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)	How effective is teamwork really? The relationship between teamwork and performance in healthcare teams: A systematic review and meta-analysis
AUTHORS	Schmutz, Jan B.; Meier, Laurenz; Manser, Tanja

VERSION 1 – REVIEW

REVIEWER	Maastricht University, CAPHRI Care and Public Health Research Institute Maastricht, the Netherlands
REVIEW RETURNED	21-Dec-2018

GENERAL COMMENTS	The question "how effective is teamwork really?" is a very relevant question for health care teams. However, the paper left several questions to me about the conceptualization, methods and outcomes of the study. The focus of my feedback is on the rationale for the study objective, the conceptualizations, and the inclusion criteria. These aspects needs more clarity for a good understanding of the outcomes of the statistical analysis. Study objective: You want to investigate the relationship between teamwork and clinical performance and potential moderating variables of this relationship. It is about uniprofessional and interprofessional care teams. What is your definition of a team - when could one speak about 'team members'? It was surprising that only teams in acute hospital care were included in the analysis (emergency, anesthesia, surgery, intensive care). What about teams in the extramural / chronic care? What about the differences between acute care and chronic care / teams in primary and secondary care? I miss exclusion criteria in your search term. I also miss relevant literature about interprofessional teamwork like the systematic reviews of S. Reeves, and the recent meta-ethnographic review of O Petit dit Dariel about interprofessional teamwork in hospitals. Are their findings in line with your study? What kind of studies did they include in comparison to your study? Teamwork effectiveness is an even more complex and broad concept than teamwork. A very helpful framework is the Integrated Team Effectiveness Model (ITEM) as described by Lemieux-Charles (2006) about the complex relationships between between team context, structure, processes, and outcomes. Her conclusion is that context variables and collaboration, conflict resolution, participation, and cohesion are very crucial for team effectiveness. Based on this knowledge, my specific questions are:

	The IPO framework seems to me very simplistic (input process output). Why have you chosen for this framework? How do you exactly define 'performance'? Do you also differentiate between objective and subjective process/outcome measures? What about the contextual and methodological factors that might moderate the effectiveness of teamwork (page 6, line 27). Did you define these moderators before you started the review or during the review process? Why these moderators? Could you explain the search terms. Why 'decision making' and 'leadership' and nothing about effectiveness or performance or process or outcomes? Why did you choose for a meta-analytical study approach? Would a scoping review give more insight into the aspects of teamwork that have a positive impact on team performance? Most studies included in the review are observational studies. The aim of the papers, their design, rating scales, and settings are very different. Are they comparable and are the effects quantifiable? Since the information about the studies are very limited (design, rating scales), I'm not sure how to interpret the mean correlation and the tests regarding the moderators. "We provide strong evidence that teamwork contributes considerably towards quality of care – or in other words, poor teamwork significantly increases the risk for unsafe care and even patient harm", discussion page 23, l19 Your findings suggest that better teamwork improves the quality of care, but what is your basis for the statement regarding unsafety and patient harm? Minor issues Page 5, 17: I would say 'generate evidence' instead of strong evidence Page 6, 55: I do not understand the assumption Page 7, 28: I do not understand the expectation Page 7, 40: teamwork is more important or ' effective' teamwork is more important? Page 9, 26: I do not see clinical performance in the search term, page 39

REVIEWER	Philip Chilibeck University of Saskatchewan, Canada
REVIEW RETURNED	21-Jan-2019

GENERAL COMMENTS	I was asked to provide a statistical review of the manuscript. I have included in the review some comments about the systematic review and other minor suggestions. Page 2, abstract, lines 25-26: Change "two individual" to "two individuals" Page 6, line 10: Change "team members experience" to "team members' experience" Page 9, line 35: It is stated as part of the search strategy: "a manual forward search to identify studies that cite the studies we

	included in our meta-analysis" How were these studies identified (i.e. what database was used to identify these studies; i.e. was it Web of Science or another database?). Page 9-10: The description of inclusion criteria for the included studies seems very brief. Please ensure you follow the PRISMA guidelines in detail here. Is it relevant to describe the "PICOS" here (i.e. Population, intervention, comparator, outcomes, study type)? In the statistics section (pages 19-20): Please provide reference to how you classified the effect sizes (i.e. as "small", "medium", "large", etc.) with an indication of the cut-offs used. Please indicate the statistical test you used to assess whether heterogeneity was present. Did you do any assessments for study quality (i.e. bias)? I know there are a number of tools used to assess quality of randomized controlled trials (i.e. Cochrane tool, Jadad score) – are there any tools available to assess quality of the types of studies you included in your meta-analysis? Page 20, lines 29-33: I think this description of how outliers were identified should be moved to the statistics section, rather than being described in the results section. Figure 1: In the boxes of the flowchart, clarify where titles and abstracts were screened.

REVIEWER	Ashley Hughes University of Illinois at Chicago, Chicago, IL, USAd
REVIEW RETURNED	21-Feb-2019

GENERAL COMMENTS	This review provides a much needed synthesis on the quantitative relationship between teamwork and clinical performance outcomes. I applaud the authors for the clarity of the writing, methodology employed, and overall flow of the paper. Performing this type of meta-analytic integration which spans multiple fields is no small feat. The paper is well written and provides and excellent translation of the literature on team science from multidisciplinary journals to healthcare. I have a few areas for improving the manuscript prior to its publication in BMJ open: My primary comments have to do with the methodology used for the study. First, in my understanding of the literature on teamwork in healthcare, these articles tend to be published in a variety of outlets, as the authors mention in the introduction (Lines 28-30). Why is pubmed the only database utilized for identifying articles? Small k is an issue in investigating some of your interesting moderator analyses so expanding the search could prove advantageous.

	Second, I appreciate the inclusion of forest plots and I2 statistics for heterogeneity. However, I do not understand the rationale for selecting Fischer's Z scores used to create composites rather than use of averages or other composite creating techniques (Nunnally, 1978) which are used in other team metas (DeChurch & Mesmer-Magnus, 2010)? Please clarify the advantages or rationale motivating this approach. References DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-analysis. Journal of Applied Psychology, 95(1), 32-53. Nunnally, J. C. (1978). Psychometric theory. New York, NY: McGraw- Hill. Third, clinical context is surprisingly mentioned nowhere within the approach. I agree with the authors' conclusions that the nature of the team task interdependence would vary based on patient acuity, there are a variety of other factors that would be different based on clinical context (types of tasks, patients, and associated staffing structure). Is it possible to explore this further or otherwise provide some guidance as to what this assessment might look like? I'm surprised by this as you are submitting for publication in a medical journal rather than a business or psychology outlet without much mention of exactly what constitutes clinical performance or the clinical context. Fourth, out of pure curiosity, the authors mention F-tests as part of the statistics encountered in primary studies. However, I'm guessing that the authors' choice in using a correlational meta-analysis was due to lack of experimental or quasi experimental designs. Further, the authors describe the rationale for their meta-analysis as stemming from a need to establish a direct relationship between teamwork and clinical performance. Please clarify. Also, I appreciate that levels of analysis for coding were taken into consideration and that the authors chose team level as the level of analysis. This is appropriate and makes findings that much more relevant and compelling. Interesting dilemma about the reliability measures employed. I empathize with the lack of Cronbach's alphas reported; however, a potential source of moderation could be measurement criterion (i.e., observation versus survey). Real or simulated patient- how were patient actors characterized? Arguably, this may be different from a human patient simulator, particularly in regards to ability to demonstrate teamwork. Also, we are talking about teamwork demonstrated within the care team? Is the patient included? Why would having a real patient matter in the context of team process within the clinical care team? Please elaborate. Double coding… why only 25%? Correlations are low. I don't see much discussion on why this could be or connection to what this would mean in context… Bare bones meta-analysis will produce lower than typical correlations with higher than usual standard deviations. To a clinical audience, this correlation is going to seem LOW. Why does this correlation matter? How does this translate? Couch your findings in the context of what it means to improve quality. What it means to patients and what problems may be addressed through improved team process. Connect to the teams literature as well as the clinical care. References

	Arthur, W., Jr., Bennett, W., & Huffcutt, A. I. (2001). Conducting meta-analysis using SAS. London, UK: Psychology Press. Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis:Correcting error and bias in research. Thousand Oaks, CA: Sage. This is just a preference to ease interpretation of your results- you report 95% confidence intervals for the main correlational finding. Unfortunately, at first glance, the range (lower to upper) looks like a negative sign. I recommend choosing a different format as allowed by the journal. Rather than stating an upfront limitation of a possible file drawer effect, let's test the likelihood of it. Comprehensive meta-analysis as a program can assist with this- Hunter and Schmidt (2004) and Duvall and Tweedie (2000) offer some guidance on this as well. There are other ways too of testing for presence of file drawer effect or publication bias. References Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Introduction to meta-analysis. West Sussex, UK: Wiley. http://dx.doi.org/10.1002/9780470743386 Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics, 56, 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research. Thousand Oaks, CA: Sage. Model and moderation clarification- you mention testing models. I am familiar with random effects models as well as model-based methods in meta-analysis (Hunter & Schmidt, 2004). However, I do not see coding or analytic procedures for model-based testing (see Cheung, 2008) as I am used to seeing it. Please clarify the approach and why it was chosen. The same goes for testing the presence of a significant moderating effect. Currently, the approach differs from those cited as motivating the current work. Why was the current method for testing moderators chosen over using an approach such as Whitener non-overlapping confidence intervals (Whitener, 1990) or Zou's (2007) confidence interval significance test(s) References Cheung, M. W. L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. Psychological Methods, 13(3), 182-202. Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research. Thousand Oaks, CA: Sage. Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. Journal of Applied Psychology,75, 315–321 Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. Psychological Methods, 12, 399–413. http://dx.doi.org/10.1037/1082-989X.12.4.399 Namely, in reviewing the metafor package for clarification on omnibus techniques, these appear to test linear coefficients in models. While I am unclear as to how the coding would allow the model to be constrained in terms of accounting for interrelatedness amongst moderators, I am definitely unclear on the use of omnibus

	significance tests for categorical moderators in meta-analysis that claims to use a model-based approach.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:Reviewer: 1Reviewer Name: Anneke van Dijk PhDInstitution and Country: Maastricht University, CAPHRI Care and Public Health Research Institute, Maastricht, the NetherlandsPlease state any competing interests or state 'None declared': None declaredPlease leave your comments for the authors belowThe question "how effective is teamwork really?" is a very relevant question for health care teams. However, the paper left several questions to me about the conceptualization, methods and outcomes of the study. The focus of my feedback is on the rationale for the study objective, the conceptualizations, and the inclusion criteria. These aspects needs more clarity for a good understanding of the outcomes of the statistical analysis.

R1: Study objective: You want to investigate the relationship between teamwork and clinical performance and potential moderating variables of this relationship.It is about uniprofessional and interprofessional care teams. What is your definition of a team - when could one speak about 'team members'?

Response: Thank you for this feedback. We would like to highlight here that the professional composition of teams is only one potential moderator. Our point is that healthcare is mostly interprofessional especially in acute care settings. However, some studies or team trainings still use uniprofessional teams, which often does not represent the reality. As we note on p. 7 diverse educational paths may lead to different values, beliefs, attitudes and behaviors therefore making collaboration more difficult. Therefore, we think it is important to add the professional composition as a potential moderator.
Based on your comment we now also added a paragraph to the introduction where we provide a clear definition of what a team is. We did not provide an explicit definition of "team members" since a) the individual is not the focus of the study and b) we think that this should now be addressed by our definition of "team".

R1: It was surprising that only teams in acute hospital care were included in the analysis (emergency, anesthesia, surgery, intensive care). What about teams in the extramural / chronic care? What about the differences between acute care and chronic care / teams in primary and secondary care? I miss exclusion criteria in your search term.

Response: Thank you for this feedback. We did indeed not mention the criteria that we excluded long term care. We added the sentence to the "inclusion criteria" section: "We excluded articles investigating long term care since the dynamics of teamwork over a longer period of time are different." on p. 11. We hope that this makes clear that teamwork over longer period of time is difficult to compare with more acute settings, where the vast majority of the studies about teamwork was conducted.

R1: I also miss relevant literature about interprofessional teamwork like the systematic reviews of S. Reeves, and the recent meta-ethnographic review of O Petit dit Dariel about interprofessional

teamwork in hospitals. Are their findings in line with your study? What kind of studies did they include in comparison to your study?

Response: Thank you for this suggestion. We had a close look at the suggested literature and hope we found the right papers. The review of S. Reeves about interprofessional education1 was less relevant for our meta-analysis since they did not measure teamwork and we did not look at educational concepts or the effect thereof. The focus of the review was on interprofessional education interventions. In our meta-analysis we were interested in the direct relationship between teamwork and performance. As we state in the methods section we included papers only that measured at least one teamwork process variable. We did not include interventions that do not directly measure teamwork and establish a relationship between a teamwork variable and performance. Teamwork interventions are already the focus of other meta-analysis like Hughes et al. 2016 and others.
After reading the review of Petit dit Dariel & Christofalo (2018) it became clear why we did not include it in our study. In their method section, they mention "…articles were included if they examined interprofessional teamwork within a hospital using a qualitative methodology." Since in our meta-analysis we are interested in quantifying the relationship between teamwork and performance we did not include any qualitative studies. We also thought about comparing the results of our meta-analysis with the review of Petit dit Dariel & Christofalo. However, after a closer look it made less sense to compare the two reviews. Petit dit Dariel & Christofalo state their main results are "Interprofessional teamwork …found to be influenced by systems perpetuating power imbalances, organizational practices that interfered with interprofessional interactions, representations of teamwork and leadership". In their review, they focused on antecedents of teamwork and not the outcomes of teamwork. Therefore, it makes it difficult to compare the two studies and we chose not to include the review into our paper.

1Reeves, S., Zwarenstein, M., Goldman, J., Barr, H., Freeth, D., Hammick, M., & Koppel, I. (2008). Interprofessional education: effects on professional practice and health care outcomes. Cochrane Database of Systematic Reviewsl

R1: Teamwork effectiveness is an even more complex and broad concept than teamwork. A very helpful framework is the Integrated Team Effectiveness Model (ITEM) as described by Lemieux-Charles (2006) about the complex relationships between between team context, structure, processes, and outcomes. Her conclusion is that context variables and collaboration, conflict resolution, participation, and cohesion are very crucial for team effectiveness. Based on this knowledge, my specific questions are: The IPO framework seems to me very simplistic (input process output). Why have you chosen for this framework?

Response: Thank you for this feedback. The question about the chosen framework is very relevant. There is an ongoing debate in the organizational psychology literature about various teamwork/team effectiveness models. The big critique of the IPO model is that it does not take into account the dynamics of team processes2. However, to date it is still the most influential and most used framework to describe team processes. Especially for our purposes—to define and describe teamwork processes—it is in our opinion an adequate model. Since we are not concerned with team dynamics in our meta-analysis the simplicity of the model helps the reader to understand our conceptualization of teamwork.
The ITEM is a very helpful framework for team effectiveness, thank you for this suggestion. However, also this model is an extension of a IPO model. Figure 1 on page 2673 describes task design (Task type, task features, team composition) as inputs influencing processes which in turn influences objective and subjective outcomes. In addition, Lemieux-Charles describes team-psycho-social traits as potential influencing factors of team processes. Also, the author adds another layer by including organizational context to the model. Since we are interested in the relationship Process-Outcome only we think it would confuse the audience to describe a more complex model.

Nevertheless, based on your comment we added two sentences on page 6 acknowledging the simplicity of the model and also referencing to more complex models in the literature including Lemieux-Charles model.

2Kozlowski, S. W. J. (2015). Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. Organizational Psychology Review, 5(4), 270–299. http://doi.org/10.1177/2041386614533586

3Lemieux-Charles, L. (2006). What Do We Know about Health Care Team Effectiveness? A Review of the Literature. Medical Care Research and Review, 63(3), 263–300. http://doi.org/10.1177/1077558706287003

R1: How do you exactly define 'performance'? Do you also differentiate between objective and subjective process/outcome measures?

Response: In the manuscript, we define performance multiple times. First on page 6 we state: “Team performance is often described in terms of inputs, processes and outputs (IPO).[21,24-26] Outputs like quality of care, errors or performance are influenced by team related processes (i.e. teamwork) like communication, coordination or decision making…”. On page 8 we distinguish between process and outcome performance and refer to the relevant performance literature: “The literature usually differentiates between process- and outcome-related aspects of performance.[33,34] Process performance measures are action-related aspects and refer to adequate behaviour during procedures (e.g. adhering to guidelines), making them easier to assess. Outcome performance measures (e.g. infection rates after operations) follow team actions, with assessment occurring later than process measures.
We did not, respectively were not able to differentiate between objective and subjective process or outcome performance. Outcome performance seem to be more objective in general since it includes patient outcomes (e.g. infection rate, mortality, morbidity) that can be objectively observed and counted. Process performance measure mainly include published checklist based rating systems. Here it is possible to argue about the objectivity of the measure. Unfortunately, most papers did not provide any quality indicators for the process performance measurement.
We also added an additional paragraph about clinical performance to the discussion where we discuss the different forms of performance measures depending on the context. With that we hope we could adequately address your feedback.

R1: What about the contextual and methodological factors that might moderate the effectiveness of teamwork (page 6, line 27). Did you define these moderators before you started the review or during the review process? Why these moderators?

Response: Yes, we did define these factors before we started the review. We chose the most relevant variables based on the literature and our experience. Unfortunately, the word limit of BMJ Open does not allow for a thorough theoretical rational for every moderator like we would do it in a psychology or management outlet with no word limits. Also, we had to limit ourselves to variables that are reported in the selected papers. As we state on p. 29/30 we were not able to code certain variables (e.g. team climate) that might be relevant because the selected literature did not report them.

R1: Could you explain the search terms. Why 'decision making’ and ‘leadership’ and nothing about effectiveness or performance or process or outcomes?

Response: We did include “decision-making” and “leadership” because these are the teamwork processes we are interested in based on current models (e.g. Reader et al. 20094).
We did indeed not include performance in our search terms. This has an important reason. During the search, we noticed that we would limit ourselves too much if we include performance as a search term and potentially miss relevant literature. Especially studies measuring outcome performance (e.g. mortality) do not necessarily mention performance in their key words. Therefore, we did not include the term in our search. This resulted in many more results that we then manually screened and selected according to relevant performance measures.

4Reader, T. W., Flin, R., Mearns, K., & Cuthbertson, B. H. (2009). Developing a team performance framework for the intensive care unit. Critical Care Medicine, 37(5), 1787–1793.

R1: Why did you choose for a meta-analytical study approach? Would a scoping review give more insight into the aspects of teamwork that have a positive impact on team performance? Most studies included in the review are observational studies. The aim of the papers, their design, rating scales, and settings are very different. Are they comparable and are the effects quantifiable? Since the information about the studies are very limited (design, rating scales), I’m not sure how to interpret the mean correlation and the tests regarding the moderators.

Response: Thank you for this suggestion. We agree that a scoping review would have some advantages compared to a meta-analysis. However, there are already many reviews published that could be considered as a scoping review and synthesize the literature about teamwork or various aspects of teamwork in healthcare.5-9 Because of this we think another review might not advance the field much. As we state in the introduction “A meta-analytical approach moves beyond existing reviews on teamwork in healthcare[9,15-18] and quantitatively tests if the widely advocated positive effect of teamwork on performance holds true”. We acknowledge the power of a more qualitative analysis of the literature. However, we see today that hospital policy makers value quantitative data more than qualitative results. Also, as Borenstein et al. (2011)10 notes in their book “Introduction to Meta-Analysis” that reviews become less useful as more information becomes available. The author of a review is required to capture the findings reported in each selected study, assign an appropriate weight to the finding, and then to synthesize the findings across all studies. This process becomes difficult and eventually untenable as the number of studies increases. Therefore—despite its disadvantages—we think that a quantitative approach to synthesize the literature is highly needed and missing in the literature about teamwork in healthcare. Also, we would like to highlight that we provide more qualitative information that you would traditionally see in a systematic review in Table 1. The reason why most studies are observational probably comes from our research question. We are interested in a correlation between two variables not in interventions. Also, a clear measurement of teamwork was an inclusion criteria and it might be more difficult to assess teamwork and relate it to an outcome with questionnaires. Nevertheless, that is what the literature search resulted in. We did not limit ourselves to observation studies.
The question about the comparability and quantifiably of the studies is a valid point and a critique that all meta-analysis’ have to deal with. Robert Rosenthal, one of the pioneers in meta-analytic procedures, was once asked if it makes sense to perform a meta-analysis, given that the studies differ in various ways and the analysis amounts to combining apples and oranges. Rosenthal answered that this makes sense if your goal is to produce a fruit salas.9 We are certainly not saying that all studies in our meta-analysis are identical and it is rarely the case in meta-analytical studies. The main result of our study is the common effect or mean effect that teamwork processes have on clinical performance. As we outline in the background section teamwork is a multi-facetted construct including various variables and therefore we had also to include a variety of studies that conform with our framework of teamwork. The mean correlation provides the mean effect that teamwork processes have on clinical performance on average. This is an estimation of the true effect in the population. The tests for moderation is investigating if this relationship is stronger or weaker under certain conditions.

We also added one paragraph to the discussion section how to interpret the correlation. I hope we could address your questions and concerns.

5Schmutz J, Manser T. Do team processes really have an effect on clinical performance? A systematic literature review. Br J Anaesth 2013;110:529–44.
6Manser T. Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. Acta Anaesthesiol Scand 2009;53:143–51.
7Fernandez Castelao E, Russo SG, Riethmüller M, et al. Effects of team coordination during cardiopulmonary resuscitation: A systematic review of the literature. J Crit Care 2013;28:504–21.
8Dietz AS, Pronovost PJ, Benson KN, et al. A systematic review of behavioural marker systems in healthcare: what do we know about their attributes, validity and application? BMJ Qual Saf 2014;23:1031–9.
9Flowerdew L, Brown R, Vincent C. Identifying nontechnical skills associated with safety in the emergency department: a scoping review of the literature. Ann Emerg Med 2012;59:386–94.
10 Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to Meta-Analysis (pp. 1–413). Hoboken, NJ: John Wiley & Sons, Ltd.

R1: “We provide strong evidence that teamwork contributes considerably towards quality of care – or in other words, poor teamwork significantly increases the risk for unsafe care and even patient harm”, discussion page 23, l19Your findings suggest that better teamwork improves the quality of care, but what is your basis for the statement regarding unsafety and patient harm?

Response: Thank you for this comment. You are right, strictly speaking we cannot make this statement based on our data. We deleted the sentence. Minor issues

Page 5, 17: I would say ‘generate evidence’ instead of strong evidence

Response: We deleted “strong”.

Page 6, 55: I do not understand the assumptionPage 7, 28: I do not understand the expectation

Response: We had another close look at the paragraph and where not sure what aspect we needed to change. Since it is listed as “minor issues” we do not expect to change the whole reasoning. Also, the other reviewers did not provide any comments concerning these two paragraphs. We are happy to make any changes in the next round based on more specific feedback what specific aspects or words are unclear.

Page 7, 40: teamwork is more important or ‘ effective’ teamwork is more important?

Response: We use important here as the expression for the moderation. We expect the effect to be stronger in non-routine situations making it more important. We do not want to use effective here because the reader might confuse this with the literature about team effectiveness.

Page 9, 26: I do not see clinical performance in the search term, page 39

Response: Yes, that is true. We realized that if we use “clinical performance” as a search term we would potentially miss relevant literature. For example, we consider patient outcomes (e.g. infection rate, length of stay) as performance. But some papers don’t add the keyword “clinical performance” nor do the authors use the words “clinical performance” in their manuscript. Therefore, such a paper would not show up in the search. So, we decided to leave it open and getting more search results and

manually select the relevant clinical performance measures when we screened title and abstracts.

Reviewer: 2

Reviewer Name: Philip Chilibeck Institution and Country: University of Saskatchewan, Canada Please state any competing interests or state 'None declared': None declared Please leave your comments for the authors below I was asked to provide a statistical review of the manuscript. I have included in the review some comments about the systematic review and other minor suggestions.

R2: Page 2, abstract, lines 25-26: Change "two individual" to "two individuals"

Response: Done, thank you for this comment.

R2: Page 6, line 10: Change "team members experience" to "team members' experience"

Response: Done. Thank you for this comment.

R2: Page 9, line 35: It is stated as part of the search strategy: "a manual forward search to identify studies that cite the studies we included in our meta-analysis" How were these studies identified (i.e. what database was used to identify these studies; i.e. was it Web of Science or another database?).

Response: Yes, we used Web of Science. We added this to the sentence.

R2: Page 9-10: The description of inclusion criteria for the included studies seems very brief. Please ensure you follow the PRISMA guidelines in detail here. Is it relevant to describe the "PICOS" here (i.e. Population, intervention, comparator, outcomes, study type)?

Response: Thank you for this suggestion. We had again a closer look at the PRISMA statement. PRISMA states "Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.". We think Figure 1 in combination with the text in the methods section provides the necessary information. If something important is missing we would be happy to add this information if you could point us into the right direction.
PICOS is less relevant here since we do not include any intervention studies, because we are interested in the relationship between teamwork and clinical performance. Information about population, outcomes and study type can be found in Table 1 and 2.

R2: In the statistics section (pages 19-20): Please provide reference to how you classified the effect sizes (i.e. as "small", "medium", "large", etc.) with an indication of the cut-offs used.

Response: Thank you for this comment. We provide a reference in the text to Cohen and now also to Bosco et al (2015)11 . Both provide guidelines about effect size classification. Unfortunately, we are not able to provide clear cut of values, because depending on the source they differ. Bosco et al. provide an interesting overview about this topic. They created a figure summarizing the recommendations of different studies how to classify a "medium" effect. As you can see in the figure below the common effect of our meta-analysis is clearly considered within or above the ranges of most studies recommendations for a "medium" effect. Please see attached PDF for Figure.

11 Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. Journal of Applied Psychology, 100(2), 431.

R2: Please indicate the statistical test you used to assess whether heterogeneity was present.

Response: To test for heterogeneity, we calculated Q and I2. The significant Q value ($Q = 53.73$, $p < .05$) indicates that is unlikely that all studies share a common effect size; however, it does not inform us about the amount of dispersion of the true effect sizes. For this purpose, we calculated I2 ($I2 = 45.96$). According to Higgins et al.; 2003), a value of 50 (%) might be considered as moderate amount of true variance.

12 Higgins, J., Thompson, S.G., Deeks, J.J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. BMJ, 327, 557–560.

R2: Did you do any assessments for study quality (i.e. bias)? I know there are a number of tools used to assess quality of randomized controlled trials (i.e. Cochrane tool, Jadad score) – are there any tools available to assess quality of the types of studies you included in your meta-analysis?

Response: Thank you for this suggestion, certainly a valid point. Indeed, there are a few tools in order to assess the quality of intervention studies. However, tools to rate descriptive observation studies are rare. We address this on page 21: “Since we included only descriptive studies and no interventions we only included the sample size of the individual studies as a potential bias into the meta-analysis.”. Also, if we were able to create a quality indicator (e.g. 1-10) for each study this would not inform the meta-analytical procedure. The meta-analysis itself is considering a larger margin of error in studies with smaller sample sizes which can be considered as a quality indicator that we included in the analysis. R2: Page 20, lines 29-33: I think this description of how outliers were identified should be moved to the statistics section, rather than being described in the results section.
Response: Thank you for this comment. We moved one sentence into the methods section.

R2: Figure 1: In the boxes of the flowchart, clarify where titles and abstracts were screened.Response: Thank you, we missed that. We added “titles and abstract” to Figure 1.

Reviewer: 3 Reviewer Name: Ashley Hughes Institution and Country: University of Illinois at Chicago, Chicago, IL, USAd Please state any competing interests or state ‘None declared’: None declared Please leave your comments for the authors below See attached file

R3: This review provides a much needed synthesis on the quantitative relationship between teamwork and clinical performance outcomes. I applaud the authors for the clarity of the writing, methodology employed, and overall flow of the paper. Performing this type of meta-analytic integration which spans multiple fields is no small feat. The paper is well written and provides and excellent translation of the literature on team science from multidisciplinary journals to healthcare.
Response: Thank you very much for this positive feedback.

I have a few areas for improving the manuscript prior to its publication in BMJ open:

My primary comments have to do with the methodology used for the study.

R3: First, in my understanding of the literature on teamwork in healthcare, these articles tend to be published in a variety of outlets, as the authors mention in the introduction (Lines 28-30). Why is pubmed the only database utilized for identifying articles? Small k is an issue in investigating some of your interesting moderator analyses so expanding the search could prove advantageous.

Response: Thank you for this feedback. We focused on PubMed since it is the most common database to access papers that potentially investigate medical teams. PubMed includes 30’000 journals from the field of medicine, psychology and management and all the most impactful journals are represented. We are fairly confident that through the additional inclusion of relevant reviews and forward and backwards search, our results represent an accurate representation of what is in the literature. Adding another database to our search is at this point very difficult for us given the time frame for the resubmission of the manuscript (28 days). However, we do acknowledge this fact in the limitation section now.

PubMed List: https://www.nlm.nih.gov/bsd/serfile_addedinfo.html

R3: Second, I appreciate the inclusion of forest plots and I2 statistics for heterogeneity. However, I do not understand the rationale for selecting Fischer’s Z scores used to create composites rather than use of averages or other composite creating techniques (Nunnally, 1978) which are used in other team metas (DeChurch & Mesmer-Magnus, 2010)? Please clarify the advantages or rationale motivating this approach.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-analysis. Journal of Applied Psychology, 95(1), 32-53.

Nunnally, J. C. (1978). Psychometric theory. New York, NY: McGraw-Hill.

Response: Thank you for this feedback. Altough DeChurch and Mesmer-Magnus (2010) note that an “average correlation was computed» (p. 37), we are not sure about the exact way they did this. Using simple arithmetic average (i.e., correlations will be summed and divided by the number of coefficients) is problematic because the distribution of r becomes negatively skewed as the correlation is larger than zero. As a result, the average r tends to underestimate the population correlation. As a solution to this problem, scholars have suggested to convert r to Fisher's z scores, to average the zs, and then to backtransform it to r (Rambo, Chomiak, & Price, 1983). As shown by Silver and Dunlap (1987), the average z backtransformed to r is less biased than the average r. We therefore used Fisher’s z scores to average multiple correlations derived from the same sample.

Rambo, W. W., Chomiak, A. M., & Price, J. M. (1983). Consistency of performance under stable conditions of work.Journal of Applied Psychology, 68, 78-87.
Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used?. Journal of Applied Psychology, 72(1), 146–148.

R3: Third, clinical context is surprisingly mentioned nowhere within the approach. I agree with the authors' conclusions that the nature of the team task interdependence would vary based on patient acuity, there are a variety of other factors that would be different based on clinical context (types of tasks, patients, and associated staffing structure). Is it possible to explore this further or otherwise provide some guidance as to what this assessment might look like? I'm surprised by this as you are submitting for publication in a medical journal rather than a business or psychology outlet without much mention of exactly what constitutes clinical performance or the clinical context.

Response: Thank you for this comment. We provide more information about clinical context in various parts of the paper. First of all, we added a paragraph to the limitations section where we acknowledge that teamwork needs to be interpreted in the light of the clinical context. We definitely want to avoid that the reader—based on our non-significant findings of the moderations—thinks that teamwork is a one-size-fits-all concept. Also, we provide more background about how clinical performance measures are context specific in the discussion. We hope with this we could address your comments.

R3: Fourth, out of pure curiosity, the authors mention F-tests as part of the statistics encountered in primary studies. However, I'm guessing that the authors' choice in using a correlational meta-analysis was due to lack of experimental or quasi experimental designs.

Response: That is exactly right. Most studies reported a correlation between a teamwork variable and performance and did not conduct any intervention. Therefore, we chose using a correlational meta-analysis.

R3: Further, the authors describe the rationale for their meta-analysis as stemming from a need to establish a direct relationship between teamwork and clinical performance. Please clarify.

Response: Thank you for this comment. We think the teamwork-performance relationship is important since this defines the effectiveness of teamwork. Every new medication that is released is tested if it is effective or not and what the effect is. We see the teamwork-performance relationship as an indicator of how important teamwork is in relation to performance. In our opinion this is especially important since many teamwork training studies establish a relationship between training and performance but the do not establish a relationship between training and teamwork, so the exact mechanisms are unknown. With our meta-analysis we add to the understanding of the team process-performance relationship.

R3: Also, I appreciate that levels of analysis for coding were taken into consideration and that the authors chose team level as the level of analysis. This is appropriate and makes findings that much more relevant and compelling.

Response: Thank you.

R3: Interesting dilemma about the reliability measures employed. I empathize with the lack of Cronbach's alphas reported; however, a potential source of moderation could be measurement criterion (i.e., observation versus survey).

Response: Interesting idea, thank you. As we note on page 23 there were only 3 studies who used survey methods and one of them did not go into the analysis because it is considered as an outlier (Cooper & Wakelam, 1999). Nevertheless, we were curious and added survey vs. observation as a moderator but could not find any significant effect (p = .39).

R3. Real or simulated patient- how were patient actors characterized? Arguably, this may be different from a human patient simulator, particularly in regards to ability to demonstrate teamwork. Also, we are talking about teamwork demonstrated within the care team? Is the patient included? Why would having a real patient matter in the context of team process within the clinical care team? Please elaborate.

Response: Thank you for this feedback. We can see that the explanation on page 12 about patient realism can be misleading. In the beginning, we thought about assessing simulated patients (patient actors) as “simulated”. We did not find any study that used patient actors though. Therefore “simulated” means the study used a patient simulator (manikin). We changed the wording in the methods section so that is clear that simulated means a training manikin and not a human being. Our reasoning why it should matter to distinguish between real and simulated patient comes more from a methodological point of view. We provide 2 arguments for including this moderator on p. 8. First, studies conducted with a patient simulator might be more standardized than studies with real patients and therefore less influenced by confounding variables. This would potentially result in a stronger measured effect between teamwork and performance. Second, there is always the fear that healthcare workers behave differently in a simulated setting compared to a real setting. This could go into both directions. Either individuals put more effort into the case because they see the simulation as a kind of test or they question the realism of the situation and are less motivated to participate and collaborate. Both cases would be a potential confounding variable. In fact, we are happy that we did not find a moderation because this means, as we state in the discussion, that the effects observed in a simulated setting are comparable with the effect in real life. Which provides a strong argument for the realism of simulation as well as using simulation for teamwork research.

R3: Double coding… why only 25%?

Response: Since PRISMA does not provide any specific guidelines about the amount of data that needs to be double coded we correspond to published reviews that coded around 20% of the data. Eg.

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta-analysis of teamwork processes: tests of a multidimensional model and relationships with team effectiveness criteria. Personnel Psychology, 61(2), 273–307. 12 out of 83 papers.
Castelao, E., Russo, S. G., Riethmüller, M., & Boos, M. (2013). Effects of team coordination during cardiopulmonary resuscitation: A systematic review of the literature. Journal of Critical Care, 28(4), 504–521. 20%
Deneckere, S., Euwema, M., Van Herck, P., Lodewijckx, C., Panella, M., Sermeus, W., & Vanhaecht, K. (2011). Care Pathways Lead to Better Teamwork: Results of a Systematic Review. Social Science and Medicine. 10%

R3: Correlations are low. I don’t see much discussion on why this could be or connection to what this would mean in context… Bare bones meta-analysis will produce lower than typical correlations with higher than usual standard deviations. To a clinical audience, this correlation is going to seem LOW. Why does this correlation matter? How does this translate? Couch your findings in the context of what it means to improve quality. What it means to patients and what problems may be addressed through improved team process. Connect to the teams literature as well as the clinical care.

Response: Thank you for this comment. We agree that at the first glance the correlation seems small, however it is considered as a medium not a small effect and needs to be recognized as such. We agree that we did not do a good job in putting this number into perspective. We added a whole paragraph about the interpretation of the correlation in the discussion section.
A correlation of .28 corresponds to an odds ratio of 2.81. Of course, this transformation seems artificial and simplifies the correlation because teamwork and often the outcomes are not simple dichotomous variables that can be divided into an intervention vs control group. However, an OR of 2.8 illustrates that teams with good teamwork are 2.8x more likely to show good performance. If you imagine that teamwork is only one predictor of probably many this should not be underestimated. If we then take into account that clinical performance is often associated with patient outcomes or patient safety no one would say “no” to a process that would increase the likelihood of good performance about 2.8 (OR) times.
We also added a paragraph about clinical performance measures and what they mean in different contexts. This should also help interpreting the results. Finally, we also added literature about CRM and highlight, that based on our results teamwork should not only be trained in crisis situations but also in more routine situations.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to Meta-Analysis (pp. 1–413). Hoboken, NJ: John Wiley & Sons, Ltd.

Arthur, W., Jr., Bennett, W., & Huffcutt, A. I. (2001). Conducting meta-analysis using SAS. London, UK: Psychology Press.

Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis:Correcting error and bias in research. Thousand Oaks, CA: Sage.

R3: This is just a preference to ease interpretation of your results- you report 95% confidence intervals for the main correlational finding. Unfortunately, at first glance, the range (lower to upper) looks like a negative sign. I recommend choosing a different format as allowed by the journal.

Response: We did make the adjustment in the text and now write “95% CI: .20 to . 35”

R3: Rather than stating an upfront limitation of a possible file drawer effect, let’s test the likelihood of it. Comprehensive meta-analysis as a program can assist with this- Hunter and Schmidt (2004) and Duvall and Tweedie (2000) offer some guidance on this as well. There are other ways too of testing for presence of file drawer effect or publication bias.

Response: Thank you for this suggestion. We did test for publication bias with the Egger test. The results indicate that there is no asymmetry in the funnel plot ($z = 1.79$, $p = .074$), suggesting that there is no publication bias. We added one sentence to the methods section on page 22.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.) Publication bias in meta-analysis: Prevention, assessment, and adjustments (pp. 99–110). Chichester, England: Wiley.

R3: Model and moderation clarification- you mention testing models. I am familiar with random effects models as well as model-based methods in meta-analysis (Hunter & Schmidt, 2004). However, I do not see coding or analytic procedures for model-based testing (see Cheung, 2008) as I am used to seeing it. Please clarify the approach and why it was chosen. The same goes for testing the presence of a significant moderating effect. Currently, the approach differs from those cited as motivating the current work. Why was the current method for testing moderators chosen over using an approach

such as Whitener non-overlapping confidence intervals (Whitener, 1990) or Zou's (2007) confidence interval significance test(s)

References

Cheung, M. W. L. (2008). A model for integrating fixed-, random-, and mixed-effects meta-analyses into structural equation modeling. Psychological Methods, 13(3), 182-202.

Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in

research. Thousand Oaks, CA: Sage.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. Journal of Applied Psychology,75, 315–321

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. Psychological Methods, 12, 399–413. http://dx.doi.org/10.1037/1082-989X.12.4.399

Namely, in reviewing the metafor package for clarification on omnibus techniques, these appear to test linear coefficients in models. While I am unclear as to how the coding would allow the model to be constrained in terms of accounting for interrelatedness amongst moderators, I am definitely unclear on the use of omnibus significance tests for categorical moderators in meta-analysis that claims to use a model-based approach.

Response: We agree with you that different approaches to test moderator effect in meta-analysis have been suggested, and we are sorry for not being clear about the rationale of our choice of the technique proposed by Viechtbauer (2010). As you mentioned, an alternative approach would be to compare the estimates of the subgroups, using non-overlapping CI or CI significance tests
Our decision to use Viechtbauer's approach was driven by two reasons. First, we not only have categorical moderators (e.g., professional composition) but also a continuous moderator (team size). The problem with dichotomizing a continuous moderator variable is well-established (Cohen & Cohen, 1983; Stone-Romero & Anderson, 1994) and subdividing a continuous moderator into subgroups should be avoided. Second, moderators might share substantial variance and hence should not be analyzed separately (Viswesvaran & Sanchez, 1998). As noted by Steel and Kammeyer-Mueller (2002, p. 97), "the danger of ignoring issues of multicollinearity is not only to complicate issues, by failing to eliminate redundant variables, but also to misdirect researchers' attention." We, therefore, decided to examine the effects of the moderator variables simultaneously.

For this purpose, we tested a mixed-effects model including professional composition, team familiarity, team size, task type, patient realism, and performance measure as moderators. The omnibus test of the coefficients of the moderator variables was not significant ($p = .98$), indicating that the null hypothesis that all of these coefficients are equal to zero cannot be rejected. In line with that, all individual coefficients were not significant ($ps > .45$) in this model.
Given that some scholars might disagree with Viswesvaran and Sanchez (1998) and Steel and Kammeyer-Mueller (2002) and argue that a test with multiple moderators is overly conservative, we also examined the moderators individually (i.e., a separate analysis for each moderator variable). None of these separate moderator analyses was significant (professional composition: $p = .84$; team familiarity: $p = .55$; team size: $p = .87$; task type: $p = .60$; patient realism: $p = .83$; performance measure: $p = .20$), confirming the result of the multiple moderators model.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlational analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. Journal of Applied Psychology, 87(1), 96-111.

Stone-Romero, E. F., & Anderson, L. E. (1994). Relative power of moderated multiple regression and the comparison of subgroup correlation coefficients for detecting moderating effects. Journal of Applied Psychology, 79, 354–359.

Viswesvaran, C., & Sanchez, J. I. (1998). Moderator search in meta- analysis: A review and cautionary note on existing approaches. Educational and Psychological Measurement, 58, 77–87.

VERSION 2 – REVIEW

REVIEWER	Anneke van Dijk - de Vries Maastricht University, CAPHRI Care and Public Health Research Institute
REVIEW RETURNED	12-Apr-2019

GENERAL COMMENTS	The revised paper has been improved a lot, my compliments. My residual suggestion is to mention in the title and/or abstract that the focus of the review and meta-analysis is on acute health care teams.

REVIEWER	Philip Chilibeck University of Saskatchewan, Canada
REVIEW RETURNED	08-Apr-2019

GENERAL COMMENTS	The authors have adequately addressed all my concerns

REVIEWER	Ashley Hughes University of Illinois at Chicago USA
REVIEW RETURNED	08-Jun-2019

GENERAL COMMENTS	Overall, I applaud the authors for their efforts and noticeable improvements in the manuscript. I appreciate the detailed response(s) to my particular points and elaborate on a few points, mainly indicating where the manuscript could benefit from further clarification. * Abstract: Methods: Line 12-13: "data sources were searched up to June 2018 and included PubMed" --> this sentence sounds like the authors have searched several data sources which included PubMed, but in fact, they only used PubMed. Thus, I would suggest the abstract would reflect the fact that only 1 data source was searched in June 2018 without a timebound (e.g.: Pubmed was searched in June 2018 to identify potential articles). I realize this is an acceptable practice for healthcare related outlets; however, I am disappointed to see that my prior comment regarding the addition of other databases was not incorporated.

	Line 19-20: as the authors wanted to focus on health professional readers, it would be better to give examples for "team process" and "performance measure" like what was done for "moderator variables". Results: Line33-34: providing correlation "r" in full text, while in the response to reviewer letter, the authors stated that they provided OR for better interpretation of the results, which should be shown in the abstract too. * Full text: Methods: Search strategy: As they do not limit the date of publication when searching for articles, the actual date of the last search should be provided rather than "last search June 2018" because even with a small amount, there would be papers published in late June were not included in this study if the search ended in early June. Furthermore, talking about updating this review in the future, it would be better to have an end date of the previous search to continue. Discussion: Limitations and future directions: page20, line 20-22: it would be better to explain why they only search the PubMed database right in the text rather than in the response to reviewers letter. Readers need to know why also. Table 3: footnote for * B. Concepts that would benefit from clarification 1. Regarding simulation vs. real-life settings (discussion: bottom of page 18 and the top 2 lines/ page 19), I would like to address that patient care in real life involves many specialties, not just different professions. Thus, the relationship of specialties among patient care may affect how clinical performance was assessed. For example, surgery and anesthesia: anesthesia is the upstream part as surgery can't be done without anesthesia. Thus, the clinical outcome of a patient underwent the care of a surgery team, were affected by both the team works of anesthesia and surgery teams. Would the "simulated" studies with "standardized" scenarios be generalized about this? I think it'd better be careful when saying about "generalizing" simulation to real life settings. 2. Is "accuracy of diagnosis" a clinical outcome or process measure? It seems to be "process" to me as it related to HOW the provider performed, and in fact, some inaccurate diagnoses won't lead to "measurable" clinical outcome. I saw the authors classified "accuracy of diagnosis" as "outcome" measure then. Statistical considerations There were several areas of the statistical analysis that appeared unclear. I have these areas detailed below for the authors' consideration and further clarification of the manuscript. It's clear that outliers were identified; however, the method used should be clarified. Arthur and Huffcutt is typically associated with the meta technique ?? Reference

	Huffcutt, A. I., & Arthur, W. (1995). Development of a new outlier statistic for meta-analytic data. Journal of Applied Psychology, 80(2), 327-334. Double check that your PRISMA flow diagram matches the numbers that you mention in the paper Standardized residual score appears high. Please explain Only four studies assessed outcome performance measures. Measures included accuracy of diagnosis, postoperative complications and death, surgical morbidity and mortality, ventilator-associated pneumonia, bloodstream infections, pressure ulcers and acute physiology and chronic health evaluation score. Significant heterogeneity even in the presence of moderator analysis ; I see that the referenced paper Hughes et al (2016) does not report heterogeneity; Reference Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., Marlow, S. L., Lacerenza, C. N., ... & Salas, E. (2016). Saving lives: A meta-analysis of team training in healthcare. Journal of Applied Psychology, 101(9), 1266-1304. What does * stand for in your tables? I'm used to this referring to a p-value; yet, it is only applied to your Q-statistics, which are easily influenced by sample size. Almost all of your findings demonstrate significant heterogeneity, which currently is not clearly expressed in the table. Please clarify "We excluded articles investigating long-term care since the dynamics of teamwork over a longer period of time are different" Healthcare teams literature: Coordination of care for chronically ill patients, such as those in longterm care facilities, have unique team task interdependencies Teams literature: Yes, teams over time exhibit different teamwork competencies; namely, it allows for iterative performance cycles which allow for emergent states, typically strengthening certain processes or at least states. The sentence as written doesn't make sense in the context of the literature. The sentence needs to be supported References Kianfar, S., Carayon, P., Hundt, A. S., & Hoonakker, P. (2019). Care coordination for chronically ill patients: Identifying coordination activities and interdependencies. Applied Ergonomics, 80, 9-16. Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. Academy of Management Review, 26(3), 356-376.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Anneke van Dijk - de Vries Institution and Country: Maastricht University, CAPHRI Care and Public Health Research Institute Please state any competing interests or state 'None declared': None declared Please leave your comments for the authors below

The revised paper has been improved a lot, my compliments. My residual suggestion is to mention in the title and/or abstract that the focus of the review and meta-analysis is on acute health care teams.

Response: Thank you very much. We added this information to the abstract. Also, we highlight again now on page 12 again why we excluded teams from long-term care.

Reviewer: 3

Reviewer Name: Ashley Hughes Institution and Country: University of Illinois at Chicago, USA Please state any competing interests or state 'None declared': None declared Please leave your comments for the authors below

Overall, I applaud the authors for their efforts and noticeable improvements in the manuscript. I appreciate the detailed response(s) to my particular points and elaborate on a few points, mainly indicating where the manuscript could benefit from further clarification.

Response: Thank you.

* Abstract:

Methods: Line 12-13: "data sources were searched up to June 2018 and included PubMed" --> this sentence sounds like the authors have searched several data sources which included PubMed, but in fact, they only used PubMed. Thus, I would suggest the abstract would reflect the fact that only 1 data source was searched in June 2018 without a timebound (e.g.: Pubmed was searched in June 2018 to identify potential articles). I realize this is an acceptable practice for healthcare related outlets; however, I am disappointed to see that my prior comment regarding the addition of other databases was not incorporated.

Response: Thank you for this suggestion. We added the sentence, as you suggested to the abstract: "PubMed was searched in June 2018 without a limit on the date of publication.". So, it becomes clear that there was only one database searched. We acknowledge that the difference in practice across research communities can sometimes be frustrating. However, we did our best to describe our methods as accurately and transparent as possible to allow for readers to consider this potential limitation when interpreting our findings.

Line 19-20: as the authors wanted to focus on health professional readers, it would be better to give examples for "team process" and "performance measure" like what was done for "moderator variables".

Response: We added examples to "team processes" and "performance measures".

Results: Line33-34: providing correlation "r" in full text, while in the response to reviewer letter, the authors stated that they provided OR for better interpretation of the results, which should be shown in the abstract too.

Response: Thanks for this suggestion. We added to the abstract: “corresponding to an odds ratio of 2.8” in brackets.

Full text: Methods: Search strategy: As they do not limit the date of publication when searching for articles, the actual date of the last search should be provided rather than "last search June 2018" because even with a small amount, there would be papers published in late June were not included in this study if the search ended in early June. Furthermore, talking about updating this review in the future, it would be better to have an end date of the previous search to continue.

Response: We agree, thanks for this suggestion. We added the exact data to the text on page 11.

Discussion: Limitations and future directions: page20, line 20-22: it would be better to explain why they only search the PubMed database right in the text rather than in the response to reviewers letter. Readers need to know why also.

Response: Thank you, we added the following sentence to this paragraph: “PubMed is the most common database to access papers that potentially investigate medical teams and includes approximately 30’000 journals from the field of medicine, psychology and management. We are confident that through the additional inclusion of relevant reviews and forward and backwards search, our results represent an accurate representation of what can be found in the literature.”

Table 3: footnote for *

Response: Thank you for reading the manuscript so carefully. We now added in the footnotes that the asterisk indicates that the Q value is significant.

B. Concepts that would benefit from clarification 1. Regarding simulation vs. real-life settings (discussion: bottom of page 18 and the top 2 lines/ page 19), I would like to address that patient care in real life involves many specialties, not just different professions. Thus, the relationship of specialties among patient care may affect how clinical performance was assessed. For example, surgery and anesthesia: anesthesia is the upstream part as surgery can't be done without anesthesia. Thus, the clinical outcome of a patient underwent the care of a surgery team, were affected by both the team works of anesthesia and surgery teams. Would the "simulated" studies with "standardized" scenarios

be generalized about this? I think it'd better be careful when saying about "generalizing" simulation to real life settings.

Response: Thank you for raising the issue of multiple specialties in healthcare teams and its implications for teams being composed of multiple crews. This issue is more relevant to performance in lower acuity settings where these teams collaborate over longer periods often in a desynchronised and non-collocated manner. However, our review focused on acute care teams and the contribution of all members of the interprofessional, interdisciplinary team to a joint outcome. The simulations in the included studies recreate the tasks from the clinical environment including multiple specialties and were appropriate. Thus, we believe that the generalizability that has also been highlighted in several studies investigating the feasibility of simulation as a research environment for acute care teams can be assumed. We added "…in acute care teams" to the sentence to make clear that we cannot draw any conclusions for long-term care teams.

2. Is "accuracy of diagnosis" a clinical outcome or process measure? It seems to be "process" to me as it related to HOW the provider performed, and in fact, some inaccurate diagnoses won't lead to "measurable" clinical outcome. I saw the authors classified "accuracy of diagnosis" as "outcome" measure then.

Response: We see your point here. We are assuming you are referring to Tschan, Semmer, Gurtner et al. 2009. In this paper, the authors conceptualized diagnostic accuracy as the outcome and linked this with team processes. In this paper, the teams had time to talk and exchange information (team process) and after that they had to provide their assumption what the diagnosis is. The act of stating the diagnosis was done after the case was handled. Also, it does not explicitly describe a procedure to follow. Therefore, we considered it as outcome performance.

Tschan, F., Semmer, N. K., Gurtner, A., Bizzari, L., Spychiger, M., Breuer, M., & Marsch, S. U. (2009). Explicit reasoning, confirmation bias, and illusory transactive memory: A simulation study of group medical decision making. Small Group Research, 40(3), 271–300. http://doi.org/10.1177/1046496409332928

Statistical considerations. There were several areas of the statistical analysis that appeared unclear. I have these areas detailed below for the authors' consideration and further clarification of the manuscript. It's clear that outliers were identified; however, the method used should be clarified. Arthur and Huffcutt is typically associated with the meta technique ??

Reference Huffcutt, A. I., & Arthur, W. (1995). Development of a new outlier statistic for meta-analytic data. Journal of Applied Psychology, 80(2), 327-334.

Response: We apologize for not being clear in the earlier versions of this manuscript and now explain that we followed Viechtbauer and Cheung's (2010) advice and used the studentized deleted residuals to identify outliers. One case (Carlson et al.,[9] r = .89, n = 44, studentized deleted residuals = 4.26) was identified as outlier and therefore excluded from further analyses,

Double check that your PRISMA flow diagram matches the numbers that you mention in the paper.

Response: We checked Figure 1 again and added one more sentence to the text: “After duplicates were removed 1988 articles were screened using title and abstract.” So the logical flow in the text reflects Figure 1.

Standardized residual score appears high. Please explain

Response: As noted above, we used the studentized deleted residuals to identify outliers. The study by Carlson et al. had an extraordinarily high value and hence was excluded for further analyses.

Only four studies assessed outcome performance measures. Measures included accuracy of diagnosis, postoperative complications and death, surgical morbidity and mortality, ventilator-associated pneumonia, bloodstream infections, pressure ulcers and acute physiology and chronic health evaluation score.

Response: That is correct. That is how we explain it in the text. It seems that this is an excerpt from our manuscript on p. 23 and we are not sure what we should address here.

Significant heterogeneity even in the presence of moderator analysis; I see that the referenced paper Hughes et al (2016) does not report heterogeneity; Reference Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., Marlow, S. L., Lacerenza, C. N., ... & Salas, E. (2016). Saving lives: A meta-analysis of team training in healthcare. Journal of Applied Psychology, 101(9), 1266-1304.

Response: Also here we are not sure what we should address and we are missing a question. We checked the paper and we do not mention Hughes et al. in relation with heterogeneity. Do we miss something? We are happy to address this comment with further specific instructions.

What does * stand for in your tables? I’m used to this referring to a p-value; yet, it is only applied to your Q-statistics, which are easily influenced by sample size. Almost all of your findings demonstrate significant heterogeneity, which currently is not clearly expressed in the table. Please clarify

Response: We apologize for not mentioning this in the previous version of the manuscript. As you assumed, the asterisk indicates that the value is significant. We realized that it was misleading to use the asterisk only for the Q- but not the r-values. We changed the table accordingly and now also indicate which r values are significant ($p < .05$). That said, we believe that the reported I2 values clearly indicate that there is considerable heterogeneity in the estimates.

“We excluded articles investigating long-term care since the dynamics of teamwork over a longer period of time are different”Healthcare teams literature: Coordination of care for chronically ill patients, such as those in longterm care facilities, have unique team task interdependencies Teams literature: Yes, teams over time exhibit different teamwork competencies; namely, it allows for iterative performance cycles which allow for emergent states, typically strengthening certain processes or at least states. The sentence as written doesn’t make sense in the context of the literature. The sentence needs to be supported

References

Kianfar, S., Carayon, P., Hundt, A. S., & Hoonakker, P. (2019). Care coordination for chronically ill patients: Identifying coordination activities and interdependencies. Applied Ergonomics, 80, 9-16.Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. Academy of Management Review, 26(3), 356-376.

Response: Thank you for this suggestion. We added this information and the literature to the paragraph and now say: “We excluded articles investigating long-term care since the coordination of care for chronically ill patients has to consider the unique team task interdependencies in this setting.[48] Also, teams working together over longer periods of time are more likely to develop emergent states (e.g. team cohesion) that influence how a specific team works tog

VERSION 3 – REVIEW

REVIEWER	Ashley M Hughes University of Illinois at Chicago
REVIEW RETURNED	23-Jul-2019

GENERAL COMMENTS	1. Limitation of using only Pubmed: this was addressed in the "limitations" section. In the letter to reviewers, page 9/14, the authors said that they were "fairly confident that their results represented an accurate representation of what is in the literature", while in the full text, they omitted the word "fairly". I don't know if you would like to suggest they keep the word "fairly" to tone down a bit. I do think they should keep the work "fairly" unless they did search in other databases. 2. The use of Fischer's Z scores to create composites rather than using averages or other composite creating techniques. --> The authors had explained why they used Z scores. However, they did not address this in the text. Do you think they should say a bit about why they decided to use Z score? I think adding an explanation would help readers to understand why they choose this technique (a chance to learn then). 3. Request to clarify the approach and why chosen the "model-based testing" & "testing for the presence of a significant moderating effect" + Why chose the current method for testing moderators over suing other approaches?

	--> The authors had explained in the letter but not in the text. I think they should mention this in the text.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 3
Reviewer Name: Ashley M Hughes

Institution and Country: University of Illinois at Chicago

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below
1. Limitation of using only Pubmed: this was addressed in the "limitations" section. In the letter to reviewers, page 9/14, the authors said that they were "fairly confident that their results represented an accurate representation of what is in the literature", while in the full text, they omitted the word "fairly". I don't know if you would like to suggest they keep the word "fairly" to tone down a bit. I do think they should keep the work "fairly" unless they did search in other databases.

RESPONSE: We added the word “fairly” to the sentence.

2. The use of Fischer's Z scores to create composites rather than using averages or other composite creating techniques.
--> The authors had explained why they used Z scores. However, they did not address this in the text. Do you think they should say a bit about why they decided to use Z score? I think adding an explanation would help readers to understand why they choose this technique (a chance to learn then).

RESPONSE: Thank you for this comment, we added a footnote to explain why we did use the Z score.

3. Request to clarify the approach and why chosen the "model-based testing" & "testing for the presence of a significant moderating effect" + Why chose the current method for testing moderators over suing other approaches?
--> The authors had explained in the letter but not in the text. I think they should mention this in the text.

RESPONSE: Thank you for this feedback. We did consider adding the explanations about the mentioned issues to the manuscript. However, we feel that this part will be very technical and we would need at least 400-500 additional words to thoroughly explain the choice of the tests. This would lengthen the whole manuscript, that is already above the word count of an average paper in BMJ Open. Further, the focus of our paper is not the methods itself and all the necessary references about why we chose the methods are cited in the text. We hope that the editor agrees with our assessment. If not we will be happy to add an additional paragraph to the manuscript.

Thank you very much for the opportunity to revise our manuscript.