist in digital appendix for final version of paper? Could you please have a look which one of these statements is wrong?

These statements do not refer to our including of the CONSORT checklist in our appendices. Rather they refer to whether the RCT authors included a CONSORT checklist with their published manuscript. One did and one did not.

Please enter your name: Lena Schmidt

Job Title: Research Associate in Research Synthesis

Institution: University of Bristol

Reviewer: 3

Comments:
Overall, a rigorous systematic review which addresses the intended aims and clearly guides future research. Methodology is transparent and in keeping with PRISMA reporting guidelines. Reasonable attempt has been made to include relevant publications.

We thank the reviewer for their positive comments.

There are a few areas which I felt required clarification:
1) Why did the review only include papers from 2010 onwards?
2) Page 18, line 34-36: "arguably exaggerated claims regarding equivalence with… clinicians": Although many studies were clearly limited, we cannot necessarily infer that claims were therefore "exaggerated".
3) The fact that 62% did not state that further studies/trials were required has perhaps been overemphasised – although an important negative finding, we cannot infer that the 62% felt that existing studies alone are adequate.

For point one, deep learning came to the fore from around 2014 onwards. The earliest paper included by us was from 2016. It was not felt necessary to extend the search prior to 2010 on the basis of discussions with various stakeholders in the deep learning research arena.

For point two, we agree that this is a tricky area for language and therefore used 'arguably' to add a degree of uncertainty. We agree that study limitations do not in and of themselves mean that claims are exaggerated but a failure to acknowledge the limitations and frame study conclusions within that context certainly makes exaggeration or overstatement more likely.

For point three, it is true that we cannot infer what the authors thought if they have not stated anything about whether or not further prospective work is important. This is unfortunately an area in which press releases and media attention may well 'fill in the gaps'. We hope this paper generates a robust debate about the extent to which authors can reduce the chances of their findings being misinterpreted by couching them in more cautious language and caveats.

Finally, I agree wholeheartedly with the concluding call for further reproducible RCTs, crucially, involving expert comparison groups, prior to AI algorithm use in a clinical setting.

Yes, this will be important going forward.

Please enter your name: Dr Louise C Yates

Job Title: Clinical Radiology Specialty Registrar

Institution: University Hospitals Birmingham

Reviewer: 4

Comments:
This is an interesting review of the quality of and claims made by studies that compare deep learning with clinicians for the evaluation of medical images. Deep learning has become a buzzword that leads to overly optimistic expectations. I have some main issues, and then minor issues and details that mainly relate to further information or clarification of matters that were unclear to me

We thank the reviewer for their kind comments and hope we have now addressed the issues they highlight.

MAIN COMMENTS

1. The paper summarizes claims made by the paper, and describe that most papers make positive claims in favour of deep learning. In that respect, please add information on the reported performance for the observational studies. Basically, an overview table of the 81 papers, including their design (retrospective, prospective, real world), development and/or validation nature, sample size, performance, and claims made would be useful. In contrast to this, reported performance for the two RCTs was described in the text.

We have constructed a table as requested that covers the design type, development/validation nature, sample size, performance and claims made in abstract as well as many other interesting categories of information on the studies which have been touched on by the reviewer in other points. However, given the large number of studies the table is quite large and may work better as an online electronic supplementary table that interested parties can explore. We are happy to be guided by the editors on which format they think would work best for the paper and adjust accordingly.

2. I did not see much overlap between the bullet points about 'what this study adds' and the 5 key findings described in the discussion. Anyway, consistent with the 'what this study adds' section, the high risk of bias and poor reporting of many studies (mainly non-randomized) is a key finding in itself (now hidden in the 5th finding in the discussion).

We have now added the missing key findings from the discussion section into the 'what this study adds' section.

3. On p14, the risk of bias in non-randomized studies is very shortly discussed. It deserves more information, because this is a very important issue. E.g. what does 'analysis domain' refer to? What did the studies at high risk of bias do? In general, the detailed information regarding the content of and results regarding the PROBAST items is too limited/abstract in my view. Please expand.

We have added information to clarify which specific PROBAST items were most deficient into the results section on page 14: *"The major deficiencies in the analysis domain related to PROBAST items 4.1 (were there a reasonable number of participants), 4.3 (were all enrolled participants included in the analysis), 4.7 (were relevant model performance measures evaluated appropriately) and 4.8 (were model overfitting and optimism in model performance accounted for)."*

4. Cf. Appendix 3: The review excluded studies in which clinicians were also involved in determination of the ground truth. However: these studies are also published, and claims are made in these studies as well. I do not understand why these are excluded. These studies should be included and listed as at high risk of bias.

Our aim in this paper was the comparison of AI against clinicians. We felt that the fairest comparison between AI and clinicians would be one in which both groups were compared to an independently ascertained gold standard and therefore the clinicians couldn't be involved in the establishment of this gold standard.

5. The impact of expertise is very relevant (cf fourth key finding in discussion):
   - The definition of expert vs non-expert may be elaborated on more.
   - Did any study try to assess the effect of clinician expertise?

Arguments can certainly be made that expertise is very relevant just as similar arguments can be made for the other points. As mentioned in the methods section, we defined an expert for the purposes of this review as: *"…an appropriately board-certified specialist/attending or equivalent."* **A few studies did try to assess the effect of clinician expertise by for example splitting the human group into various sub-groups (medical student, junior resident, attending for example).**

MINOR ISSUES

- P8, the definition of a 'clinical problem' is vague. Isn't it mostly about diagnosis of some condition?

The main purpose of this criterion was to exclude segmentation tasks (e.g. merely tracing the outline of a cancer in a CT scan) where a patient is not specifically seeking management of their health. A clinical problem could instead be incorporation of outlining the cancer so that the algorithm can then (a) define/diagnose it as a cancer if for example above a certain size and/or (b) provide a management decision (biopsy, surveillance, excise) etc.

- Table 1 is hard to read, please try to rearrange. Also, it is often not easy to know what kind of medical images are being used, e.g. when the intervention is described as 'AI assisted clinic'. It is unclear what is done.

We have made the table landscape as requested by another reviewer. All references to AI-assistance refer to deep learning imaging diagnosis embedded into the usual clinic pathway that would ordinarily consist of human only interpretation.

- P11, what does 'accuracy for treatment recommendation' refer to? What is the ground truth here?

We have added to the text to clarify this. Accuracy for treatment recommendation was the decision for surgery or follow-up and the ground truth was the decision from a separate gold standard group of cataract experts.

- P11-12: please add sample size when describing the RCTs in text.

We have added these as requested.

- P13: the 7 studies that claimed that the algorithm could now be used in clinical practice, where these mainly the prospective real world validations?

Not necessarily. Only 2 of 7 were assessed in prospective real-world validations. We have added this to the manuscript text.

- P13: what kind of sample size calculations did these 14 studies do? Otherwise, this claim means little to me.

Detailed information on the exact calculations performed was not recorded during the data collection stage. As per item 8 of the TRIPOD guidance, we only recorded whether the study *"…explain[ed] how the study size was arrived at…"* **rather than how exactly this was done. A large amount of data was collected during the course of this systematic review at the cost of many person hours of labour. This required finding a balance between collecting items that were important to our study aims and that readers would be interested in without having an overburdensome data collection form.**

- P13: did studies using split sample always split into training, validation, and test parts? This seems to be implied, but is not clear. And when providing median sample sizes, how were studies with CV or bootstrapping (instead of split sample) handled? (see also overview in Appendix 5)

Not all studies split neatly into training, validation and testing. Some were development and internal validation only, others were external validation only. This can be appreciated more readily from the newly added supplementary electronic table. Sample size for studies with CV or bootstrapping was the original number. The same applied for studies where data augmentation was used to boost sample size.

- Data augmentation: it would be interesting to add more information on how often different techniques were used. (see also overview in Appendix 5)

Detailed information on the different data extraction techniques was not recorded during the data collection stage (see response above on balance of data-collection versus resource/time cost).

- P14: how many studies are 'the vast majority'? And what happened in other studies? Were all images then rated by 1 of the clinicians?

The data on how many studies had the entire dataset rated by each expert was not recorded (see response above on balance of data-collection versus resource/time cost). Some studies had so many images to be rated that for (presumably) logistical reasons the dataset was partitioned, sometimes in an overlapping nature, between experts. So of 2,000 images, 600 would be rated by each of 6 experts. In other studies, all 2,000 might be rated by each of the 6 experts.

- Nearly two thirds of studies failed to recommend further studies: is this related to whether the study was retrospective, prospective, or a real world study?

It is difficult to say if they are related given how few real-world studies there were.

- Limitations in discussion: did you consider to add ad hoc items to TRIPOD or PROBAST that would be relevant for this review?

We did not consider adding ad hoc items to either TRIPOD or PROBAST as this is likely to be assessed in far more detail in the working groups that are preparing AI/ML extensions to these reporting guidelines.

- Limitations in discussion: What did you mean when saying that generalizing to other types of AI is not appropriate? When modelling electronic health record data, the aim is rarely to compare with clinicians (as you mention in appendix 3)?

Our findings on potentially overexaggerated findings and deficiencies in reporting would not necessarily apply to non-imaging studies (for example: a paper using reinforcement learning to predict fluid and vasopressor dose requirements in sepsis [Komorowski et al. Nat Med. 2018]). It may be that there are very similar issues in many other types of AI paper. However, we cannot say this from our findings as we only assessed medical imaging studies.

- Figure 1: the first reason for excluding full texts (34 excluded because of it) was unclear to me. It seems like a mixture of things?

These were slight variations on the theme of whether there was an independent gold standard with a non-historical human comparison group.

- Appendix 3: The issue on overfitting is unclear to me. I do not understand what you are referring to here.

We meant that there is potential for overfitting of a predictive algorithm to the index dataset used for development.

- Appendix 3, on expertise level: do you mean that detailed level of expertise was not recorded? Confusing, because you do make a distinction between experts and non-experts.

We meant that our definition of expertise was: *"an appropriately board-certified specialist/attending or equivalent".* **However, within this there could be an 'expert' who had only just become board-certified versus another who was an internationally renowned leader with several decades of clinical experience. This latter degree of detail was not consistently reported and not extracted by us during data collection.**

- Appendix 5: 9 studies were 'development only' studies? This deserves more attention, it seems very problematic. But then later, when TRIPOD study types are presented, 0 studies were 'development only'. This was confusing, please clarify.

We have addressed an identical point in our responses to reviewer 2: *"We agree that the terminology is not always clear. In the appendix, we have stuck to the official phrasing used in the TRIPOD Explanation and Elaboration document. TRIPOD study types 1a and 1b are both technically 'development'. We have changed the wording in the manuscript to clarify this."* **See page 13:** *"…compared to development only with validation through resampling…"*

- Appendix 5: the external dataset testing bullet is unclear to me.

This refers to a dataset that is completely separate from the index dataset. In some cases this was domestic data and in some cases a foreign dataset was obtained.

- Appendix 5: Training set events: give an overview of outcomes. If outcome was non binary, was it always nominal? How many studies had a binary outcome?

We have briefly mentioned this in the results text on page 13 where the number of studies that use a binary outcome (62) as opposed to multi-class classification (19) is described.

- Appendix 5 on code availability was too cryptical for me. Does 'modelling' mean that the model was made available somehow?

Modelling refers to the code used to construct the actual deep learning algorithm (as distinct from pre-processing code used to sort and label the data prior to modelling).

DETAILS FOR CONSIDERATION BY THE AUTHORS

- P6 'raw data': perhaps be more specific here, the rest of the sentence seems to refer to medical images.

We have elected not to make this change to keep the flow of the sentence.

- P6, the second step of the path to implementation involves RCTs to evaluate real world impact and usefulness. True, but it is a general problem of prediction models that these studies rarely happen. This deserves a comment, I feel. Detail: at this point, was unclear to me how these two steps related to the aim of the current paper. Perhaps in the last paragraph you may mention that you aim to search for randomized and non-randomized studies on the evaluation of medical images.

We have elected not to make this change here as we discuss this point in more detail later in the discussion section. We appreciate that there are practical and resource reasons why RCTs will not always occur but this is still a reasonable standard to strive for as exemplified by the two deep learning RCTs that have been published.

- P9, about study selection and extraction of data:
  i. the fourth person (DR) for abstract screening was not mentioned yet (only later, in acknowledgements).
  ii. Full text assessment: was this done by the same people who screened the abstract?

For the first part, DR was involved in study selection but not data extraction and had no further part in the research. As criteria for authorship were not satisfied, the contribution by DR was duly noted in the acknowledgements section.

For the second part, yes, the full text assessment was done by the same people who had screened the abstract.

- P10, data synthesis: perhaps mention the pre-specified and post hoc features for descriptive analysis?

These are detailed in appendix 3.

- P11-12, only risk of bias for blinding: this is not easy to avoid?

It is true that blinding may be difficult in some situations. However, this remains part of the assessment system for the Cochrane tool and the fact that one of the trial groups was planning to repeat their study with blinding suggests it is possible (albeit difficult).

- P14, lines 8-10: sentence (about volume and granularity etc) was unclear to me.

This refers to the amount of separate data available about the expert group. Some papers included little details other than stating that some humans were board certified. Other papers added how many years experience the clinician add and whether there was any specific sub-specialty experience for example.

- P16: internal review refers to review by FDA staff?

Yes, this is internally by FDA staff as far as we are aware.

- Search strategy in appendix: typo in 'deuplicate'

We have corrected this error.

- Appendix 3 is very nice, thank you. Is this referred to in the text?

We have added a reference to this in the first line of the methods section: *"…with any deviations from the protocol detailed in the Supplementary Appendix."*

- Appendix 3: 'we may have missed studies in non-imaging areas': so what was the initial aim? Any study comparing any AI/machine learning method with clinician judgment?

We initially planned to include deep reinforcement learning algorithms as well but later decided against this for the reasons given in appendix 3.

- Appendix 5: What is meant with non-random split? Chronological or geographical split?

This could be any form of split that was not explicitly randomised. The two provided by the reviewer would be common examples.

- Appendix 5: Validation set size: available in 37 of X studies

Some studies used only development and testing.

- Appendix 5: 'requirement for prospective +/- trials': what does +/- mean?

This refers to the authors stating a requirement for prospective work (or if already prospective then RCTs). We have added extra words to clarify in appendix 5.

- Appendix 5: write GPU in full

We have made this correction

Please enter your name: Ben Van Calster

Job Title: Associate Professor

Institution: KU Leuven (Belgium)

Reviewer: 5

Comments:
Summary:
In this interesting and timely report, Nagendran et al. run a systematic review of the literature on deep learning (AI) in medical imaging applications. The primary finding is that there were few RCTs, and the data usage standards for studies based on observational data were generally low. They conclude that the the current quality of the evidence base for the non-inferiority or superiority of AI-based disasse detection or diagnosis compared to clinical experts, is poor, and that conclusions in the literature are generally exaggerated, considering this weak evidence base.

Comments:
I enjoyed reading this review and it comes at an appropriate time in the development of this nascent field. While we are all hopeful about the promise of such technology, I entirely agree with the author's motivation in performing this review, and the execution and conclusions are, to my judgement, generally sound.

We thank the reviewer for their kind comments.

I have only one major comment which I hope will improve the quality of the study. The study nicely reports on several critical aspects of deep learning in imaging: sensitivity/specificity, internal vs. external train/test, prospective vs. retrospective, and randomized vs. observational data, reproducibility and methodological reporting.

However, it omits the functional robustness testing of these algorithms, which is another critical layer of evidence alongside RCTs, external validation etc. Broadly speaking, this tests whether the algorithms function as claimed, e.g. if the 'algorithm can diagnose skin cancer from digital images of nevi', then it must be able to do just that, as plainly understood by common sense. This is distinct from the question of how well it performs that function, which is (partly) addressed by statistical questions of e.g. diagnostic accuracy according to the application.

We agree that functional testing forms an important component of the pathway from development to validation and ultimately real-world implementation.

The need for this testing arises from the manifest "brittleness" of deep learning which is capable of memorizing the training data and therefore largely operates in the non-statistical, "interpolation" regime. Therefore, it is at extremely high risk of merely leveraging deterministic confounding information such as patient traits or image acquisition device properties inadvertently captured in the training data. This is (apparently) something quite unique to deep learning, and thus requires special consideration distinct from reviews of other statistical prediction methods.

Functional robustness is another critical "layer" of evidence on top of the gold standard of RCTs, since an algorithm could perform well in an RCT yet the design of the intervention may not probe for the kinds of functional tests discussed above. This is because an RCT is limited to fixing the entire distribution of the data apart from randomizing the controlled variable. Whereas, functional robustness testing measures the change in algorithm performance along multiple variables without performing explicit interventional experiments.

These tests are also different from tests of statistical performance under "real-world" clinical conditions, since real-world clinical testing does not expressly probe for factors as in the above examples and thus provides limited, if any, information about how the algorithm responds to changes in individual variables.

We agree with these points. Real-world testing in and of itself may not uncover certain issues that arise during implementation. This is not wholly dissimilar to non-AI clinical trials. An example might be a surgical

trial in which the real-world success of a procedure is markedly lower than in the RCT as the RCT operators were surgeons from high volume centres with more throughput of cases. Functional testing of AI speaks to its external validity.

To my awareness, several such functional robustness studies of medical deep learning algorithms have been published, although the evidence base is extremely sparse (much as with very few algorithms being subjected to RCTs):

- In this study, a deep learning algorithm to detect pneumonia in chest radiographs was subjected to functional testing in terms of the robustness of the algorithm to changes in hospital setting:
https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683

- Similarly, here, a deep learning algorithm for melanoma detection was tested for robustness to surgical skin markings:
https://jamanetwork.com/journals/jamadermatology/article-abstract/2740808

- A deep learning system for detecting fractures from radiographs was subjected to several tests addressing robustness to hospital process variables and patient traits:
https://www.nature.com/articles/s41746-019-0105-1

Addressing will likely require re-examining the papers under review to evaluate whether they have performed these kinds of functional robustness tests. Some potential examples of functional tests I can envisage: If the algorithm is presented with comorbid conditions, can it ignore these and focus on the information pertinent to the primary diagnostic task? Does the algorithm still work if different clinical protocols are applied (such as, changes to protocol about the use of sizing rulers on tumour radiographs)? Does the algorithm produce the correct decision for normal anatomical variants? What happens when the digital imaging resolution changes? When there is expert disagreement, how does the algorithm change if competing labels are used to train it? These are just suggested ideas.

The suggested ideas are very interesting and actually overlap with future work that we are planning in this sphere. Although we could look through our existing papers to try and answer these questions, the workload would be prohibitively large and require a significant time investment. Given that the search was performed last in June, we would prefer to get the important messages from this paper out into the research and public domains (as emphasised by other reviewers and editors who asked about our dissemination plans and highlighted the timeliness and importance of this paper). We have added a comment to the discussion section to acknowledge this issue: *"Even in an RCT setting, ensuring that functional robustness tests are present is crucial. For example, does the algorithm produce the correct decision for normal anatomical variants and is the decision independent of the camera or imaging software used?"*

Please enter your name: Max A. Little

Job Title: Senior Lecturer

Institution: University of Birmingham

Reviewer: 6

Comments:
In order to give some context to my review I need to point out at the outset that I have no technical knowledge of systematic review and found some of the research challenging to read and understand. That said I view my role as a Patient and Public Reviewer to comment on whether the review/ subject is of interest to patients, has involved them meaningfully and are the conclusions relevant to patients and the public.

This is a very interesting topic of particular relevance to me as I have had extensive imaging including yearly, monitoring scans. I agree with the authors about media headlines fuelling hype in this area and was shocked, but not surprised when the review exposed the lack of actual evidence base. Of particular concern is the FDA approving 16 deep learning algorithms for marketing despite only one RCT registered in the USA.

We thank the reviewer for their kind comments and agree with their observation regarding the FDA.

Relevance to Patients
There is no doubt that this topic is of great relevance to patients, access to and accurate, timely reporting of imaging can improve diagnosis, monitor treatment and contribute to disease management. Inflated claims in the press regarding AI, without the evidence base to back them up, are potentially damaging. I was particularly concerned to read about the website address in one abstract that allowed patients to upload their eye scans and use the algorithm themselves.
The authors are to be congratulated for so clearly outlining the lack of RCT's in this area, highlighting the need to temper conclusions, improve reporting, transparency (especially around sharing the algorithm and data used), eliminate bias and cut through the exaggerated claims.

We agree that the issues raised in our paper have major relevance to patients.

Missing Topics/ Issues

The major flaw for me in this research is the absence of any patient/ carer involvement which would have contributed to the research and the issues it raises being " patient insight driven " as opposed to surmising what patients think. That might have raised issues such as accountability, who has ultimate responsibility for results when " machines " read a scan? Does AI help the problem of so called " incidental findings "? Does AI improve the interpretation of scan results in the context of ongoing symptoms, what amount of clinical information is optimum and should patients have a role in describing their symptoms to potentially improve diagnosis? Is it only clinical signs and symptoms that are helpful in this context, if patients don't get to input their data/ symptoms/ experience are we missing a vital piece of the diagnostic jigsaw? I have a rare disease, my experience has been that the interpretation of imaging can be challenging and can affect diagnosis and treatment .This touches on the authors definition of expert as " appropriately board certified specialist attending or equivalent ", do they mean any qualified radiologist? Is the definition relevant when considering specialist scans or imaging in rare/ complex disease? If there are relatively few specialist radiologists in areas such as neuro radiology could AI potentially improve the reporting of imaging in geographical areas that don't currently have access to them and so improve patient care? The authors talk about protecting patients and make assumptions about what patients think yet sought no patient insight or involvement.

The reviewer makes excellent points on the nuances of the interactions that AI will in future have once implemented into real time clinical pathways. The example given by the reviewer of their own condition being rare and subject to variation in interpretation is especially poignant as AI proponents hope that such variation in practice will be reduced in future by AI though this is by no means guaranteed. As we mention below, it was an oversight at the design stage of this project not to involve patient representatives to ascertain what features are important to them and whether existing research considers these issues.

In conclusion this is a robust systematic review which highlights many issues and makes a series of excellent recommendations none of which I would take issue with. Its very timely given the media interest in this field and the potential for more trial results in this area being reported in the next few years.
The AI-TREE collaboration to guide best practice in AI for healthcare is interesting and the authors are right to highlight its importance, its disappointing that there is no patient involvement in this initiative, its confined to clinicians, methodologists, statisticians, data scientists and healthcare policy makers.

We agree that patient involvement in such collaborations is important although, to our knowledge, the AI-TREE group does not include patient involvement. In raising awareness of this paper, we hope to bring more patients into the conversation so their views form the basis for designing future research in this area.

I am left with the feeling of a missed opportunity to incorporate the views of patients and carers in this important debate, AI is new and innovative yet once again patients have been relegated to the role of passive recipients in this debate rather than active participants, another thing done to us rather than with us. I am left wondering how some patient involvement and insight may have altered the conclusions and recommendations of this otherwise excellent review, a wasted opportunity?

We agree with the reviewer that this is unfortunately a missed opportunity and if we could have gone back to the planning stages, we would have made greater effort to recruit patient representatives into the project team to help ensure a more patient-centred focus to some of the elements as described above by the reviewer. We are planning future work in this sphere that will take account of this feedback and definitely involve patient representatives in the team.

Please enter your name: Lynn Laidlaw

Job Title: Reviewer

Institution: BMJ PPS

Reviewer: 7

Comments:
This is an interesting and timely article. I have a few minor comments.

Study selection - how were the "clearly irrelevant" records removed? That is, how many reviewers screened them?

These were removed manually by 4 reviewers where the title was clearly not at all relevant to the review (for example the term AI picked up studies of aromatase inhibitors (AI) in oncology).

I'm not sure why predictor variables are irrelevant here. They are not applicable if the algorithm does not use any data other than images as the input, but we very well can build algorithms that take both images and other variables as predictors. It seems to me that the eligibility criteria described here are clear that this review is limited to the former where the sole input to algorithms is images. I wouldn't ask you to revise your assessment of studies but I'd ask you to state that you did not use this item in TRIPOD. This applies to items 1, 5c, 7a, 7b, 9, 10a, 10b, 11, 12, 13b.

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. We have stated the exact items and sub-items used in Appendix 2.

I am not certain of the relevance of reporting hardware configurations. It is pertinent in the context of anticipated run-time delays, so information specific to run-time performance is more informative than the hardware configuration. From a peer reviewer's perspective, I would like to know what computing power was necessary to run the algorithms because it can give me a sense of whether it is realistic as opposed to having been made up. I understand that hardware information is sometimes reported, more so in the engineering literature, but what matters for a clinical application is what to expect in run-time delay. I also think that investigators developing and validating the algorithm are not required to evaluate and report information on expected run-time performance. This relates to the comment on studies that show non-inferior but quicker performance than other methods.

We agree that there are differing interpretations on the relevance of hardware configurations. One view is that they are important for reproducibility and full details should be included. Another view is that they are only important for the validation and testing hardware rather than the initial development of the algorithm. Another view is that they are not important and it is the functional characteristics of the implementation that matter. We haven't taken a specific view and a full discussion would be beyond the scope of this paper but we simply point out that from a reproducibility angle, this would be an interesting area to consider when new reporting guidelines are being formed (for example TRIPOD-ML extension).

The fifth item in the discussion section regarding "descriptive phrases" - did the authors consider assessing the manuscripts for spin?

This is an interesting idea for future work but we did not consider this for the current paper.

I suggest the authors discuss the US FDA guidance on software as a device (reference #28). They are pertinent in the context of which version of an algorithm is evaluated in a randomized controlled trial, and how it is documented in a published report of the trial.

We have added a comment to the discussion section about the FDA guidance and their acknowledgement that the current guidance may not be fit for the AI era: *"The FDA does however recognise and acknowledged that their traditional paradigm of medical device regulation was not designed for adaptive artificial intelligence and machine learning technologies."*

Figure 3 - labels on x-axis are too small to read. Also, is "completeness" of reporting defined in the Methods section?

As mentioned in response to another reviewer: *"We wonder if a numbered key would improve readability and make the image less crowded and would be very happy to work with the BMJ editorial and production staff on their style preferences if the paper is accepted for publication."* **Completeness of reporting is not specifically defined in the methods section - we used the term in a general sense rather than completeness specifically being defined by a level (e.g. 80% or 90%).**

I'm afraid it is not clear to me how the authors evaluated "real world testing". What criteria would make a study deserve the label of "real world testing"?

We considered real-world testing to involve implementation of the algorithm prospectively in an active clinical pathway. For example, a diabetic retinopathy screening programme involving real patients who have their images taken by humans and assessed by the AI algorithm which then outputs a diagnosis or management recommendation.

Please enter your name: S. Swaroop Vedula

Job Title: Assistant Research Professor

Institution: Johns Hopkins University