Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-21-25

Important Note

We greatly appreciate the constructive feedback provided by both reviewers. We noticed that many comments addressed aspects that would be important to include in a systematic review-type article. However, we would like to clarify that we conducted a comprehensive literature review, following the guidelines for a 'Review Article' as noted on the JMAI Guidelines for Authors document. This literature review presents a timely, comprehensive analysis of the factors that influence healthcare professionals' trust in medical AI. We apologize for any miscommunications or lack of clarity regarding the article type.

We believe a literature review to be more appropriate than a systematic review at this time as there currently does not exist a comprehensive consolidation of available literature on this topic. As such, a literature review article is an important primary step to synthesize the current literature and consolidate what is already known about this subject. Since this has not been previously performed, it will enable identification of knowledge gaps and contribute to further understanding by summarizing relevant evidence, which is important to do before a systematic review is completed. By first collating information in the form of a literature review, we aim to facilitate an understanding of the landscape in which this information applies to make valuable contributions to the field of medical AI. A systematic review could be a recommended next step, however, initially we wanted to broadly understand the way trust is conceptualized to enable fine tuning of the perception of each factor, ensure no errors of omission etc.

Since this is a literature review-type article, many of the comments/suggestions for edits do not apply.

Reviewer A

Comment 1: Selection bias in studies: one point that struck me is that evidential robustness (e.g. clinical trials) or replicability/reproducibility in studies were not mentioned as factors, conducive to trustworthiness. I find this omission surprising, given the buzz that surrounded the release of the SPIRIT and CONSORT AI reporting guidelines for clinical studies on AI systems.

Reply 1: Thank you for this feedback. The evidential robustness of the selected studies was not mentioned as a factor conducive to trustworthiness because this paper only identified the factors contributing to trust in medical AI as specifically stated by health care professionals in the selected articles. The use of CONSORT-AI reporting guidelines in papers on medical AI was not a factor mentioned in the articles that healthcare professionals considered as contributory to their trust in medical AI and hence it was not mentioned. Further, there were very minimal, if any, clinical trial-based studies used that would require the use of such framework.

Trust in the replicability of studies about "trust in AI" affecting a subjects' concept of trust was not studied. Rather this affects our trust in the concepts we identified as relevant for "trust in AI". So, the issue appears to require clarity that we are "evaluating the concepts relevant for a user to trust AI with which they are working" and not "trust in the papers published in the realm of AI".

Changes in the text: We have tied specific elements from CONSORT-AI to the trust framework by adding that the use of these reporting guidelines is something to consider for clinical trials involving AI to ensure a comprehensive evaluation of the AI technology before deployment and integration into clinical environments (see page 18 lines 305-312). Since this is a literature review, a bias assessment is not necessary.

Comment 2: Moreover, at least a factor that increases my trust in AI applications is that there are useful self-correction mechanisms in medicine, such as meta-analyses which systematically reveal the shortcomings of existing studies. Hence, my concern boils down to the question: would slightly different inclusion/search criteria have led to very different results?

Reply 2: Thank you for this observation. The search criteria of the manuscript captured the broad discourse in the medical community surrounding trust in AI. None of the included or reviewed papers mentioned that the use of meta-analyses was influential in a users' trust in medical AI. As such, this was not included as a contributing factor. Nevertheless, we understand that exposing shortcomings of existing studies can contribute to increasing one's trust in medical AI. Therefore, we did not exclude studies based on design and used a broad search strategy to capture relevant articles on the topic.

Changes in the text: We added an acknowledgment that a useful next step for future research may be a systematic review now that we have consolidated and synthesized the available literature to better understand the research landscape and knowledge gaps (page 27 lines 503-506). We also clarified that we did not exclude articles based on study design (page 7 line 108).

Comment 3: Lack of conceptual clarity: the paper distinguishes between 'transparency', 'interpretability' and 'explainability', without making the differences clear. Since in the ML community, these concepts are oftentimes used interchangeably, it would be helpful to know, how the authors delineate the very concepts. Maybe it would be helpful in general, if the key terms are briefly defined, since their meanings can vary considerably, depending on the context.

Reply 3: We appreciate this feedback. We recognize that since there are many discrepancies in the nomenclature used in literature, there is a lack of clarity in definitions. As such, it is difficult to create an evidence-based definition, as there is much variation. This paper is one mechanism to describe the breadth of such variation and provide a foundational contribution to an eventual evidence-based conceptual understanding. We agree that it is important to acknowledge the lack of clarity before introducing our approach to these definitions based on select relevant papers that make the delineation between concepts.

Changes in the text: The definitions of select terms as they are described in engineering versus medicine are included in the Discussion section, however, we have made the definitions clearer by including them under their own subheader in the Methods section instead, so the reader can better understand the concepts before reading the results (page 7,8 lines 112-129). We have also included a statement in the introduction (page 5 lines 55-58) that acknowledges the challenges in the realm of AI research regarding the inconsistency/lack of universally accepted definitions of these key terms. Given that these are terms anticipated to be relevant for understanding the concept of trust, we were obligated to accept that they may be applied inconsistently in the literature.

Reviewer B

Comment 4: Although a PRISMA checklist is presented, it is not that clear systematic review methodology was properly followed and the methods, as reported, are not currently replicable... It needs to be clear: is this a literature review or a systematic review? If a literature review: why? what will this add to the literature, if methods can't be followed and aren't replicable? If a systematic review, why have you not followed reporting criteria properly?

Reply 4: See *Important Note*

Changes in the text: Justification of why a literature review was performed, and the importance of this approach was added to the Introduction (page 5,6 lines 61-77) and statement regarding the importance of a systematic review as a potential next step has been added to the Discussion (page 27 lines 503-506).

Comment 5: PRISMA checklist line numbers do not correspond with those in the manuscript.

Reply 5: Thank you for bringing this to our attention. The PRISMA diagram has been included as a figure (JPG file). As such, we are unable to have line-by-line numbering.

Changes in the text: None at the moment, however, if there is a preferred way of line numbering a JPG-file type Figure, we are more than happy to adjust accordingly.

Comment 6: I would expect to see: inclusion/exclusion criteria based on PICO or SPIDER criteria and clearly defined study designs which were eligible

Reply 6: Thank you for this feedback. We acknowledged that we did not exclude based on study design as the nature of a literature review is to get a sense of the wide span of available works and it would thus not necessarily be relevant to list out all study designs. We stated that we did distinguish studies by study type (qualitative or quantitative) but we did not exclude based on any specific design.

Changes in the text: We have applied the SPIDER criteria to determining our research question and have noted this in the methodology section (page 6 lines 86-89). See page 7 line 108 for clarification that we did not exclude articles based on study design.

Comment 7: I would expect to see: a full replicable search string; information on who did the screening and how many search results were screened in duplicate; what data was extracted from studies

Reply 7: We have detailed our search terms and strategy (pages 6,7); See *Important Note* as this information does not apply to a literature review.

Changes in the text: N/A

Comment 8: I would expect to see: the method for mapping mentioned trust topics into themes or concepts e.g. thematic synthesis

Reply 8: Thank you for this suggestion. The mapping method was very briefly alluded to, however, we recognize this description may not be very clear.

Changes in the text: Thematic synthesis was added as a description of our approach to mapping (page 9 line 160) to increase clarity.

Comment 9: I would expect to see: quality or risk of bias appraisal

Reply 9: See *Important Note* as this information does not specifically apply to a literature review.

Changes in the text: N/A

Comment 10: In the results I would expect to see study characteristics (including information on study design and participants included), quality appraisal.

Reply 10: We appreciate this feedback. See *Important Note* as this information does not specifically apply to a literature review. As a literature review, conventionally a select few seminal papers are discussed in more detail. Generally, and in accordance with our aims, we focus on providing an overview of the discourse surrounding trust in medical AI according to healthcare professionals, and as such, it may not be relevant to list this information for all 57 included articles.

Changes in the text: We added general statements in the inclusion criteria that participants are certified healthcare professionals to better describe the participants included (page 8 line 89).

Comment 11: In the results I would expect to see search results, quality appraisal, and a thematic or narrative synthesis.

Reply 11: See *Important Note* as quality appraisal does not specifically apply to a literature review. Search results are in the PRISMA flowchart (page 12) as well via the list of trust concepts (pages 13,14,15 lines 226-266) and in Figures 2 and 3 in the Results section (pages 17, 18). The list of influential trust factors in the Results section serves to make the concepts that medical professionals consider as impactful to their trust in AI very clear and readable, as this was a specific aim of the literature review. This list is complemented by a more narrative description at the end of the Results section (page 15, 16 lines 268-285). This narrative description and analysis of results is continued in the Discussion section.

Changes in the text: We clarified the thematic synthesis as it relates to the consistency between papers in the way the terms are defined (page 9 line 160).

Comment 12: I am not sure what the simple quantitative presentation of number of papers mentioning concepts contributes to our understanding of the topic. I took little away from it. Firstly, it doesn't seem to match the aims, which suggested authors were aiming to "elucidate the discourse in the medical community regarding key factors related to trust".

Reply 12: Thank you for this perspective. We respectfully disagree, as quantitative presentation of the number of papers identifies the areas of current focus of research endeavor. If we identify: 1) what are the key factors, and 2) how commonly each are found, then we have identified the most common and least common (or therefore possibly overlooked) concepts. What we have not done is identify concepts that are relevant but not included in prior studies. Therefore, all the concepts identified will be included in a future research project, but subjects will be asked also to identified themes not present or under-represented that they feel are still important for trust in AI.

Changes in the text: We added a justification to clarify the relevance of a quantitative presentation of number of papers that mention AI trust concepts (page 13 line 216-219 and page 15 lines 273-275).

Comment 13: I disagree that you can assume that because a topic is highly investigated by the research community (who may not be clinicians) that it is of key importance in the clinical environment.

Reply 13: Thank you for noting this perspective. This appears to assume that the research community outside clinicians is secondary. A large AI research community could better serve the clinical needs if it better understood the areas that were of concern to clinicians. This approach demonstrates highly investigated items, which may or may not reflect importance. It is clear that if researchers and funding are only directed to certain areas at the expense of others it can be assumed that there is a perceived correlation between frequent investigation and importance of a topic. It reflects a degree of perceived importance; however, we agree that clarification regarding unknown unknowns is merited. This paper is partly meant as a review to draw attention to areas that others who are purely focusing on another topic may not currently recognize.

Changes in the text: We have elaborated to clarify that although it is likely that highly investigated topics reflect some level of importance, it is also possible that the current research focus is misdirected, or that concepts deemed important in one discipline would not necessarily translate to those deemed relevant in another discipline using the same AI. We acknowledged that this does not identify the unknown unknowns, and have reinforced the rationale behind our approach (page 26 line 478-483).

Comment 14: The mapping method between quotes in papers (or results of questionnaires? it is not clear what type of data was assessed) and final concepts is not clear (or replicable). The Table given in Appendix A is inadequate for the reader to understand how the concepts were decided on. A more narrative approach would have given the reader more information about how the factors identified meaningfully influence clinicians' views and trust of AI.

Reply 14: Thank you for this feedback. The mapping method was solely used for AI trust concepts that were alluded to in articles via direct participant quotes or implicit definitions. These concepts were 'mapped' to an explicit definition so they could be accounted for in our quantitative analysis. Our explanation of how the factors identified meaningfully influence clinicians' views and trust of AI is included in the Discussion. The concepts were decided on based on the results of included articles. We agree that a narrative description specifically regarding the mapping method would support the readers' understanding and replicability of the process.

Changes in the text: We have more clearly outlined our approach to mapping concepts and selecting factors from included articles. We explain how the methods section of identified papers was reviewed and mention of key themes was recorded. In paper mentioning multiple themes, each them was individually recorded, etc. (page 9, 10 lines 155-168). We recognize there may be bias in this process, and acknowledge this in the Limitations section (page 26 line 483-487).

Comment 15: If the authors feel that there is plenty of narrative or qualitative literature on this topic, and that a quantitative summary is what is needed now (for a particular reason), that should have been made clear in the introduction and aims.

Reply 15: Thank you for this feedback. We agree that the importance of a current quantitative summary should be alluded to in the aims.

Changes in the text: See page 5 lines 61-64, page 13 line 216-219, and page 15 lines 273-275 for clarification of importance of quantitative summary and another reason as to why this literature review is important and contributes to the current body of literature; AI researchers and developers need some input to better direct their work and make the outcomes more clinically relevant.

Comment 16: I would have expected some of these problems to be addressed in the limitations section, but they are not acknowledged.

Reply 16: We appreciate this observation, however, not performing a systematic review is not a limitation of having done a literature review.

Changes in the text: The significance of performing a literature review rather than a systematic review was addressed in the Introduction (page 5,6 lines 68-77).