Author’s response to reviews

Title: Collective intelligence in medical decision-making: a systematic scoping review

Authors:

Kate Radcliffe (kate.radcliffe@ucsf.edu; kategradcliffe@gmail.com)
Helena Lyson (helena.lyson@ucsf.edu)
Jill Barr-Walker (jill.barr-walker@ucsf.edu)
Urmimala Sarkar (urrimala.sarkar@ucsf.edu)

Version: 2 Date: 23 Jul 2019

Author’s response to reviews:

Please find attachment as Supplementary Material

July 23, 2019

Dirk Krüger, PhD
Editor
BMC Medical Informatics and Decision Making

Dear Dr. Krüger and the Editorial Board at BMC Medical Informatics and Decision Making,

We thank you for providing the opportunity to revise our manuscript (MIDM-D-18-00328) entitled “Collective intelligence in medical decision-making: a systematic scoping review” for consideration as a review manuscript in BMC Medical Informatics and Decision Making.

We are pleased to address the comments from the reviewers and editorial board below. We have revised our manuscript by incorporating your recommendations throughout. Your comments are reproduced in bold with our response following. We reference page and line numbers referring to the exact location of the updated text in the clean version of the manuscript and, when appropriate, we included the updated text in italics in this letter.

All authors have read and approved submission of the manuscript. Authors claim no conflicts of interest. The manuscript has not been published and is not being considered for publication elsewhere.
Thank you for considering our manuscript for publication in BMC Medical Informatics and Decision Making.

Sincerely yours,

(electronically submitted on behalf of all authors)

Corresponding Author
Urmimala Sarkar, MD, MPH
Associate Professor, University of California San Francisco
Zuckerberg San Francisco General Hospital, 1001 Potrero Ave, Building 10, Ward 13
San Francisco, CA 94143
Urmimala.sarkar@ucsf.edu
415-206-4273
Manuscript MIDM-D-18-00328

Reviewer 1: Ralf H. J. M. Kurvers

P2, L23-25. "Studies of radiological…" I would still suggest to clarify better what this inclusion/exclusion criterion constitutes. What is a complex problem, and what is a non-complex problem? Please clarify, I would also start this sentence with the general statement, and only in the end mention the specific examples of radiological scans. (Same for P6, L17).

Thank you for this comment. We endeavor to clarify this as follows:

We include studies of complex medical decision-making rather than identification of a visual finding, as in radiology or pathology. We differentiate between medical decisions, in which synthesis of multiple types of information is required over time, and studies of radiological scans or pathological specimens, in which objective identification of a visual finding is performed. [P2, L10]

Our goal was to uncover the utility of collective intelligence in diagnosis and decision-making, in which multiple sources of objective and subjective data generate a diagnosis over time, rather
than in binary decisions such as identification or absence of a finding in radiology or pathology. [P6, L13]

P3, L11. Is it really true that the average in Galton's experiment (note that Galton used the arithmetic mean to determine the average) was better than all individual estimates? I read the original paper again, but I did not find such a statement. Also, why not cite the original study, rather than a study referring to it?

Thank you for this comment, we agree the original study should be cited. We have added it to P3 L4. We keep Surowiecki’s study as we believe it provides a worthwhile interpretation of the results in the context of wisdom of the crowds.

As to the interpretation of Galton’s results, we revisited the original text and found this statement: “It appears then, in this particular instance, that the vox populi is correct to within 1 per cent of the real value, and that the individual estimates are abnormally distributed in such a way that it is an equal chance whether one of them, selected at random, falls within or without the limits of -3.7 per cent and +2.4 per cent of their middlemost value.” With this in mind, we revised the statement for clarity:

He found that the average of all the estimates was correct within 1% of the actual weight, and the individual estimates were more likely to be incorrect. [P3, L4]

P3, L57. What is mean with "a specific correct and objective "answer""? Please clarify, or give concrete examples to illustrate what is meant with this

Thank you for this clarification. The intention is to describe that diagnosis is the result of multiple streams of information, which makes the process complex, as opposed to binary determination done in pathology or radiology.

For example, as described by the National Academies, which we cite in text, “Performing a clinical history and interview, conducting a physical exam, performing diagnostic testing, and referring or consulting with other clinicians are all ways of accumulating information that may be relevant to understanding a patient's health problem. The information-gathering approaches can be employed at different times, and diagnostic information can be obtained in different orders. The continuous process of information gathering, integration, and interpretation involves hypothesis generation and updating prior probabilities as more information is learned.”

The diagnostic process is highly complex and requires accumulation of multiple streams of information at various time points, whereas we characterize radiology and pathology as one important piece of that process. We clarify as follows:

At one end of the spectrum, collective intelligence can be applied to objective identification of abnormalities on images, whether they are pathologic slides or radiologic scans, and recent research supports collective intelligence in these settings. The pathological and radiological
determinations are critical to contributing objective information to a clinician’s determination about the diagnosis. In contrast, the diagnostic process in the clinical setting synthesizes subjective data, like clinical history and patient-reported information, with objective pathological and radiological findings, to continually generate new hypotheses. [P3, L22]

P4, L17. I do not comprehend this sentence. This sentence seems to make an inference about how other studies used certain concepts. What (I think) the authors mean to state is that they consider these terms interchangeable. If so, this conflicts with P3, L22, in which they seem to suggest that crowdsourcing is not the same as the other concepts. Though the manuscript is much easier to follow after the revision in terms of terminology use, I still find a number of puzzling statements scattered throughout the text.

Thank you for this comment. Our intent is to describe that we examined the methodology of studies and made inclusion determinations regardless of the terminology used. For example, some studies refer to their work as “crowdsourcing” when in fact the current understanding of crowdsourcing is different from what they describe. We believe this is due to the absence of a shared understanding of the terms among the community performing this work. We clarify as follows:

This review does not make inclusion or exclusion determinations based on the terminology used in the studies, because different studies employ terminology differently. For example, investigations may use terms such “wisdom of the crowd,” “crowdsourcing,” or “collective insight,” to describe their work. We included the studies as long as they examine medical decision-making among medical professionals. [P4, L8]

P10, Outcomes. This section has improved, but I still think it could benefit from some more structure. It is a list of discussion points without any structure, making it hard to grasp the essence and navigate this part. Since it is also not foreshadowed what will be discussed in this section (nor why certain aspects are discussed and others not), I find it hard to judge this section. Having headers, or an explanation of why certain aspects will be discussed could help here I think.

Thank you for these remarks. We have added subsection titles to help clarify to which studies we are referring while discussing their outcomes. The headers are: “Group processes” [P10, L20], “Individual processes,” [P11, L11], “Diagnostic accuracy,” [P12, L1]. We believe it should be clear now that we are discussing those studies with similar initial decision tasks.

We note that the nature of our included studies is quite heterogeneous, making it difficult to create section headers that are more specific. By chunking the studies based on their decision task, we hope to give clarity. We include the diagnostic accuracy header due to the fact that we believe information related to diagnostic accuracy will be of interest to readers.
There is also a lot of unnecessary repetition in this paragraph. For example, P11 L37, P11 L57, P12 L30 mentions three time the study in which pairs and individuals were compared. Also other conclusions are repeatedly stated. I would suggest to remove this repetition. I think this also happens because of the general lack of structural clarity in this section.

We agree that there is a fair amount of repetition which makes this section confusing. We removed the final paragraph of this section as we had previously mentioned all those studies, making it redundant.

P11, L37. In reference 26 medical students were used, not physicians.

Thank you for noticing this, we correct the language as such:

Finally, one study compared the diagnostic accuracy of medical student pairs with individual medical students and found that pairs were more accurate in their diagnoses. [P11, L25]

P11, L57. Again, reference 26 did not use physicians. I really urge the authors to accurately report the studies they are citing to avoid confusion. If the aim is to provide a framework in this area, then it does not help that several of the statements are simply incorrect.

We agree and apologize for this oversight. Language corrected throughout:

As compared to novices, expert physicians had better diagnostic accuracy and faster decision times(30), and novice pairs were more accurate than individuals those working alone(27). [P12, L8]

P11, L36 + L41. Diagnostic performance and diagnostic accuracy are used in sentences following each other. Why use different terminology to indicate the same thing?

We thank the reviewer for this note and agree that consistent terminology makes more sense. We have endeavored to correct this across the manuscript.

Finally, one study compared the diagnostic accuracy of medical student pairs with individual medical students and found that pairs were more accurate in their diagnoses. [P11, L25]

P12, L51. Typo.

corrected!

I found it surprising that in the Conclusion the 3 pillars of the framework are not all mentioned. That seems strange as this seems to be one of the major conclusions of the study.

We appreciate this feedback. We acknowledge that we did not provide enough context and have added the following:
In this review we present a novel framework to describe investigations into collective intelligence. Studies examined two distinct forms of the initial decision task in collective intelligence: individual processes that were subsequently aggregated, versus group synthesis in which the diagnostic thinking was initiated in a group setting. The initial decision task is followed by aggregation or synthesis of opinions to generate the collective decision-making output. When a group jointly develops their initial decision, synthesis occurs as part of the initial input, whereas in individual processes, manual or IT methods are required to generate a collective output from the individual inputs that experts contribute. The final collective output can then be routed back to the decision-makers to potentially influence patient care. [P14, L2]