PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Reporting quality of studies using machine learning models for medical diagnosis: a systematic review.
AUTHORS	Yusuf, Mohamed; Atal, Ignacio; Li, Jacques; Smith, Philip; Ravaud, Philippe; Fergie, Martin; Callaghan, Michael; Selfe, James

VERSION 1 – REVIEW

REVIEWER	Alison Leary London South Bank University University of South Eastern Norway
REVIEW RETURNED	13-Oct-2019

GENERAL COMMENTS	Thank you for this very interesting and timely paper. I think it is a valuable contribution to the field. My comments on the manuscript are primarily on the some style and content issues. I think it is important that such reviews are published in medical (as opposed to informatics etc) journals. Abstract is clear but the Conclusion could be clearer and more impactful. At the moment the Conclusion seems to be one very long sentence. Introduction You do explain what ML is and that models need training but for a medical/health it would be helpful to say exactly what you mean by ML even though it might seem obvious (I have found many different definitions outside of computer science, particularly in medicine!). One of the things you refer to is the need to train but it might be helpful to refer explicitly to the role of data quality (ie top of page 5) and the risk of garbage in garbage out otherwise. Literature search I did wonder why you did not do a wider search but this is explained in the search methods. I do not think you can make the assumption that because journals dont target clinicians, clinicians would not see these papers. I think the limited searching, even though you explain the rationale, is a limitation. It might also be helpful to say something about why TRIPOD should be used (again for general medical readership) The subsequent parts read well and the discussion brings it together. Some of the language could be more precise (ie line 21 pg 10 "A high number of...." Line 29 is somewhat perplexing but seems to be saying something important. What was unclear? As a data scientist I think the point from line 17 onwards is well made-did the papers use "off the shelf" platforms or were

	methodological, statisticians, epidemiologists or data scientists involved? From the figures it looks like you utilised PRISMA which seems sounds but I do not recall seeing that in the text.

REVIEWER	Gilmer Valdes University of California San Francisco, Radiation Oncology
REVIEW RETURNED	05-Nov-2019

GENERAL COMMENTS	In the article: “Reporting quality of studies using machine learning models for medical diagnosis: a systematic review.” the authors performed a meta-analysis/review of the quality of the reports of the articles using machine learning for diagnostic tasks. The objective is to evaluate how these findings intended to be used in clinical applications can be evaluated and reproduced. Given the extensive advises provided on the TRIPOD guidelines and the fact that these guidelines seem to have been accepted by the community given the number of references they have, the articles use the adherence to these recommendations as a proxy for “goodness” of the report. Their main finding is that a big proportion of articles do not adhere to the recommendations and that studies lacked adequate detail on the participants on which the diagnostic task was evaluated. These findings are well supported as illustrated on table 3. The selection criteria for the articles included in the review/meta-analysis is also well documented and represent several different specialties (See Table 2.) Although this article does not provide a unique scientific contribution per se (it is a review article), the same can serve editors and the readership as an additional guideline to TRIPOD highlight common mistakes while reporting results. The same is also well written and organized. Give that it is a review article without unique scientific contributions, I defer the judgment of whether it should be published on the BMJ journal to the editors. Below find small comments/questions: P3 L 41-> Missing parenthesis after 2019. P5 L10-11: The sentence: “The importance of transparent….” is not complete. P9 L42-45: It would be nice to know how many authors acknowledged this limitation in their texts. Since ML models are built using retrospective data, shifts on the distributions (referred in this article as differences between training data and clinical settings where the models will be used) will affect the majority of models in a bigger or lesser degree.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

1) I think it is a valuable contribution to the field. My comments on the manuscript are primarily on some style and content issues. I think it is important that such reviews are published in medical (as opposed to informatics etc) journals. Abstract is clear but the Conclusion could be clearer and more impactful. At the moment the Conclusion seems to be one very long sentence.

We thank the reviewer for their comments and feedback. We agree with you that this review should be within the medical literature and that it is important to raise these issues and iterate the need for higher quality in the reporting of studies using ML methods for clinical diagnosis. We have followed your suggestion and have now streamlined our conclusions in the abstract as suggested…

> *"All studies in this review failed to use reporting guidelines, and a large proportion of them lacked adequate detail on participants, making it difficult to replicate, assess and interpret study findings."*

2) You do explain what ML is and that models need training but for a medical/health it would be helpful to say exactly what you mean by ML even though it might seem obvious (I have found many different definitions outside of computer science, particularly in medicine!).

We thank the reviewer for their constructive comment. We have provided the definition of ML whilst noting the lack of consensus for ML within the field.

> *"While there is no consensus on the definition, a key principle of ML models is that they are developed based on the automatic extraction of patterns from data [7]. In contrast to traditional statistics, whereby models are explicitly programmed based on statistical theory and assumptions, ML models learn from examples without the need for explicit rules to make decisions [8]"*

3) One of the things you refer to is the need to train but it might be helpful to refer explicitly to the role of data quality (ie top of page 5) and the risk of garbage in garbage out otherwise.

This is another astute suggestion from the reviewer that certainly helps clarify few assumptions. Often times we talk about methods and conduct of ML with the assumption that data quality is a given standard, therefore, this is certainly worth clarifying within the text. As suggested, this has now been added to this section:

"As ML models are only as good as the data used to train them, it is vital to emphasise the importance of data quality [9]."

4) I did wonder why you did not do a wider search but this is explained in the search methods. I do not think you can make the assumption that because journals dont target clinicians, clinicians would not see these papers. I think the limited searching, even though you explain the rationale, is a limitation.

Due to the ever-growing ML literature within medicine, we chose these journals to really narrow our scope and use them as an exemplar to explore the reporting quality of studies using ML methods for clinical diagnosis. You are right in saying that clinicians are also likely to see articles within journals that are non-clinically focused, however, it is highly likely that those ML articles that are published within the Clinical core journals are targeted towards clinicians. So, these set of journals present us with a good signal-to-noise ratio. All that said, we acknowledge it as a limitation as it doesn't widely examine the reporting quality within the literature, and we do highlight this limitation within our discussion.

"However, it is worth noting that we have not included all medical journals and therefore our findings may not be applicable to all journals. Despite this, we have included studies published within the Medline Core Clinical Journals, these journals cover all areas of clinical and public health."

5) It might also be helpful to say something about why TRIPOD should be used (again for general medical readership)

We thank the reviewer for their astute comment. Even though TRIPOD was developed for traditional multivariable prediction models in medicine, it is the most rigorous and relevant guideline for evaluating the use of ML for medical diagnosis. Further, most items in the TRIPOD Checklist apply to studies using machine learning, and in our case, it contains the relevant items which are attributed the study populations. This has now included within the methods section:

"For studies developing, evaluating or updating clinical prediction, the TRIPOD provides guidance on reporting the key items. As it stands, TRIPOD is the most rigorous and relevant guideline for evaluating the use of ML methods for medical diagnosis. As such, an extraction list based on the TRIPOD checklist was developed."

6) The subsequent parts read well and the discussion brings it together. Some of the language could be more precise (ie line 21 pg 10 "A high number of...." Line 29 is somewhat perplexing but seems to be saying something important. What was unclear?

We thank the reviewer for their input and suggestion for more precise language. We have streamlined our language in line 21, the sentence reads as follows:

> *"In more than half of the studies, it was unclear whether the study population corresponded to the setting in which the ML diagnostic system will be used in. However, in a third of the reviewed studies, the test populations did not correspond to the populations in which tests were hoped to be applied to, further limiting their generalisability"*

The sentence beginning from line 29 refers to the four studies that were unclear on their data source. This is important in determining how and where the studies samples are chosen from. Some studies conveniently choose from an already available database. Whereas in other studies, they prospectively select samples by using a random sampling method. Knowing this information in studies using ML diagnostics system is key as it tells us a lot about the samples and whether systematic bias such as selection bias exist. What is unclear has now been clarified within the document:

> *"Information on data source was unclear in four studies; this is vital in evaluating the source and methods used to derive study samples. Information on data source is vital in evaluating the source and methods used to derive study samples. In diagnostic studies the use of different methods to derive the evaluation sample from the wider population could lead to more or less accurate estimation of the diagnostic performance. The ideal method for sampling should be based on probability and not convenience, as this allows for a representative sample to be selected from a sampling frame whereby all eligible individuals have an equal chance of being selected."*

7) As a data scientist I think the point from line 17 onwards is well made-did the papers

use "off the shelf" platforms or were methodological, statisticians, epidemiologists or data scientists involved?

The review did not evaluate this particular aspect. However, we thought it was important to iterate this within the discussion as the use of ML methods to solve domain specific problems is cross disciplinary. Therefore, requiring the consultation and involvement of experts within these different disciplines, in our case, methodologists, statisticians or epidemiologists.

8) From the figures it looks like you utilised PRISMA which seems sounds, but I do not recall seeing that in the text.

This is already reported on the first paragraph of the methods section:

"The framework used for this methodological systematic review is Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guideline for Systematic reviews [19]."

Reviewer 2

1) In the article: "Reporting quality of studies using machine learning models

for medical diagnosis: a systematic review." the authors performed a meta-analysis/review of the quality of the reports of the articles using machine learning for diagnostic tasks. The objective is to evaluate how these findings intended to be used in clinical applications can be evaluated and reproduced. Given the extensive advises provided on the TRIPOD guidelines and the fact that these guidelines seem to have been accepted by the community given the number of references they have, the articles use the adherence to these recommendations as a proxy for "goodness" of the report. Their main finding is that a big proportion of articles do not adhere to the recommendations and that studies lacked adequate detail on the participants on which the diagnostic task was evaluated. These findings are well supported as illustrated on table 3. The selection criteria for the articles included in the review/meta-analysis is also well documented and represent several different specialties (See Table 2.) Although this article does not provide a unique scientific contribution per se (it is a review article), the same can serve editors and the readership as an additional guideline to TRIPOD highlight common mistakes while reporting results. The same is also well written and organized. Give that it is a review article without unique scientific contributions, I defer the judgment of whether it should be published on the BMJ journal to the editors.

We thank the reviewer for these comments. In this systematic review, we found that studies developing or validating machine learning based systems for clinical diagnosis failed to use reporting guidelines and the studies lacked adequate detail on the participants on which the diagnostic task was evaluated on, thus making it difficult to replicate, assess and interpret study findings.

As reviewer one highlighted, this is a review on a timely issue. We have utilised a research on research approach to identify and determine the common themes surrounding the reporting quality of studies using ML methods for medical diagnosis. Regarding the scientific contribution of this review, it is important to understand the relevancy of research on research studies. Such studies may not provide direct scientific knowledge from the reviewed studies as traditional systematic reviews; however, they provide invaluable research evidence on the way science is conducted, reported and disseminated. This is so that groups such as funders stakeholders, policymakers and guideline developers can make evidence-based decisions to inform improve the conduct, reporting and dissemination of science. In the case of this systematic review, the aim was to identify common reporting deficiencies within diagnostic studies using ML methods, using studies published within Medline Core Clinical Journals as an exemplar. Just like in Christodoulou et al. review (2019), this review identified some important issues surrounding the reporting quality of studies using ML for medical diagnosis. Such findings may contribute to the development of relevant reporting guideline, such as, TRIPOD-ML which is currently underway (Collin and Moon, 2019).

Below find small comments/questions:

P3 L 41-> Missing parenthesis after 2019.

This citation style has changed, this is therefore no longer a problem.

"Over the past decade, access to large amounts of clinical data and the development of new ML techniques has led to a rise in the application of ML methods to medicine [2 3]."

P5 L10-11: The sentence: "The importance of transparent…." is not complete.

This has now also been resolved within the text:

"Highlighting the importance of transparent and rigorous reporting of clinical predictions models accuracy studies, particularly as the diagnostic prediction models of an instrument can vary greatly due to factors such as population characteristics…"

P9 L42-45: It would be nice to know how many authors acknowledged this limitation in their texts. Since ML models are built using retrospective data, shifts on the distributions (referred in this article as differences between training data and clinical settings where the models will be used) will affect the majority of models in a bigger or lesser degree.

Studies developing ML-based diagnostic methods should be put under the same rigour as typical diagnostic methods/tools within medicine. As previously highlighted by reviewer one, data quality is vital and without it this would have quite an implication on the usability, generalisability and safety of these diagnostic tools. As the current review focused on the reporting quality of the methodology and findings of the reviewed studies, why these items were not reported was beyond the scope of the review.

VERSION 2 – REVIEW

REVIEWER	Alison Leary London South Bank University, 412928
REVIEW RETURNED	02-Dec-2019

GENERAL COMMENTS	Thank you for making the amendments. I think it will be clearer to a non technical reader

REVIEWER	Gilmer Valdes University of California San Francisco, Radiation Oncology
REVIEW RETURNED	22-Dec-2019

GENERAL COMMENTS	The authors have addressed all my concerns and I here recommend the same for publication.