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 (urmimala.sarkar@ucsf.edu)

Version: 1 Date: 30 Apr 2019

Author’s response to reviews:

Please find our Response to Reviewer Comments Letter attached as supplementary information

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
BERKELEY * DAVIS * IRVINE * LOS ANGELES * RIVERSIDE * SAN DIEGO * SAN FRANCISCO
SANTA BARBARA * SANTA CRUZ

UCSF Division of General Internal Medicine
San Francisco General Hospital
BOX 1364, SFGH BLDG 10, WARD 13
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
SAN FRANCISCO, CA. 94143 - 1364
Urmimala.Sarkar@ucsf.edu

April 26, 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
A review on collective intelligence in medical diagnostics is no easy feat given the heterogeneity of the field, the often ill-defined terms, the existence of many related concepts, and the different research communities working on this topic. I think this manuscript can be helpful in clarifying the concept of collective intelligence in medical diagnostics, and provide some structure to the field, though I feel it needs to be revised to achieve this goal. In its current form, I think it is not yet clear enough, especially regarding (i) the main aim of the paper, (ii) the definitions of the terms used (especially collective intelligence is ill defined), (iii) the content of the studies it is summarizing, and (iv) the inclusion criteria used.

Thank you for these comments and will address them specifically below. We have included the Appendix 1 in this submission, which includes details of our search strategy.

We have made significant revisions to the manuscript to clarify our aim. As you note, the field is heterogeneous and often ill defined. Therefore, we sought to focus on collective intelligence as it applies to complex medical decision-making, which we describe below.

Abstract

P3, L4/5. "Collective intelligence..., refers to the insights of groups".

This is vague. I would give a more intuitive description of what collective intelligence entails.

We have revised this statement to clarify, within the confines of the abstract character limit.

Collective intelligence, facilitated by information technology or manual techniques, refers to the collective insight of groups working on a task and has the potential to generate more accurate information or decisions than individuals can make alone. [page 2, line 2]

P3, L10/11. "characterize the current state of research with respect to collective intelligence and medical decision-making".

This goal is not very concrete. Is it not possible to have a more concrete goal in mind with this scoping review?
We clarify this statement to emphasize that our goal is to describe a framework that can unify diverse studies in collective intelligence.

We aim to characterize the current state of research with respect to collective intelligence in medical decision-making and describe a framework for diverse studies in this topic. [page 2, line 5]

P3, L20/21. This sentence seems to be incorrect.’

We revise to clarify the types of cases.

Studies that examined medical decisions such as diagnosis, treatment, and management in the context of a actual or theoretical patient case were included [page 2, line 9]

P3, L32/33. "Initial decision task (group vs. individual)".

Without any explanation, it is unclear what is meant with this. It would be helpful to either use clearer terminology, or explain this.

P3, L36/37. "Availability of collective intelligence to participants".

Again, please clarify. According to the definition given earlier in the abstract, this means: "availability of the insights of groups to participants". But what does this mean?

We understand that given the brief nature of the abstract, it is not yet clear what is meant by the categories presented. We made brief revisions to this text in the abstract, but due to character constraints, we aim to describe the framework in much greater detail within the Results section of our manuscript. In the abstract it is not yet necessary for the reader to fully understand the framework, but it is illuminated in the body of the text beginning in the subsection “application of collective intelligence.” [page 7, line 18]

We present a framework to characterize these diverse studies, and future investigations, based on how they operationalize collective intelligence for medical decision-making: 1) how the initial decision task was completed (group vs. individual), 2) the method of aggregation or synthesis of opinions (information technology vs. manual vs. in-person), and 3) the availability of collective intelligence to participants. [page 2, line 16]

P3, L37 "a better understanding of collective intelligence and its goals may improve medical decision-making". The goal of collective intelligence? What do the authors have in mind with this statement?

We have moved this statement to the discussion and revised to clarify that we are interested in how collective intelligence can be applied to medical decision-making.
…a better understanding of collective intelligence and its applications to medicine may improve medical decision-making. [page 2, line 24]

P3, L42/43. "promising tool". I am not sure how the preceding text in the abstract justify the conclusion that collective intelligence is a promising tool. (I do not disagree with the statement, but I do not see how it logically follows from the rest of the abstract.)

We revise to the following:

Collective intelligence in medical decision-making is gaining popularity as a means of advancing medical decision-making, thereby improving patient outcomes. [page 2, line 21]

P4 When introducing the concept of collective intelligence, a number of classical studies on collective intelligence are not cited. Please add general studies on this topic. (e.g. Krause et al. 2010 TREE Swarm Intelligence in Animals and Humans; Surowiecki 2004 The Wisdom of Crowds, etc).

P4 Generally, the background section is relatively short and could be improved by discussing collective intelligence in a broader fashion. It is currently very narrow in scope.

Thank you, these studies are now included. We have also made some revisions to broaden our description of collective intelligence.

Collective intelligence, in contrast to individual aptitude, is the ability of a group to perform a wide variety of tasks.1-3 This concept can be referred to as “the wisdom of crowds,” and the classic example is Galton’s experiment in asking people with a range of expertise to look at a cow and estimate its weight. He found that the average of all the estimates was closer to the cow’s actual weight than any individual’s response.3 Studies have demonstrated that groups using collective intelligence4 in different cognitive tasks have high performance and can generate more accurate outcomes than the decisions of individuals alone. 4-6 Terms such as “collective intelligence,” “wisdom of the crowds,” and “crowdsourcing,” which are broad in scope, have been utilized to describe group decision-making in fields such as medicine, business, and ecology.3 7 8 [page 3, line 1]

P4, L6/7 "Groups using collective intelligence".

This is ambiguous. What is meant with this? Can there also be groups not using collective intelligence? In general, I think the word collective intelligence is used a bit too broadly throughout the manuscript, and it would help to use this specific word less often, and be more concrete in what it meant in several places in the manuscript when using this word. This, for
example, is also true in the abstract when stating "Availability of collective intelligence to participants", the meaning of which is unclear.

In the context of our review, “groups using collective intelligence” refers to the fact that individuals are attempting to achieve a collective insight rather than working alone. We state in the first line that this is in contrast to individual aptitude, in which an individual is working alone without taking advantage of the insight of others. We believe our use of Galton’s concept as mentioned above helps to clarify this statement.

Studies have demonstrated that groups using collective intelligence4 in different cognitive tasks have high performance and can generate more accurate outcomes than the decisions of individuals alone. 4-6 [page 3, line 5]

P4, L37. "and its implications for expert medical decision-making lack clarity in the literature"

Please give a few concrete examples to clarify what is meant with this rather vague statement.

We have made additions to help clarify our statement. Additionally, we hope the text below will help to clarify our purpose and our definition of “complex medical decision making” which is the basis for the search.

Nevertheless, collective intelligence remains poorly characterized in the medical setting and its implications for medical decision-making lack clarity in the literature. Medical diagnosis and decision-making encompasses a range of complexity and certainty. 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.4,11 In contrast, diagnostic decisions can require synthesis of multiple types of information over time and often do not result in a specific correct and objective “answer.” Currently, little is known about collective intelligence in complex medical decision making, although early results with simulated cases are promising.16 Recent research heralds the application of collective intelligence to radiology or pathology as proof that collective intelligence will improve accuracy across all medical specialties, yet there are no reviews of the application of collective intelligence in a typical diagnostic medical setting. Therefore, we conducted a systematic scoping review to both synthesize and characterize the current state of research on collective intelligence in complex medical decision-making. [page 3, line 20]

P4, L46 "inter-expert collaboration".

Collaboration is usually seen as individuals working directly together (whereas the aggregation of independent decisions is usually not considered collaboration; the authors use a similar distinction on P. 6, L. 12). Therefore, I would not state here that the review focuses on "collaboration" only.
Thank you for this important clarification. Revision:

Due to the relative novelty of the term “collective intelligence” in medicine, our review focuses on studies that describe efforts to make medical decisions through the combination of expert insights with an array of interventions.

Methods

P5, L28/29 "did not search for specific group activities that may constitute collective intelligence". Please clarify what this means. I also do not understand why "specific examples" of collective intelligence were not directly searched for, given that the authors state, "our search strategy aimed to capture any type of collaborative decision-making"

Thank you for clarifying. Our intention was to indicate that although we know that certain activities are collective intelligence (case conferences, etc) we did not include those terms in the search strategy but included them if they came about from our broad search, which was designed to capture such an activity.

Given the incipient nature of this field, there are no prerequisite methodologies for generating collective intelligence in medicine. As such, we did not limit our search to specific methods for generating collective insight, such as case conferences or the use of computational rules (“majority”, “quorum”, and “weighted quorum”) in our search strategy. However, we included this literature if it met our search criteria. [page 4, line 25]

P5, L55. This implies that grey literature does not include research on clinical settings. But is this really true?

Agreed. Clarification:

We did not use grey literature because of our interest in research subjected to peer-review. [page 5, line 14]

P6, L5. Please report the values of the inter-rater reliability score, so it is possible to evaluate the reliability of the search. (Rather than reporting that this was done).

In accordance with good practice for systematic reviews, we performed a double-screening procedure to result in a consensus and had a third reviewer resolve discrepancies. We detail this procedure and cite the guiding principles. We maintain that this practice is in accordance with standard systematic review procedure and are not planning to report a IRR score at this time.
Two reviewers (KR & HCL) independently screened a random sample of 181 studies (10% of the overall total) by title and abstract and collaboratively reviewed screening decisions to ensure inter-rater consensus, in accordance with the current recommended standards for study selection.19 20 [page 6, Line 11]

P6, L14. "we included studies that applied a collective intelligence to any aspect". Applying a 'collective intelligence' sounds ambiguous. Please clarify this sentence.

We have clarified our statement to describe the utilization, rather than application, of collective intelligence.

We included studies in which participants examined real or simulated patient cases and made a judgement either collaboratively or individually. Because the diagnostic process involves complex medical decision-making before and after a diagnosis is made,21 we included studies that utilize collective intelligence in any aspect of the medical decision-making process, including diagnosis, treatment, or management. [page 5, line 19]

P6, L41/42. When reading the section on data extraction, I was surprised to read that the number and type of raters were not extracted. But then in Table 1, I learn this has been done. Please clarify.

We apologize for the confusion. This was done and we have corrected:

A standardized form was created to extract data in the following areas: 1) study setting, 2) study type and methodology, 3) characteristics of the intervention (e.g. intervention type, participant characteristics, and outcome measures used), and 4) results on primary outcomes as well as accuracy. [page 6, line 17]

In the methods section it would be helpful to have a better overview of the inclusion/exclusion criteria. Perhaps even have a list with all the criteria. Criteria are now often not clearly defined (e.g., P6, L35 "consideration of complex, non-binary data"; but how exactly was the decision made whether a study was "complex" or not?). Moreover, new criteria are being introduced in the results section (e.g., P7, L. 16 "no studies were included that included non-medical experts"). It is thus not easy to follow the literature search process.

We have compiled all of our inclusion and exclusion criteria into a new sub-section and ensured that all the criteria are mentioned.

Studies were included if they aggregated the medical opinions of at least two medical experts (physicians or trainees), with respect to specific clinical cases. We included studies in which participants examined real or simulated patient cases and made a judgement either
collaboratively or individually. Because the diagnostic process involves complex medical decision-making before and after a diagnosis is made, we included studies that utilize collective intelligence in any aspect of the medical decision-making process, including diagnosis, treatment, or management. Included studies make a judgment based on a specific, individual-level, actual or simulated patient case, rather than examining clinical syndromes in general (for example, expert opinions on hypertension guidelines). We did not limit our search to studies that detailed an analysis of the accuracy of collective intelligence. Due to our interest in characterizing the state of the literature surrounding collective intelligence, we did not limit the primary outcomes under investigation by included studies. Collective intelligence may be generated by a group of experts who make a collective decision, or may be the result of aggregation of the insights of multiple individuals. Therefore, even though these two processes differ in their methods and outcomes, both constitute collective intelligence in the current literature and both are included in this review.

We excluded studies in non-English languages, with no full-text and those that did not include physicians or medical students. Studies that were secondary analyses of previously reported data were excluded due to our interest in primary data. Studies were also excluded if they assessed the opinions of radiologists and pathologists, or examined radiological scans or pathological specimens. Our goal was to uncover the utility of collective intelligence in diagnosis and decision-making, rather than in binary decisions such as identification or absence of a finding in radiology or pathology. [page 5, line 17]

Results

P8, L32. "into a single collective intelligence".

I think it would be better to simply talk about a collective decision or a collective outcome. I would not use the phrasing "collective intelligence" here.

We update our language to “collective output” where appropriate.

Next, collective intelligence for both individual and group processes requires an aggregation or synthesis of the insights of individuals or a group into a final decision-making output, which we refer to as the collective output. [page 8, line 18]

We also present additional clarification in our background to help the reader better understand our intention when describing collective intelligence in this review:

Due to the relative novelty of the term “collective intelligence” in medicine, our review focuses on studies that describe efforts to make medical decisions through the combination of expert
insights with an array of interventions. This review seeks to inform future studies that aggregate the insights of multiple individuals to improve patient care and safety. This review includes investigations that describe their work as “wisdom of the crowd,” “crowdsourcing,” or “collective insight,” as the same phenomenon. For the purposes of this analysis, we utilize the term “collective intelligence” to describe interventions that utilize group insight to accomplish a task, with the understanding that such interventions may vary in methods and outcomes. Due to the diverse nature of the research in this budding field, our review seeks to inform future studies by describing a framework to which future investigations may be applied. [page 4, line 5]

P8, L44 "expert discussion". This sounds very similar as the "open discussion among group members as discussed in the section on "initial decision task". Perhaps clarify that (if I understand this correctly) how these two discussion types are different (or happening in different phases of the process).

This sentence simply serves to describe how IT can facilitate aggregation in both individual and group processes. You are correct that group processes in the initial decision task are the same as expert discussion. The difference here is that we are describing the method of aggregation of these initial decision tasks. IT can aggregate both individual and group processes. We hope this is clearer if you reference Figure 2.

P8, L51 "In four of six individual-based studies, technology facilitated the aggregation of individual opinions into a collective intelligence 17,22,26,27"

When looking at some of these studies, I am not certain if this statement is correct. Take reference 27. This study compares the accuracy of single physicians versus computers. I do not see how this is related to collective intelligence; nor how this is an example of studies in which technology facilitated the aggregation of individual opinions into a collective intelligence.

Thank you for this clarification. You are correct there is a comparison between physicians and computers, however, the distinction is that it is a group of physicians (ie, collective intelligence) being compared to computers. The use of IT does not reference the comparison of the physicians to the computers, but rather the method that the multiple insights of the physicians were aggregated to create collective intelligence. We have gone over our Tables to ensure that it is clear why each of the studies were included. This is a very heterogeneous field as you note, thus we had to be very careful to ensure that each of the studies met our criteria. We feel confident that every study included constitutes collective intelligence.

P9, L35 "Availability of collective intelligence output"
Please explain what is meant with this? I think it would help to reformulate this, because it is unclear what it means to have collective intelligence 'available'. Perhaps collective outcome? Or collective decision?

We have revised this paragraph to clarify what we mean with this portion of the framework and why it bears importance. We hope you will find these revisions to your liking.

The final collective output has the potential to inform patient care decisions when it is available to participating individuals for review. For example, in group processes of collective intelligence (8 studies), the contributing group members are of the final results of consensus as they have participated in the consensus-making. On the other hand, through individual processes, the contributors may not be aware of the final decision-making output. In the cases, the study team or IT platform must decide whether to share the final consensus with the individuals. When contributors are aware of the final consensus, the collective intelligence results have the potential to impact future patient care. Our results show that the collective intelligence was available to participants in one of the six studies using individual processes, and in the study examining both group and individual processes. For example, Douzgou et al. describe a procedure a final Expert Case Report, the product of collective intelligence, which generates information about diagnosis or treatment, is routed back to the soliciting clinician. The availability of collective intelligence to clinicians will have important implications in whether or not collective intelligence can impact and improve patient care in real-time. [page 9, line 21]

P10, L3 "but did not necessarily investigate the accuracy of collective intelligence or its benefit over individual decision-making."

It would be helpful to already state in the introduction that this scoping review is not only focusing on collective accuracy as compared to individual accuracy. In almost all studies that talk about collective intelligence, accuracy is a central concept, so I was surprised that this was not a hard selection criterion. I would thus already in the introduction sketch this broad approach to clarify the scope better to the reader.

We thank you for this helpful comment, and have included this in our Inclusion and Exclusion criteria subsection.

We did not limit our search to studies that detailed an analysis of the accuracy of collective intelligence. [page 5, line 24]
P10, L8. "For example, information needs among participants at a case conference, a traditional means of generating collective intelligence, were resolved through group processes of collective intelligence".

Were resolved through group processes of collective intelligence. Please clarify what this means, as this is hard to comprehend.

Thank you for this clarification. We have revised:

For example, participants at a case conference were able to resolve their questions about a patient’s case by collaborating in a group process. Such findings lend credence to the fact that generating collective insight among groups may improve patient care. [page 10, line 17]

P10, L23 "Finally, one study demonstrated that a collaborative technology platform could enhance a group's collective intelligence".

Please explain what this means. Better diagnostic accuracy? Again, the word collective intelligence is used so broadly throughout the text that it is really hard to understand its meaning.

Due to the fact that our review has a limited focus on diagnostic accuracy, we have updated this language to reflect the goal of this cited study which is enhanced cognitive skills.

Finally, one study demonstrated that a collaborative technology platform could enhance a group’s cognitive skills. While these studies may seem overly simplistic compared to the sophistication of recent automated methods or collective intelligence platforms, they are practical applications of “wisdom of the crowds” to complex medical decision-making and are relatable applications of collective intelligence to usual practice. [page 10, line 24]

P10, L40/42. When looking at references 20 and 22 I fail to see how they are connected to the concept of collective intelligence. These studies investigate how different groups of raters have different use of ranking. They do not (as far as I can see) relate to anything related to group decision making. Also when looking at the statements in the text I fail to see how this is related to collective intelligence: "finding that specialists varied systematically in the diagnoses they reached, and that experts were more accurate than medical students”. I get a bit worried about the suitability of these studies to be included in a review about collective intelligence in medical diagnostics.

We have updated this language to better characterize the implications of this finding. Regarding the suitability of studies, we will address your comment in full at the closing of this response letter:

Studies that utilized individual processes of collective intelligence also compared the diagnostic ability of different physicians groups27 finding that specialists varied systematically in the
diagnoses they reached, and that experts were more accurate than medical students. 29 These findings imply that while collective intelligence may be a useful tool in diagnosis, it is important to consider the level of expertise and specialty of participants generating the collective intelligence. [page 11, line 12]

P10, L58. "As compared to novices, expert physicians had better diagnostic accuracy and faster decision times."

Again, how is this statement relevant to collective intelligence?

Thank you for this clarification. Revision:

As compared to novices, expert physicians had better diagnostic accuracy and faster decision times,29 and pairs were more accurate than individuals working alone.26 This finding has implications in future investigations, which may choose to use the combined insights of experts rather than novices, such as students, as participants in collective intelligence. [page 11, line 22]

Generally, in the results section, I found several claims that, in my opinion, are not backed up by the specific literature references provided. Moreover, when looking at several of these references I do not see an obvious link to collective intelligence. It is possible that the authors have a radically different view of what collective intelligence entails, but then it would be helpful if this view is presented early on so that it becomes clear which aspects are being considered a part of collective intelligence. I think the results section needs to be clarified, and the claims need to follow the cited studies more rigorously and, at places, more concretely.

In order to address this important comment, we have added the following text to our background:

Due to the relative novelty of the term “collective intelligence” in medicine, our review focuses on studies that describe efforts to make medical decisions through the combination of expert insights with an array of interventions. This review seeks to inform future studies that aggregate the insights of multiple individuals to improve patient care and safety. This review includes investigations that describe their work as “wisdom of the crowd,” “crowdsourcing,” or “collective insight,” as the same phenomenon. For the purposes of this analysis, we utilize the term “collective intelligence” to describe interventions that utilize group insight to accomplish a task, with the understanding that such interventions may vary in methods and outcomes. Due to the diverse nature of the research in this budding field, our review seeks to inform future studies by describing a framework to which future investigations may be applied. [page 4, line 5]

We also add this text to our methods:

Given the incipient nature of this field, there are no prerequisite methodologies for generating collective intelligence in medicine. As such, we did not limit our search to specific methods for
generating collective insight, such as case conferences or the use of computational rules (“majority”, “quorum”, and “weighted quorum”) in our search strategy. However, we included this literature if it met our search criteria [page 4, line 25]

Regarding reviewer’s concern about included studies:

We have clarified the final column, “relevance to collective intelligence” in table 1 of our manuscript. We agree that these studies are heterogeneous. However, for this reason we included a column that describes in brief why they met inclusion criteria. In each of these studies, a group of medical experts contributes to a collective output related to a case. For some studies, the reviewer mentions that “ranking” does not appear to constitute collective intelligence. However, in that study in particular, the experts are making complex decisions related to medical care and fit our definition of complex medical decision-making (prognosis, diagnosis, or treatment). We hope that by reviewing our framework diagram and the tables, our intentions will become clearer.

Table 1. Collective intelligence study design

<table>
<thead>
<tr>
<th>Study author, year</th>
<th>Description of experts</th>
<th>Study design</th>
<th>Relevance to collective intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gagliardi, 2007</td>
<td>20 general surgeons, 2 pathologists, 1 medical oncologist, 1 radiation oncologist</td>
<td>Qualitative observational study to explore the role of multidisciplinary cancer conferences in practice</td>
<td>Describe collective output generated in multidisciplinary cancer conferences</td>
</tr>
<tr>
<td>Douzgou, 2016</td>
<td>Physicians with patients with malformation syndromes</td>
<td>Descriptive study of a consultation tool which generates collective insight</td>
<td>Assess a collective intelligence tool</td>
</tr>
<tr>
<td>Sternberg, 2017</td>
<td>International colleagues with urologic expertise</td>
<td>Use Twitter as a potential collective intelligence tool</td>
<td>Describe social media as a collective intelligence tool</td>
</tr>
<tr>
<td>Sims, 2014</td>
<td>Clinicians affiliated with academic departments: 28 from pediatrics, 27 from neurology, 10 from internal medicine, 4 from psychiatric, 11 from pediatric neurology, 5 others</td>
<td>Descriptive study of a clinical consultation system which generates collective insight and qualitative evaluation of the tool</td>
<td>Describe a collective intelligence tool</td>
</tr>
<tr>
<td>Nault, 2009</td>
<td>5 spinal deformity surgeons</td>
<td>Feasibility study of a surgical decision-making tool as compared to a group of experienced surgeons</td>
<td>Compare collective intelligence generated by experts with a technology tool</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Participants</td>
<td>Study Type</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Alby, 2015</td>
<td>1 oncologist and others from hematology, anesthesiology, surgery, and nephrology</td>
<td>Qualitative observational study of conversations about cancer cases between the chief oncologist and other physicians at a hospital</td>
<td>Characterize collective intelligence generated in usual practice</td>
</tr>
<tr>
<td>Kattan, 2013</td>
<td>24 urologists and oncologists</td>
<td>Analysis of physician group accuracy as compared to a nomogram</td>
<td>Compare collective intelligence generated by experts with a technology tool</td>
</tr>
<tr>
<td>Kunina-Habenicht, 2015</td>
<td>283 medical students, 20 expert physicians</td>
<td>Descriptive study of the development of a computerized test to assess diagnostic accuracy; results were compared among medical students and expert physicians</td>
<td>Compare computer-generated collective intelligence of experienced physicians to medical students</td>
</tr>
<tr>
<td>Lajoie, 2012</td>
<td>14 third-year medical students</td>
<td>Qualitative observational study of team discussions with or without a technology tool to aid collaboration</td>
<td>Optimize metacognitive activities in collective intelligence with a technology tool</td>
</tr>
<tr>
<td>Kalf, 1996</td>
<td>21 geriatricians, 21 geriatric-psychiatrists, 21 internists</td>
<td>Analysis of diagnoses generated by different specialties</td>
<td>Compare collective intelligence among different specialists</td>
</tr>
<tr>
<td>Larson, 1996</td>
<td>24 first-year interns, 24 residents, 24 medical students</td>
<td>Qualitative observational study of team diagnostic discussions when teams are exposed to different case information</td>
<td>Characterize collective intelligence generated when groups have different amounts of information about a case</td>
</tr>
<tr>
<td>Christensen, 2000</td>
<td>24 first year interns, 24 residents, 24 medical students</td>
<td>Qualitative observational study of team diagnostic discussions when given different amounts of shared and unshared information</td>
<td>Characterize collective intelligence generated when groups have different amounts of information about a case</td>
</tr>
<tr>
<td>Larson, 1998</td>
<td>48 interns and 24 third-year medical students</td>
<td>Qualitative observational study of team diagnostic discussions when teams are exposed to different</td>
<td>Characterize collective intelligence generated when groups have different amounts of</td>
</tr>
</tbody>
</table>
case information and given instructions about sharing information about a case

Semigran, 2016 234 physicians, including fellows and residents
Analysis of a collective intelligence tool as compared to the accuracy of symptom checker websites
Compare a collective intelligence tool to online symptom checkers

Hautz, 2015 88 medical students
Analysis of diagnostic accuracy when participants worked in pairs or individually
Compare collective intelligence of pairs to individual aptitude

Discussion
P11, L45 "efficiently resolve uncertainty". What is mean with resolving uncertainty? Did this also resulted in increased decision accuracy? Please clarify such statements. This is one example of ambiguous writing present in several parts of the manuscript.

Thank you. The term efficiency is unclear and unnecessary. Revision:

In particular, studies included in our review reveal that collective intelligence IT platforms can allow physicians to resolve uncertainty in diagnosis and treatment decisions,24,35 can be more accurate than online symptom checkers,34 and can be an facilitate group processes of collective intelligence36 to improve metacognitive activities and collective insight.30 [page 12, line 19]

Potential literature to include (depending on exact inclusion criteria):

Hernández-Chan et al. 2012.
Knowledge Acquisition for Medical Diagnosis Using Collective Intelligence.
Journal of Medical Systems. https://doi.org/10.1007/s10916-012-9886-3

Collective intelligence in medical diagnosis systems: A case study.


Medical Decision Making. https://doi.org/10.1177/0272989X17696998

Kurvers et al. 2018. Combining independent decisions increases diagnostic accuracy of reading lumbosacral radiographs and magnetic resonance imaging.

PLOS ONE. https://doi.org/10.1371/journal.pone.0194128

Each of these articles has been screened as part of the search and was not included as part of the search criteria. We hope our revisions to inclusion/exclusion criteria help clarify this.

Comments related to grammar and structure that have been revised:

P4, L6/7 why is reference '2' not added to references 3/4 at the end of the sentence?
P4, L9/10 Groups can also be comprised of a combination of experts and novices.
P5, P6, L54. I would not use the heading "results" given that six lines later there is the main section "Results".

L42/43. Broken sentence I think.
P5, L50. The Appendix was not available to me for reviewing.
P11, L48 Broken sentence.

Reviewer 2 (Reviewer 2): PEER REVIEWER ASSESSMENTS:

OBJECTIVE - Full research articles: is there a clear objective that addresses a testable research question(s) (brief or other article types: is there a clear objective)?

Yes - there is a clear objective

DESIGN - Is the current approach (including controls and analysis protocols) appropriate for the objective?
Yes - the approach is appropriate

EXECUTION - Are the experiments and analyses performed with technical rigor to allow confidence in the results?
Yes - experiments and analyses were performed appropriately

Statistics - Is the use of statistics in the manuscript appropriate?
N/A - there are no statistics in this study

INTERPRETATION - Is the current interpretation/discussion of the results reasonable and not overstated?
Yes - the author's interpretation is reasonable

OVERALL MANUSCRIPT POTENTIAL - Is the current version of this work technically sound? If not, can revisions be made to make the work technically sound?
Probably - with minor revisions

PEER REVIEWER COMMENTS:
GENERAL COMMENTS: The study is very interesting. It has been described with a lot of information and several conclusions have been extracted from it. I think that the authors have done an incredible job in terms of conclusion extraction since they have reached several and interesting conclusions.

REQUESTED REVISIONS:
In my opinion, most of the analysis performed is ok. It is true that probably, statistical analysis does not make sense due to the low number of final articles analyzed, so, it is impossible to, for example, analyze significance in some of the results provided. However, I think that descriptive statistics about the conclusions drawn from the analysis might help to have even a better understanding of the conclusions.
Thank you for this comment. We agree that, in addition to the low number of articles, the heterogeneity of the articles makes it futile to attempt statistical analyses. We describe the proportions of articles, which reflect our outcomes in detail, and hope this will suffice as descriptive.

ADDITIONAL REQUESTS/SUGGESTIONS:

In study selection, the authors mentioned that "Reviewers double-screened a sample of 10 percent of the total number of studies and compared results to ensure inter-rate reliability". I've not seen more information about this in the manuscript. Can you please provide more information?

Thank you for this comment, we have revised to the following, and included citations to reflect that this is an appropriate methodology:

Two reviewers (KR & HCL) independently screened a random sample of 181 studies (10% of the overall total) by title and abstract and collaboratively reviewed screening decisions to ensure inter-rater consensus, in accordance with the current recommended standards for study selection.20 21 Two reviewers (KR & HCL) completed final screening for each article to determine inclusion and presented discrepancies to US for the final determination. [page 6. Line 11]

In Results, authors state (second line) "no studies were included that included non-medical experts in the collective intelligence". You mean that you didn't include studies that used non-medical experts during the collective intelligence process decision-making process, right?

This is correct. Revision for clarity:

All 15 studies included medical experts contributing to the collective intelligence, including medical students, interns, residents, fellows, and attending physicians (Table 1), in accordance with our inclusion criteria. We did not include studies that employed laypeople’s input on medical decision-making. [page 7, line 5]