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)	Association between Measured Teamwork and Medical Errors: An Observational Study of Prehospital Care in the United States
AUTHORS	Herzberg, Simone; Hansen, Matt; Schoonover, Amanda; Skarica, Barbara; McNulty, James; Harrod, Tabria; Snowden, Jonathan M.; Lambert, William; Guise, Jeanne-Marie

VERSION 1 – REVIEW

REVIEWER	Simon Cooper Federation University Australia
REVIEW RETURNED	10-Aug-2018

GENERAL COMMENTS	Many thanks for this interesting and important paper which indicates that adverse safety events are more likely in teams with low teamwork scores. Thoughts and comments as below: Abstract Setting or participants – think you need to add some detail about the simulations here? Noting that this was a manikin based study not simulated patients etc etc Primary measures – suggest you clarify this sentence for greater clarity as one would need to read the full paper to understand this? Main paper The work should be pitched at an international audience so suggest you add background data from setting other than the USA. Also don't assume that the reader will understand terms such as 'VA' An explanation of the 'in situ' simulations is provided but in essence this does not appear to meet the definition as the simulations were run in a training centre i.e. not in a patients' home or in an ambulance etc. Interesting that you presented the cases in random order – through experiential learning this will have varied the outcomes between teams/scenarios e.g. team 1 may have made errors in scenario C which was the second scenario they completed - but team 2 who were presented scenario C as their 4th scenario will make less errors for this scenario – your thoughts? Have I got this right? Comments about the Clinical Teamwork Scale are a little superficial – what of its validity, reliability and feasibility. I note that a number of these factors don't appeared to have been measured – further some items look questionable i.e. 'patient friendliness' in an emergency situation is perhaps not the biggest priority for a team and for their rating? Did you video record simulations as a feedback mechanism and for assessment checks. What of inter-rater reliability?

	Greater explanation required occasionally for those not familiar with methods/statistical approaches e.g. “This relationship was not confounded by scenario” and in the earlier analyses section you say “We used a generalized estimate equation (GEE) model with an exchangeable correlation structure to account for potential clustering.” But there is no mention of this in the results Minor There are a few grammatical errors and/or some tortuous language in a few places e.g. strengths and limitations section and “Medical errors, the failure of a planned action to be completed as intended or the use of a wrong plan to achieve an aim” etc etc

REVIEWER	Michael Rosen Johns Hopkins University School of Medicine, USA
REVIEW RETURNED	13-Aug-2018

GENERAL COMMENTS	In general, this is a well-designed and reported study. It is unique in its context (pediatric EMS) and it is rigorous. However, there are a few minor to moderate issues detailed below that should be addressed. Most of these involve clarification of methods and overall framing of the contribution of this study. 1. Pg 4, strengths and limitations, first bullet point. This statement is not true. There have been numerous studies linking the quality of teamwork to safety, quality and other important outcomes for decades. Further, many of these studies are conducted in actual care delivery, not just simulations. This is a good study, but saying this is ‘some of the first data’ showing relationship between teamwork and safety is inaccurate. Several reviews (refs 6-8) of this literature and primary studies are even cited in this paper, though they are a bit dated. 2. Same section as above, fourth bullet point. There does not appear to be a quantification of the reliability or bias in this observational data. The authors say the CTS is a previously validated tool, but do not talk about what was done to train raters or assess inter-rater reliability. It is discussed for the measurement of errors (resolved by discussion), but not for the teamwork measurement. 3. Page 7, simulations section. “We conducted all simulations in situ at local EMS training centers…”. The use of the phrase ‘in situ’ is confusing here. Typically, in situ refers to simulations conducted in the actual unit / patient care area vs. a simulation center. But here it says ‘in situ at training centers.’ It sounds as if they were conducted at a training facility, and not really ‘in situ’, which for EMS would be in the community. Please clarify. Is this center based or truly in situ from the EMS provider perspective? 4. Page 10, data collection. For CTS scoring, it sounds as if multiple people per scenario scored teamwork, but it is not clear. If so, how were scores treated (e.g., averaged?), and can you present inter-rater reliability measures? In general, there needs to be a better description of how this tool is used. In results (pg 14) it references individual scores for communication, situational awareness, decision-making, and leadership / followership. It isn’t entirely clear if this is individual person (i.e., team member) scores or individual dimension scores for the team. The CTS paper is cited, but the scoring needs to be described in this paper more clearly.

	5. Page 13. It says there were only 170 scenarios with overall CTS scores, but previously it was stated a total of 176 scenarios completed. What happened to the six cases? 6. Page 15, link 15. “…teamwork was a significant contributor to medical errors.” This clearly expresses causality, something carefully avoided elsewhere in the paper. It would be helpful to be consistent (i.e., there is an association). The ‘contributing’ terminology is used elsewhere as well in the discussion.

REVIEWER	Catherine Best University of Stirling, UK
REVIEW RETURNED	15-Nov-2018

GENERAL COMMENTS	This paper is a well-conducted study. The manuscript is clearly written and the analyses appear appropriate to the conclusions. The study finds that teams scoring lower on a validated measure of teamwork are more likely to make errors in a simulated scenario. The only major comment I have is that this is a simple association study. The control for other potential confounding factors is largely absent. In the description of study participants in Table 1 it is clear that the authors have access to information on the constitution of the teams. Controlling for factors such as the mean number of years’ experience in a team and/or their mean proficiency in pediatric EMS, would test whether both teamwork skills and errors are correlated with experience and or technical skills. This has significant implications for whether the conclusions that you draw from this study are that teamwork is a potentially modifiable factor and that intervening in training teamwork skills will reduce errors. The alternative interpretation that has not been addressed is that teamwork and errors are both influenced by general competence factors (or alternatively that people work better with people they trust to be competent). I understand that other intervention studies have found a relationship between teamwork training and outcomes but the contribution of this study would be enhanced if other potential confounders were evaluated. Minor comments It would be useful to know how many experts evaluated each scenario. Were the experts constant throughout or did different experts evaluate different simulations/teams? I understand that the same experts evaluated errors as scored the teamwork measure. A discussion of whether this might introduce bias would be appropriate. The study is described as ‘prospective’ in the title but there is only a single time point so I would describe it as cross-sectional.

REVIEWER	Jesus Montero-Marin University of Zaragoza, Spain
REVIEW RETURNED	03-Dec-2018

GENERAL COMMENTS	The manuscript is important in the filed. Some suggestions to improve: Please, describe a bit more the characteristics of the CTS Please, refer the statistical software used for data analysis Please, provide not only p values but also the corresponding statistics (t, z, x2, ...)

	Please, present a table showing the multivariable regression analysis data How do authors interpret the absence of CTS differences in "patient friendliness"? The inclusion of more references in the discussion section to create a larger framework would be necessary. If authors can work on this suggestions the manuscript would be worth to be published.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1	
1. Abstract - Setting or participants – think you need to add some detail about the simulations here? Noting that this was a manikin based study not simulated patients etc etc	Thank you for your suggestion, we added the following to the abstract as suggested and also elaborated in text: “Simulations were conducted in situ using high-fidelity patient simulators, scene design, and professional actors playing parents and bystanders.
2. Abstract - Primary measures – suggest you clarify this sentence for greater clarity as one would need to read the full paper to understand this?	We have added details about the statistical methods used and edited for clarity as suggested.
3. The work should be pitched at an international audience so suggest you add background data from setting other than the USA. Also don’t assume that the reader will understand terms such as ‘VA’	Thank you for drawing our attention to this very important point. We have added text and references in both background and discussion that demonstrate that this is a global health issue. We believe our manuscript is much stronger for this framing. Thank you also for catching the abbreviation we have ensured that all abbreviations are also spelled out for first use.
4. An explanation of the ‘in situ’ simulations is provided but in essence this does not appear to meet the definition as the simulations were run in a training centre i.e. not in a patients’ home or in an ambulance etc.	We have provided further detail on the simulations to describe the in situ settings which included the street for the motor vehicle collision and the use of the agency’s own ambulances.
5. Interesting that you presented the cases in random order – through experiential learning this will have varied the outcomes between teams/scenarios e.g. team 1 may have made errors in scenario C which was the second scenario they completed - but team 2 who were presented scenario C as their 4th scenario will make less errors for this scenario – your thoughts? Have I got this right?	The sequence of simulation problems presented to the EMS teams were randomized to avoid the issue of progressive learning that occurs from being exposed to the series of simulations and likely experiential learning by the teams. Additionally we accounted for latent correlation at the level of team in the GEE analysis.
6. Comments about the Clinical Teamwork Scale are a little superficial – what of its validity, reliability and feasibility. I note that a number of these factors don’t appeared to have been measured – further some items look questionable i.e. ‘patient friendliness’ in an	We have added additional detail about CTS™. In validation studies, the CTS™ demonstrated substantial score concordance among raters, and excellent interrater reliability.[1-3] A systematic review of teamwork tools that have been used in obstetrics recently concluded that CTS™ was superior to other tools for measuring teamwork

emergency situation is perhaps not the biggest priority for a team and for their rating?	citing content and construct validity as well as reliability and ease of use.[4] We measure patient friendliness along with teamwork because we believe it is important to be mindful of the experience of the patient even in emergent situations. We don't consider it to be an element of teamwork per se, but we do think it is an essential element of high quality patient care.
7. Did you video record simulations as a feedback mechanism and for assessment checks. What of inter-rater reliability?	We video recorded all simulations. Because of the difficulty for capturing high quality audio and video of every team member in such dynamic environments, we consider the live observation and CTS™ to provide the most accurate rating of teamwork. CTS™ has demonstrated good reliability in several studies (see prior response)
8. Greater explanation required occasionally for those not familiar with methods/statistical approaches e.g. "This relationship was not confounded by scenario" and in the earlier analyses section you say "We used a generalized estimate equation (GEE) model with an exchangeable correlation structure to account for potential clustering." But there is no mention of this in the results	We have substantially revised the text about statistical approach to clarify our use of the GEE method to address the latent correlation of teams within the data set (i.e., each team challenged with as many as 4 scenarios).
9. There are a few grammatical errors and/or some tortuous language in a few places e.g. strengths and limitations section and "Medical errors, the failure of a planned action to be completed as intended or the use of a wrong plan to achieve an aim" etc	We have reviewed the paper and corrected any grammatical errors we could find.
Reviewer 2	
1. Pg 4, strengths and limitations, first bullet point. This statement is not true. There have been numerous studies linking the quality of teamwork to safety, quality and other important outcomes for decades. Further, many of these studies are conducted in actual care delivery, not just simulations. This is a good study, but saying this is 'some of the first data' showing relationship between teamwork and safety is inaccurate. Several reviews (refs 6-8) of this literature and primary studies are even cited in this paper, though they are a bit dated.	We acknowledge that this is not the first to associate teamwork and errors. We were not trying to imply this was the first. We were not aware of many that have found a quantitative relationship. We edited the bullet to focus on the pediatric EMS setting where we believe this is an accurate assertion. Alternatively, if it would be preferable to eliminate the phrase 'some of the first' entirely, we would be happy to change the sentence to: "This research provides important data that quantifies the relationship between clinical teamwork and the likelihood of medical errors".
2. Same section as above, fourth bullet point. There does not appear to be a quantification of the reliability or bias in this observational data. The authors say the CTS is a previously validated tool, but do not talk about what was done to train	The fourth bullet was meant to acknowledge the limitation of ultimately relying on expert judgment and that while these clinicians were not related to participants or their agencies, it is still a subjective measurement. We were unclear how to revise but are happy to respond to whatever

raters or assess inter-rater reliability. It is discussed for the measurement of errors (resolved by discussion), but not for the teamwork measurement.	the journal would suggest including, if preferred, to delete the bullet entirely. We have added additional details about CTS™, its reliability and quality assurances practices conducted during the study as requested.
3. Page 7, simulations section. “We conducted all simulations in situ at local EMS training centers…”. The use of the phrase ‘in situ’ is confusing here. Typically, in situ refers to simulations conducted in the actual unit / patient care area vs. a simulation center. But here it says ‘in situ at training centers.’ It sounds as if they were conducted at a training facility, and not really ‘in situ’, which for EMS would be in the community. Please clarify. Is this center based or truly in situ from the EMS provider perspective?	We have added details about the simulations to clarify that simulation sites included the team’s own ambulances as well as streets with vehicles etc.
4. Page 10, data collection. For CTS scoring, it sounds as if multiple people per scenario scored teamwork, but it is not clear. If so, how were scores treated (e.g., averaged?), and can you present inter-rater reliability measures? In general, there needs to be a better description of how this tool is used. In results (pg 14) it references individual scores for communication, situational awareness, decision-making, and leadership / followership. It isn’t entirely clear if this is individual person (i.e., team member) scores or individual dimension scores for the team. The CTS paper is cited, but the scoring needs to be described in this paper more clearly.	We have added detail about CTS performance as well as specific processes for this study. One of two experts trained in using CTS™ evaluated and scored teamwork. All scores were for the entire team.
5. Page 13. It says there were only 170 scenarios with overall CTS scores, but previously it was stated a total of 176 scenarios completed. What happened to the six cases?	Due to missing data for overall CTS^{TM} score, the data for 6 scenarios were not included in the regression analysis which is why data for the regression analysis was based on 170.
6. Page 15, link 15. “…teamwork was a significant contributor to medical errors.” This clearly expresses causality, something carefully avoided elsewhere in the paper. It would be helpful to be consistent (i.e., there is an association). The ‘contributing’ terminology is used elsewhere as well in the discussion.	We appreciate this observation and have endeavored to revise our phrasing to avoid conveying causality.
Reviewer 3	
1. The only major comment I have is that this is a simple association study. The control for other potential confounding factors is largely absent. In the description of study participants in Table 1 it is clear that the authors have access to	Thank you for encouraging us to clarify this potential confounder in our analysis. We used mean years of EMS experience to represent field experience and technical skill level of each team. In our multivariate analysis, mean years of EMS experience did not significantly alter our

information on the constitution of the teams. Controlling for factors such as the mean number of years' experience in a team and/or their mean proficiency in pediatric EMS, would test whether both teamwork skills and errors are correlated with experience and or technical skills. This has significant implications for whether the conclusions that you draw from this study are that teamwork is a potentially modifiable factor and that intervening in training teamwork skills will reduce errors. The alternative interpretation that has not been addressed is that teamwork and errors are both influenced by general competence factors (or alternatively that people work better with people they trust to be competent). I understand that other intervention studies have found a relationship between teamwork training and outcomes but the contribution of this study would be enhanced if other potential confounders were evaluated.	estimate of the effect of CTS™ on probability of error. We have added this finding to our Results.
2. It would be useful to know how many experts evaluated each scenario. Were the experts constant throughout or did different experts evaluate different simulations/teams?	See response to reviewer 2 comment 4. One of two evaluators expert in using CTS scored teamwork for each team running through a scenario.
3. I understand that the same experts evaluated errors as scored the teamwork measure. A discussion of whether this might introduce bias would be appropriate.	We have added more detail about the performance of CTS and the process in this study. We agree that ultimately even with robust tools, these evaluations rely on human judgment. This is addressed in bullet and discussion because we think this is important. CTS has been used around the world in different settings and clinical conditions and a recent systematic review on teamwork tools used in obstetrics found it to be superior to others available. We thank the reviewer for the comment as we believe these additional details are a helpful addition to the international audience.
4. The study is described as 'prospective' in the title but there is only a single time point so I would describe it as cross-sectional.	We find this somewhat tricky. The measurements of EMS team performance were obtained prospectively among a cohort of ems agencies in an area and during the conduct of the several simulations in time. We have removed prospective from the title.
Reviewer 4	
1. Please, describe a bit more the characteristics of the CTS	We have added additional detail about CTS that address performance and also speak to international use. In validation studies, the CTS™ demonstrated substantial score concordance among raters, and excellent interrater reliability.[1-3] A systematic review of teamwork tools that have been used in obstetrics recently concluded that CTS™ was superior to other tools for measuring teamwork

	citing content and construct validity as well as reliability and ease of use.[4]
2. Please, refer the statistical software used for data analysis	This has been included in the Statistical Analysis section.
3. Please, provide not only p values but also the corresponding statistics (t, z, x2, ...)	This has been added into the results section.
4. Please, present a table showing the multivariable regression analysis data	Table 4 has been included in the results section that addresses the multivariable regression analysis data as recommended. Now that we have substantially revised both methods and results for clarity and we clarified in the text that there were no significant differences between adjusted and unadjusted models, we wondered whether this new table was helpful. We would be fine with removing it, should the editors feel this additional Table is no longer necessary.
5. How do authors interpret the absence of CTS differences in "patient friendliness"?	We measured patient friendliness along with teamwork because we believe it is important to be mindful of the experience of the patient in emergent situations. We don't consider it to be an element of teamwork per se, but we do think it is an essential element of high quality patient care. For those reasons we were not particularly surprised that teams could have similar patient friendliness even if their teamwork differs substantially.
6. The inclusion of more references in the discussion section to create a larger framework would be necessary.	We appreciate this suggestion. We have added text and references to background, methods, and discussion to create the larger international framework as suggested.

VERSION 2 – REVIEW

REVIEWER	Simon Cooper Federation University Australia
REVIEW RETURNED	21-May-2019

GENERAL COMMENTS	Many thanks for the update on this interesting paper. It is a long time since I originally reviewed this paper so my apologies but on rereading I do believe it needs a few additions with regard to a description of the methods. At the moment the project is still not repeatable. Some key and minor points below Abstract Whilst trust is of course part of teamwork is it core? For example you may be better saying 'leadership, task management and communication' or make reference to non-technical skills early on e.g. leadership, teamwork, situation awareness and decision making" ? You also use the same terminology later Perhaps the term 'correlated' would best be avoided here as well and just keep it as a statistical terms?

Setting – you have missed the fact that there were four scenarios – that you mention later
CTS needs describing in full terms in abstract
Great summary of results here

Participants
Please could you define what you mean by ‘in situ” – just a quick statement early on as you do clarify later
Was the study ethically approved? Perhaps a statement in the text earlier – not just the statement at the end

Page 8 suggest that a description of the simulations would go well in table and please use full terms not abbreviations. Also you do not indicate how they were developed e.g. an expert clinical team with 20 years experience each etc etc etc. What was included in the pilot – how many runs did you do? How were the actors briefed/trained for the scenarios?

I don’t think you have defended why you have used CTS over other tools especially when the trials with this tool have been in obstetrics and I note that the systematic review you mention has in fact missed a number of rating tools.

Many thanks for the additional changes you have made

REVIEWER	Dr Catherine Best University of Stirling, UK
REVIEW RETURNED	24-May-2019

GENERAL COMMENTS

Second review of Association between Measured Teamwork and Medical Errors: An Observational Study of Prehospital Care in the United States

Thank you for giving me the opportunity to review the revised manuscript.
The manuscript is much improved. The additional information on the team work measure and the further exploration of covariates is useful. I have a few further comments.

1. In the abstract I would suggest using the term ‘generalized estimating equations’ rather than ‘multivariate regression model’ as it gives the reader more information about how the clustering of observations by teams was handled in the analysis and the estimation method. If I read ‘generalized estimating equations’ then I know this is a marginal model using quasi-likelihood estimation.
2. The use of the exchangeable correlation structure for the GEE assumes that the correlation between errors on the scenarios are the same for all pairs of scenarios ie those that are performed close together in time and those further apart and those on more similar topics and those on less similar topics. This assumption should not cause too many problem as GEE is generally quite robust to mis-specification of the correlation structure. Did the analysis employ Huber-White “sandwich estimator” for robust standard errors?
3. The additional information on the interrater reliability of the CTS is useful but does not quantify the degree of inter-rater agreement in this study. I would be less concerned about the fact that these were ‘humans using their best judgement, which is a method subject to bias’ than by whether the same person assessed both team work and errors for the scenarios. If the same person (out of the two evaluators) assessed CTS and errors for each team

	scenario and one evaluator is a little more likely to both identify team work deficiencies and spot errors (i.e. is more inclined to score highly on both measures) than the other, then this will induce a correlation. Could you test the effect of ‘rater’ in the model to see this affects the OR for CTS? (This might well have no effect in which case you could just add a line to the text to state this. It would be useful to see the full model in the response to these comments however). 4. In the conclusion section of the abstract add in ‘in simulated scenarios of’ before ‘caring’. 5. As I understand it, one evaluator completed the CTS for each team scenario. Did more than one person score the errors? The statement ‘When there was uncertainty over whether an action may or may not have constituted an error, the team discussed to reach consensus’ suggests more than one person observed the scenarios and rated the number of errors. Prior to this I thought one of the two evaluators scored errors and CTS for each team scenario. Please clarify. 6. Reference 26 given to support the decision to conduct a complete case analysis is ‘Hernán M, Robins J. Per-protocol analyses of pragmatic trials. N Engl J Med. 2017;377(14):1391-8.’ This paper is about the difference between intention to treat and per-protocol analyses in clinical trials. It does not refer to decisions about missing data. I agree that given the low proportion of missing data it is probably desirable to do complete case analysis. As suitable reference might be https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701793/ However there are methods that would allow imputation taking into account team and scenario effects (e.g. multiple imputation by chained equations). I agree that in this situation it is not worth it. Consider revising the discussion so that it is not suggested that it is impossible to use missing data methods because of scenario and team dependencies but that it is not appropriate given the low level of missing data and the missing data mechanism. 7. The sentence ‘This direction of research is important internationally, and may be particularly critical for countries of low economic means, who are impacted heavily by the financial burdens.’ Reads a little oddly and I’m not sure what the message is. Is this saying that improving team work is a relatively low cost intervention to improve patient outcomes and thus useful for low and middle income countries? 8. The CTS is trademarked I believe and up to 2015 this was in the name of the last author. Were any of the authors involved in its development and validation and therefore likely to receive financial benefit from its use? If so I think this should this be noted under the competing interest statement. (If no involvement then ignore) Overall the manuscript is clearly written and makes a useful contribution to this area of research.

REVIEWER	Jesus Montero-Marin University of Zaragoza, Spain
REVIEW RETURNED	15-May-2019

GENERAL COMMENTS	Authors have adequately addressed the questions raised. Thus, I think the manuscript could be publishable.

VERSION 2 – AUTHOR RESPONSE

Reviewer comments	Response
Reviewer 1	
1. Whilst trust is of course part of teamwork is it core? For example you may be better saying ‘leadership, task management and communication’ or make reference to non-technical skills early on e.g. leadership, teamwork, situation awareness and decision making” ? You also use the same terminology later	Thank you for your comments, we have updated the word “trust” to “team management” here and in the introduction.
2. Perhaps the term ‘correlated’ would best be avoided here as well and just keep it as a statistical terms?	We have changed the word “correlated” to “associated” to avoid confusion with statistical terms.
3. Setting – you have missed the fact that there were four scenarios – that you mention later	Thank you, we have updated the abstract to reflect the fact that there are 4 unique scenarios.
4. CTS needs describing in full terms in abstract	Thank you, we have added a sentence to describe the CTS.
5. Please could you define what you mean by ‘in situ” – just a quick statement early on as you do clarify later	We have added clarification for in situ.
6. Was the study ethically approved? Perhaps a statement in the text earlier – not just the statement at the end	The study was ethically approved and this was stated in the first sentence of the participants section “The Oregon Health & Science University Institutional Review Board (IRB00006942) approved the study and all subjects signed both study and video consent prior to participation.”
7. Page 8 suggest that a description of the simulations would go well in table and please use full terms not abbreviations. Also you do not indicate how they were developed e.g. an expert clinical team with 20 years experience each etc etc etc. What was included in the pilot – how many runs did you do? How were the actors briefed/trained for the scenarios?	We added a brief description in text.
8. I don’t think you have defended why you have used CTS over other tools especially when the trials with this tool	We have added our justification for selection of CTS as requested.

have been in obstetrics and I note that the systematic review you mention has in fact missed a number of rating tools.	
9. Many thanks for the additional changes you have made	Thank you
Reviewer 3	
5. In the abstract I would suggest using the term 'generalized estimating equations' rather than 'multivariate regression model' as it gives the reader more information about how the clustering of observations by teams was handled in the analysis and the estimation method. If I read 'generalized estimating equations' then I know this is a marginal model using quasi-likelihood estimation.	Thank you, we have changed the wording accordingly.
6. The use of the exchangeable correlation structure for the GEE assumes that the correlation between errors on the scenarios are the same for all pairs of scenarios ie those that are performed close together in time and those further apart and those on more similar topics and those on less similar topics. This assumption should not cause too many problem as GEE is generally quite robust to mis-specification of the correlation structure. Did the analysis employ Huber-White "sandwich estimator" for robust standard errors?	Yes, we chose to use the exchangeable correlation structure for the reasons you describe. To make our assumption explicit, we have added a sentence describing our rationale to the Methods – Statistical Analysis section. We did not use the Huber-White estimator and instead used the COVB option in SAS's GENMOD to generate our standard errors; this generalized models method is related to the Hessian (2nd derivative) matrix of the likelihood function.
7. The additional information on the interrater reliability of the CTS is useful but does not quantify the degree of inter-rater agreement in this study. I would be less concerned about the fact that these were 'humans using their best judgement, which is a method subject to bias' than by whether the same person assessed both team work and errors for the scenarios. If the same person (out of the two evaluators) assessed CTS and errors for each team scenario and one evaluator is a little more likely to both identify team work deficiencies and spot errors (i.e. is more inclined to	In more than one-half of the simulations, two subject matter experts (raters) used the CTS to evaluate team performance during and immediately after observing a simulation scenario. If a difference in scoring occurred, the two raters discussed the particular concern and reached a single consensus score. Similarly, the clinical errors observed (e.g., failure to start ECG monitoring, or incorrect medication dose calculation) were noted and if different, reconciled by discussion. Therefore it is not possible post-hoc to assess inter-rater agreement for the CTS scores or the errors. We respect the concern for the potential correlation between observing a clinical error and the rating of team-work. However, the constructs of

score highly on both measures) than the other, then this will induce a correlation. Could you test the effect of 'rater' in the model to see this affects the OR for CTS? (This might well have no effect in which case you could just add a line to the text to state this. It would be useful to see the full model in the response to these comments however).	teamwork and communication on the CTS are explicitly defined and distinct from the procedural errors observed in patient treatment.
8. In the conclusion section of the abstract add in 'in simulated scenarios of' before 'caring'.	We have updated the wording as requested.
9. As I understand it, one evaluator completed the CTS for each team scenario. Did more than one person score the errors? The statement 'When there was uncertainty over whether an action may or may not have constituted an error, the team discussed to reach consensus' suggests more than one person observed the scenarios and rated the number of errors. Prior to this I thought one of the two evaluators scored errors and CTS for each team scenario. Please clarify.	We have clarified that scenarios were scored by 2 raters. Thank you.
10. Reference 26 given to support the decision to conduct a complete case analysis is 'Hernán M, Robins J. Per-protocol analyses of pragmatic trials. N Engl J Med. 2017;377(14):1391-8.' This paper is about the difference between intention to treat and per-protocol analyses in clinical trials. It does not refer to decisions about missing data. I agree that given the low proportion of missing data it is probably desirable to do complete case analysis. As suitable reference might be. However there are methods that would allow imputation taking into account team and scenario effects (e.g. multiple imputation by chained equations). I agree that in this situation it is not worth it. Consider revising the discussion so that it is not suggested that it is impossible to use missing data methods because of scenario and team dependencies but that it is not appropriate given the low level of	Thank you for bringing this need for clarification to our attention. We agree that the reference you provided more directly addresses the issue and better supports our intended statements. As you recognized, only 6 of the 176 cases had missing variables and were dropped from our analysis. This is a relatively small proportion and their exclusion from analysis is unlikely to bias our findings in a substantive way. We agree that creating an imputed data set and conducting comparison analyses would likely not change our findings, and is not worth the effort.

missing data and the missing data mechanism.	
11. The sentence ‘This direction of research is important internationally, and may be particularly critical for countries of low economic means, who are impacted heavily by the financial burdens.’ Reads a little oddly and I’m not sure what the message is. Is this saying that improving team work is a relatively low cost intervention to improve patient outcomes and thus useful for low and middle income countries?	Yes, we have updated the wording to read more clearly. Thank you.
12. The CTS is trademarked I believe and up to 2015 this was in the name of the last author. Were any of the authors involved in its development and validation and therefore likely to receive financial benefit from its use? If so I think this should this be noted under the competing interest statement. (If no involvement then ignore)	No authors are receiving financial benefits from the use of the CTS. We have added text to make it clear that it is free.
13. Overall the manuscript is clearly written and makes a useful contribution to this area of research.	Thank you
Reviewer 4	
7. Authors have adequately addressed the questions raised. Thus, I think the manuscript could be publishable.	Thank you

VERSION 3 – REVIEW

REVIEWER	Catherine Best Faculty of Health Sciences and Sport, University of Stirling UK
REVIEW RETURNED	12-Sep-2019

GENERAL COMMENTS	The authors have responded to the previous comments on the paper. The section in the paper on statistical analysis appears to have some typos. It says ' .... Our choice to use the exchangeable correlation structure also should be robust to errors in adjacent simulation scenarios and comparisons to those farther apart in time on the testing day, as well as scenarios that may share similar characteristics (e.g., same age of pediatric patient). assumes that the correlation between errors. ' Please edit this.

	I am confused by the reponse to my comment on robust standard errors. Robust (also known as empirical, sandwich or Huber-White) standard errors are the default for the SAS GENMOD GEE. Specification of COVB option is not necessary. See https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_genmod_sect008.htm State in the methods that robust standard errors were employed for readers not familiar with SAS defaults.