PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The utility of the number needed to treat in pediatric hematological cancer randomized controlled treatment trials: A systematic review
AUTHORS	Hasan, Haroon; Goddard, Karen; Howard, A. Fuchsia

VERSION 1 – REVIEW

REVIEWER	Diogo Mendes AIBILI - Association for Innovation and Biomedical Research on Light and Image, Portugal
REVIEW RETURNED	19-Mar-2018

GENERAL COMMENTS	This study is very interesting and relevant, namely the approach used to interpret clinical significance of effect size differences found in clinical trials, using NNT estimates. Overall, the article is well structured and present tables and figures that increase readability and improve the interpretation of findings. Further a well-defined protocol was provided and followed to perform the literature search and the systematic review. A validated methodology (by Altman and Anderson) was used to obtain NNT estimates and that methodology is appropriate given the type of outcomes analyzed in the present study. Nevertheless, a few suggestions are made below. Major suggestion: The discussion should address the following: how the authors do interpret the fact that despite some comparisons show that tested interventions provide possibly clinically significant effects (only one definite clinically significant), a lot of other comparisons suggest that there is uncertainties (and even unfavorable results) regarding the clinical significance of interventions. What is the implication for clinical practice? Are those latter interventions not effective? Minor suggestions: Line 175-177: Suggestion: provide absolute numbers, not only percentages [e.g. lymphoma (N=15; 23%)]. Table 1: I believe the sentence “Percentages due not sum to a 100% due to rounding and randomized questions with absolute risk reduction equal to zero are excluded” should be deleted, since all calculations seem to have been performed against 65 randomized questions from 48 studies (i.e. after the exclusion of 2 studies with ARR = 0). Line 186: I had some difficulty in the interpretation of this sentence. The sentence should not begin with “Although”, as it may lead to a wrong interpretation. Maybe it should be rephrased (for example) as follows: “For randomized questions corresponding to NNTB (i.e. positive effect size), the NNT was less than the threshold NNT in 31% (4/13) ALL and 20% (1/5) AML comparisons”. Than we have the however (Line 187). That is fine.

	Line 188 and Line 190: Suggestion: probably “was greater than” should be used instead of “exceeded” to keep in line with the terminology used in Table 1 and Figure 1. Line 189 or line 192: I would add (for example): “In summary, a definitely clinically significant benefit was found only once, based on NNT comparisons (Freeman et al. 1997)”.

REVIEWER	Selim Corbacioglu Univ. Hospital Regensburg, Germany
REVIEW RETURNED	11-Apr-2018

GENERAL COMMENTS	Dear authors, Your manuscript is well written and adresses an important and interesting aspect of clinical trials that seemed to largely neglected in pediatric oncology trials. Minor comments: 1. Could you comment on the clinical implications post-hoc if NNTs had been implemented? 2. Why are these differences between ALL and AML?

REVIEWER	Prof. Simon Skene University of Surrey, UK.
REVIEW RETURNED	12-Jun-2018

GENERAL COMMENTS	The paper sets out to present the ‘utility’ of the number needed to treat (NNT) in pediatric haematological cancer RCTs of treatment. I feel there are two distinct facets to the paper, which should be more clearly delineated. i) The ‘utility’/usefulness of NNT as a tool/measure for helping to interpret clinical trial results ii) A systematic review covering the reporting of relevant pediatric trials and use of NNT in interpreting the results, or ability to calculate it. Recommendations should then include good practice in the calculation of NNT (or equivalently absolute risk reduction (ARR)) and their presentation in addition to planned trial analyses to aid the interpretation and generalisability of results in practice. Ie NNT is a tool to support the reporting of the primary outcome analysis and not a replacement of it. In the first point, the authors over-interpret CONSORT in ‘recommending’ both ARR and NNT are presented in trial results. This is discussed by Altmann et al [CONSORT statement requires closer examination BMJ. 2002 Dec 7; 325(7376): 1364], and remains the case – see below. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials, fives the following Item 17b. For binary outcomes, presentation of both absolute and relative effect sizes is recommended When the primary outcome is binary, both the relative effect (risk ratio (relative risk) or odds ratio) and the absolute effect (risk difference) should be reported (with confidence intervals), as neither the relative measure nor the absolute measure alone gives

	a complete picture of the effect and its implications… both doctors and lay people tend to overestimate the effect when it is presented in terms of relative risk. Item 21. Generalisability (external validity, applicability) of the trial findings Measures that incorporate baseline risk when calculating therapeutic effects, such as the number needed to treat to obtain one additional favourable outcome and the number needed to treat to produce one adverse effect, are helpful in assessing the benefit-to-risk balance in an individual patient or group with characteristics that differ from the typical trial partici¬pant That is, the importance of presenting the absolute risk difference in addition to relative differences is well made, but NNT itself is regarded only as helpful – particularly due to its ease of interpretation by the clinical community over odds/relative-risk/hazards ratios etc. Another useful reference is Altman, Confidence intervals for the number needed to treat [BMJ 1998;317:1309] which talks about the inherent difficulties of constructing confidence limits for NNTH and NNTB based on ARR and gives good recommendations about how to present such intervals where the confidence interval for the treatment effect crosses zero. The authors talk about the limitations of NNT with regard to generalisability, which are relevant, but they should also note (in discussing its 'utility') the limitations of its calculation within the trial and interpretability – see reference above. It is also true that confidence intervals for NNT are often wide limiting its true value. Having made these points, I do feel the paper could be usefully modified, and the results of the systematic review used to better emphasise and discuss the true 'utility' of NNT in aiding interpretation of (pediatric) trial results. Some specific points. 1. P3 (Limitations) "We excluded a number of trials due to reporting that precluded calculating the numbers needed to treat". This would be a useful number to include in Figure 2. (I do not seem to be able to infer it specifically. Similarly, Only 3750 /5045 records were screened. How do the authors account for the remaining 1295 papers? 2. 50 trials gave rise to 65 randomised questions. Were these factorial trials, co-primary endpoints? ie the treatment of an MCID/sample size argument relates in general only to a primary outcome. Further details would be useful. 3. It is right to focus on the MCID in relation to the treatment effect and its confidence interval, and this is not referred to enough in publications. It should also be recognised however, that the MCID in a sample size may not be universally recognised as so when it comes to generalisability. 4. The calculation of a threshold ARR and hence NNT from the MCID is useful, however due to transformations of scale they may not directly correspond. Are there any useful references with regard to this approach? If not, the authors approach is not

	undermined but attention should be brought to the further need for validation. 5. Figure 1. Useful to include an NNT entirely below zero to indicate NNTH (harm) in the spectrum of clinical significance. 6. Figure 3. The presentation of confidence intervals including negative NNT is confusing. Consider the approach advocated by Altmann (1998) mentioned above. 7. Figure 4. The first box in the No column of the flowchart in step 3 simply repeats the box above. Is this correct?

REVIEWER	Jason Oke University of Oxford, United Kingdom
REVIEW RETURNED	01-Jul-2018

GENERAL COMMENTS	I am afraid to say that I don't think I can recommend this paper for publication. The stated objective "to assess the utility of the number-needed to treat (NNT) to inform decision making in the context of pediatric oncology" is good one but I don't think this study answers this specific question. The authors seemed to have proposed a new metric/method which could be used to inform a decision of whether to adopt a new therapy but I don't think they have any evidence to show whether it is any better than current methods. For me they would to have to compare their approach with another method such as statistical significance and then assess how they differ and assess what would have changed if this approach had been adopted, this could be as simple as a survey to assess whether it really is more acceptable. Their theoretical arguments about the NNT being superior are just that, theoretical. Secondly, and perhaps more importantly, the authors are either not aware of have decided not to discuss the ongoing debate about the validity of NNT as a statistical measure. This debate started with the paper by Hutton in 2002 which criticised the use and misinterpretation of the NNT. The problems with the NNT are nicely summarised at the end of the paper:. "The NNT has poor qualities, and at best conveys only the same information as the ARR. The ARR is an absolute measure in a form which is in common use, and has good statistical properties. Therefore, it appears to us strongly preferable to base both statistical inference and scientific conclusions on the latter. If RCTs and meta-analyses are held up as the gold standard method for obtaining evidence, an unreliable statistic should not be used in interpreting this evidence." It doesn't end with Hutton, this recent paper provide a nice overview of the evidence on the problems (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1115910/) . From the paper: "However, the NNT is associated with specific weaknesses in calculation and interpretation that are not associated with other methods for integrating data from multiple trials. These weaknesses include distortions in its calculation as placebo effects approach treatment effects, with the possibility of infinite values; difficulties in estimating the precision of the NNT particularly for CI calculation; and, contrary to the original intent, difficulties in interpretation." I am sure there are counter-arguments (I do not have these to hand) and I personally think the NNT is a useful way to communicate risk but I don't think it should be used along with the even-more problematic confidence interval as a way to evaluate

	studies when we have more statistically valid measures and I would need much more evidence to convince me of otherwise. Some minor points I found the use of the term “randomised questions” confusing – is this a pediatric oncology terminology, I thought these trials compared cancer therapies. Also, not sure if this is the convention but it is confusing to refer to the lower confidence limit as the higher number e.g. “Although the NNTB is 15, the lower confidence interval is 33 and the upper confidence interval is 10. “ I understand why this is done (because higher numbers are associated with weaker effects but we don’t do this with OR’s in treatment trials.

VERSION 1 – AUTHOR RESPONSE

Reviewer Comments:

Editorial Comments:

We appreciate that reviewer 4's comments are quite negative; however, in light of the more positive comments from the other reviewers, we felt we should give you the opportunity to respond to the criticisms and revise your manuscript appropriately. Please note that we may ask any or all of the reviewers to assess your revised manuscript so urge you to address all comments as thoroughly as possible.

Editorial Requests:

1. Please revise the Strengths and Limitations section (after the abstract) to focus on the methodological strengths and limitations of your study rather than summarizing the results.

Thank you for this feedback. The strengths and limitations section following the abstract have been revised to emphasize the methodological limitations of our study.

2. The search is now out of date. Please update the search to include more recent literature.

The search has now been updated to present (August 2018).

3. A lot of the methodological details have been presented in the supplementary information. Can some of this be moved to the methods section of the manuscript so that it is more visible? We also suggest using sub-headings in the methods section.

Pertinent information from the protocol included in the supplementary file has been moved to the methods section as requested. In addition, sub-headings have been incorporated.

Reviewers' comments:

Reviewer: 1
Reviewer Name: Diogo Mendes
Institution and Country: AIBILI - Association for Innovation and Biomedical Research on Light and Image, Portugal
Competing Interests: None declared

This study is very interesting and relevant, namely the approach used to interpret clinical significance of effect size differences found in clinical trials, using NNT estimates. Overall, the article is well structured and present tables and figures that increase readability and improve the interpretation of findings. Further a well-defined protocol was provided and followed to perform the literature search and the systematic review. A validated methodology (by Altman and Anderson) was used to obtain NNT estimates and that methodology is appropriate given the type of outcomes analyzed in the present study.
Nevertheless, a few suggestions are made below.

Major suggestion:

1. The discussion should address the following: how the authors do interpret the fact that despite some comparisons show that tested interventions provide possibly clinically significant effects (only one definite clinically significant), a lot of other comparisons suggest that there is uncertainties (and even unfavorable results) regarding the clinical significance of interventions. What is the implication for clinical practice? Are those latter interventions not effective?

   We thank the reviewer for this insightful feedback. We have expanded our discussion section to include the paragraph below, which highlights the application of assessing clinical significance when considering clinical decision-making, with emphasis on recommendations on how to evaluate evidence when confidence limits traverse both harm and benefit (i.e., inconclusive clinical evidence).

   "*The aforementioned approach is recommended in light of smaller sample sizes that are often attained in pediatric oncology RCTs and rare disease trials in general, as it allows for assessment of the precision of the treatment effect as well as clinical and statistical significance. This was demonstrated in our study where the majority of randomized questions found to have a NNTB had a NNT greater than the threshold NNT, of which the upper confidence limit was less than or equal to the threshold NNT. If these RCTs were designed with higher power it is possible that definite clinical significance may have been obtained. On the other hand, based on statistical significance these findings would be considered not significant. Since statistical significance does not provide an indication on the size of the treatment effect, one would not be able to discern whether the findings could have possible clinical significance. An assessment of clinical significant therefore requires a summary measure be presented with a CI. By presenting a CI, an assessment can be made of both statistical and clinical significance, which can inform clinical decision-making. Interpreting results from RCTs based solely on statistical significance, without taking into consideration clinical significance, can result in misappraisal of evidence. Using the results of our study as an example, we demonstrated that all randomized questions, for which the NNTB was less than threshold NNT, had a lower confidence limit that was equal to, or greater than, the*

threshold NNT. Although these results were statistically significant, none had definite clinical significance and were only possibly clinically significant. These findings have clinical implications because clinicians often have to make decisions about administering treatments that are not standard of care, and rely on an accurate appraisal of evidence to inform these decisions. Inconclusive evidence, however, does not necessarily infer an ineffective intervention. Rather, inconclusive evidence (when the CI of the NNT crosses infinity as a result of the CI of the ARR crossing 0) infers that the level of clinical significance cannot be determined from the study results. The use of the NNT and the method we describe can be one more tool to support clinical decision-making within this context."

Minor suggestions:

1. Line 175-177: Suggestion: provide absolute numbers, not only percentages [e.g. lymphoma (N=15; 23%)].

   Thank you for the feedback. Absolute numbers have been added where requested in addition to the percentages.

2. Table 1: I believe the sentence "Percentages due not sum to a 100% due to rounding and randomized questions with absolute risk reduction equal to zero are excluded" should be deleted, since all calculations seem to have been performed against 65 randomized questions from 48 studies (i.e. after the exclusion of 2 studies with ARR = 0).

   The aforementioned sentence has been removed from the footnote in Table 1.

3. Line 186: I had some difficulty in the interpretation of this sentence. The sentence should not begin with "Although", as it may lead to a wrong interpretation. Maybe it should be rephrased (for example) as follows: "For randomized questions corresponding to NNTB (i.e. positive effect size), the NNT was less than the threshold NNT in 31% (4/13) ALL and 20% (1/5) AML comparisons". Than we have the however (Line 187). That is fine.

   Thank you for the feedback. The sentence has been revised as suggested.

4. Line 188 and Line 190: Suggestion: probably "was greater than" should be used instead of "exceeded" to keep in line with the terminology used in Table 1 and Figure 1.

   Thank you for the feedback. The sentence has been revised as suggested.

5. Line 189 or line 192: I would add (for example): "In summary, a definitely clinically significant benefit was found only once, based on NNT comparisons (Freeman et al. 1997)".

   We agree with the reviewer that the addition of an example to comment on the number of trials that were found to be definitely clinically significant is worthwhile. However, following updating our systematic review, all randomized questions from factorial designs were omitted as they did not adhere to the requirement of including only parallel study designs. Therefore, the randomized question which corresponded to Freeman et al. 1997 was removed and there were no randomized questions that were definitely clinically significant.

Reviewer: 2
Reviewer Name: Selim Corbacioglu
Institution and Country: Univ. Hospital Regensburg, Germany
Competing Interests: None declared

Dear authors,
Your manuscript is well written and addresses an important and interesting aspect of clinical trials that seemed to largely neglected in pediatric oncology trials.

Minor comments:
1. Could you comment on the clinical implications post-hoc if NNTs had been implemented?

Thank you for this comment. We have added a comment in the discussion to this effect. Please see response to Reviewer #1- Major Suggestion #1.

2. Why are these differences between ALL and AML?

Thank you for the feedback. The objective of our paper was to assess the utility of the NNT as a supportive tool to inform evidence-based decision-making and therefore comparing the metrics within the NNT by disease site is considered out of scope and was only provided for descriptive purposes. The small sample sizes also preclude meaningful comparisons to be drawn.

Reviewer: 3
Reviewer Name: Prof. Simon Skene
Institution and Country: University of Surrey, UK.
Competing Interests: None declared

The paper sets out to present the 'utility' of the number needed to treat (NNT) in pediatric haematological cancer RCTs of treatment.

I feel there are two distinct facets to the paper, which should be more clearly delineated.

i) The 'utility'/usefulness of NNT as a tool/measure for helping to interpret clinical trial results
ii) A systematic review covering the reporting of relevant pediatric trials and use of NNT in interpreting the results, or ability to calculate it.

Recommendations should then include good practice in the calculation of NNT (or equivalently absolute risk reduction (ARR)) and their presentation in addition to planned trial analyses to aid the interpretation and generalisability of results in practice. Ie NNT is a tool to support the reporting of the primary outcome analysis and not a replacement of it.

In the first point, the authors over-interpret CONSORT in 'recommending' both ARR and NNT are presented in trial results. This is discussed by Altmann et al [CONSORT statement requires closer examination BMJ. 2002 Dec 7; 325(7376): 1364], and remains the case – see below.

CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials, fives the following

Item 17b. For binary outcomes, presentation of both absolute and relative effect sizes is recommended

When the primary outcome is binary, both the relative effect (risk ratio (relative risk) or odds ratio) and the absolute effect (risk difference) should be reported (with confidence intervals), as neither the relative measure nor the absolute measure alone gives a complete picture of the effect and its implications… both doctors and lay people tend to overestimate the effect when it is presented in terms of relative risk.

Item 21. Generalisability (external validity, applicability) of the trial findings

Measures that incorporate baseline risk when calculating therapeutic effects, such as the number needed to treat to obtain one additional favourable outcome and the number needed to treat to produce one adverse effect, are helpful in assessing the benefit-to-risk balance in an individual patient or group with characteristics that differ from the typical trial partici¬pant

That is, the importance of presenting the absolute risk difference in addition to relative differences is well made, but NNT itself is regarded only as helpful – particularly due to its ease of interpretation by the clinical community over odds/relative-risk/hazards ratios etc.

Another useful reference is Altman, Confidence intervals for the number needed to treat [BMJ 1998;317:1309] which talks about the inherent difficulties of constructing confidence limits for NNTH and NNTB based on ARR and gives good recommendations about how to present such intervals where the confidence interval for the treatment effect crosses zero.

The authors talk about the limitations of NNT with regard to generalisability, which are relevant, but they should also note (in discussing its ‘utility’) the limitations of its calculation within the trial and interpretability – see reference above. It is also true that confidence intervals for NNT are often wide limiting its true value.

Having made these points, I do feel the paper could be usefully modified, and the results of the systematic review used to better emphasise and discuss the true ‘utility’ of NNT in aiding interpretation of (pediatric) trial results.

We thank the reviewer for their valuable feedback. We have revised the manuscript to reflect that the CONSORT 2010 statement does not recommend the use of the NNT explicitly, but recommends that it may be a helpful tool. We have also revised the discussion (please see excerpt below) to emphasize the utility and additional methodological limitations of the NNT.

"*Scenarios where the NNT results in inconclusive evidence is an evident limitation in the utility of NNT, which has been discussed by Altman[1]. To illustrate, a RCT conducted by Lange et al.[2] assessing 5-year disease free survival in pediatric AML patients in first remission after intensive chemotherapy found a 7.0% (95% CI, -19.8% to 5.8%) absolute decrease associated with the experimental treatment (interleukin-2 infused on days 0-3 and 8-17) compared to the control treatment (no further therapy). The study was powered to detect a 10% difference in 5-year disease free survival, which although not explicitly stated was assumed to be the minimal clinical importance difference and hence corresponds to a threshold NNTB of 10. The resulting NNT of the RCT was -14 (95% CI, -5 to 17) or a NNTH 14 (95% CI, NNTH 5 to NNTB 17). Off first glance, it appears as though the point estimate does not fall within the 95% CI given the disjointed confidence limits. In other studies wherein the CI traverses both harm and benefit the NNT is reported without the CI[3]. In reality, the CI encompasses values from a NNTH of 5 to ∞ and NNTB of 17 to ∞. Plotting the NNT and CI on a forest plot (Figure 3) demonstrates that a NNTH of 14 does fall within the interval and in fact, the interval is continuous. Altman therefore recommended presenting the CI of the NNT as the following (using results from Lange et al. as an example): NNTH 14 (NNTH 5 to ∞ to NNTB 17), which emphasizes continuity."*

Some specific points.

1. P3 (Limitations) "We excluded a number of trials due to reporting that precluded calculating the numbers needed to treat". This would be a useful number to include in Figure 2. (I do not seem to be able to infer it specifically.
Similarly, Only 3750 /5045 records were screened. How do the authors account for the remaining 1295 papers?

We apologize for this error. There were in fact 5,052 duplicates identified which led to 4,151 studies being screened. This error has been revised in Figure 2.

The number which corresponds to this exclusion criterion "*We excluded a number of trials due to reporting that precluded calculating the numbers needed to treat*" is include in Figure 2 (see last bullet in reasons for exclusion at full-text screen stage).

2. 50 trials gave rise to 65 randomised questions. Were these factorial trials, co-primary endpoints? ie the treatment of an MCID/sample size argument relates in general only to a primary outcome. Further details would be useful.

[1] Altman DG. Confidence intervals for the number needed to treat. BMJ 1998;317 doi: 10.1136/bmj.317.7168.1309
[2] Lange BJ, Yang RK, Gan J, et al. Soluble interleukin-2 receptor alpha activation in a Children's Oncology Group randomized trial of interleukin-2 therapy for pediatric acute myeloid leukemia. Pediatric blood & cancer 2011;57(3):398-405. doi: 10.1002/pbc.22966
[3] McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. Annals of internal medicine 1997;126(9):712-20.

We have updated our systematic review to includes studies up to 2018 and have only included parallel RCTs (i.e., factorial design RCTs were omitted). A total of 43 RCTs were included in the analysis of which two RCTs contained two parallel arms with corresponding sample size calculation (please see first excerpt below). We have provided clarity on the unit of analysis being randomized questions in the "analysis" section of the "methods" (please see second excerpt below).

"*A comprehensive literature review was performed using the databases MEDLINE (Via Ovid), EMBASE (via OVID) and Cochrane Childhood Cancer Group Specialized Register (Via CENTRAL) from inception to August 2018 to identify all superiority, parallel group, RCTs in pediatric patients diagnosed with a hematological cancer that assessed an outcome related to survival, relapse or remission and those that reported either confidence intervals (CIs) or standard errors associated with both the experimental and control estimates, or numbers of patients at risk on a Kaplan Meier curve*."

"*A randomized question is defined as an intervention comparison assessing a primary outcome for which a sample size calculation is reported. The NNT was based on the primary outcome and time point as specified in the sample size calculation. In the event that the time point specified in the sample size calculation was not reported, the information was inferred if a Kaplan Meier curve with the number of patients at risk was reported*[4]*. If the aforementioned was not provided, the time point reported in the results was used, and thus, these trials were prone to selective reporting bias. All analyses were conducted based on randomized questions to account for the fact that one study may have more than one parallel group.*"

3. It is right to focus on the MCID in relation to the treatment effect and its confidence interval, and this is not referred to enough in publications. It should also be recognised however, that the MCID in a sample size may not be universally recognised as so when it comes to generalisability.

We completely agree with the reviewer. We alluded to this in the discussion initially and have expanded this point to reflect limitations related to generalizability. Please see below.

"*An additional weakness is that the delta value in the sample size calculation was assumed to be the absolute difference that would provide an effect size that would lead to a change in clinical practice (i.e., minimal clinically important difference), if not explicitly indicated, and a proxy for the threshold ARR and NNT. This assumption, thus would lead to the possibility of effect sizes being chosen that might be more reflective of feasibility as opposed to clinical benefit and this approach and may be limited in terms of generalisability given this is not a universally recognized approach."*

4. The calculation of a threshold ARR and hence NNT from the MCID is useful, however due to transformations of scale they may not directly correspond. Are there any useful references with regard to this approach? If not, the authors approach is not undermined but attention should be brought to the further need for validation.

[4] Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. BMJ 1999;319 doi: 10.1136/bmj.319.7223.1492

The approach of assigning the threshold NNT as equivalent to the inverse of the threshold ARR (which in our study was assumed to be the minimal clinical important difference, irrespective of whether explicitly mentioned) was implemented as a method to account for instances where a threshold NNT was not determined a priori (which in our case applied to all studies). As the reviewer mentions, the ARR may not be equivalent to the NNT due to transformation of scale. The method we propose has not been validate and we acknowledged that the need for validation is required. Please see below for an excerpt from the manuscript, which reflects the aforementioned.

"*Additionally, this assumption implies that the threshold NNT is equivalent to the threshold ARR even though the NNT results in a transformation of scale and is expressed using a unit measured in patients. Therefore, a threshold ARR may not correspond to a minimal clinically important difference in terms of NNT. However, as there were no studies that reported a threshold NNT, our approach represents a feasible method to apply in the absence of a reported threshold NNT. This method is nonetheless not validated and further studies will need to be undertaken to compare whether researchers would equate the minimal clinical important difference in terms of ARR to the NNT.*"

5. Figure 1. Useful to include an NNT entirely below zero to indicate NNTH (harm) in the spectrum of clinical significance.

We thank the reviewer for this comment and have made this revision to Figure 1.

6. Figure 3. The presentation of confidence intervals including negative NNT is confusing. Consider the approach advocated by Altmann (1998) mentioned above.

We thank the reviewer for this comment. We have modified our presentation of confidence intervals to match the reporting of negative NNT (i.e., number needed to harm), whereby the values have been expressed as absolute numbers and values left of infinity have been denoted as a "number needed to harm" and "number needed to benefit" respectively. We have also revised the data to the right of the column to reflect the reporting of the confidence interval as recommended by Altman (1998) – (e.g., a number needed to treat of 10 (95% CI, -20, 4) would be reported as NNTB 10 (NNTH 20 to ∞ to NNTB 4; which emphasizes continuity).

7. Figure 4. The first box in the No column of the flowchart in step 3 simply repeats the box above. Is this correct?

We apologize for this error and this has been revised.

Reviewer: 4
Reviewer Name: Jason Oke
Institution and Country: University of Oxford, United Kingdom
Competing Interests: None declared

I am afraid to say that I don't think I can recommend this paper for publication. The stated objective "to assess the utility of the number-needed to treat (NNT) to inform decision making in the context of pediatric oncology" is good one but I don't think this study answers this specific question.

The authors seemed to have proposed a new metric/method which could be used to inform a decision of whether to adopt a new therapy but I don't think they have any evidence to show whether it is any better than current methods. For me they would to have to compare their approach with another method such as statistical significance and then assess how they differ and assess what would have changed if this approach had been adopted, this could be as simple as a survey to assess whether it really is more acceptable. Their theoretical arguments about the NNT being superior are just that, theoretical.

Secondly, and perhaps more importantly, the authors are either not aware of have decided not to discuss the ongoing debate about the validity of NNT as a statistical measure. This debate started with the paper by Hutton in 2002 which criticised the use and misinterpretation of the NNT. The problems with the NNT are nicely summarised at the end of the paper:.
"The NNT has poor qualities, and at best conveys only the same information as the ARR. The ARR is an absolute measure in a form which is in common use, and has good statistical properties. Therefore, it appears to us strongly preferable to base both statistical inference and scientific conclusions on the latter. If RCTs and meta-analyses are held up as the gold standard method for obtaining evidence, an unreliable statistic should not be used in interpreting this evidence."

It doesn't end with Hutton, this recent paper provide a nice overview of the evidence on the problems (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1115910/) . From the paper:
"However, the NNT is associated with specific weaknesses in calculation and interpretation that are not associated with other methods for integrating data from multiple trials. These weaknesses include distortions in its calculation as placebo effects approach treatment effects, with the possibility of infinite values; difficulties in estimating the precision of the NNT particularly for CI calculation; and, contrary to the original intent, difficulties in interpretation."

I am sure there are counter-arguments (I do not have these to hand) and I personally think the NNT is a useful way to communicate risk but I don't think it should be used along with the even-more problematic confidence interval as a way to evaluate studies when we have more statistically valid measures and I would need much more evidence to convince me of otherwise.

We thank the reviewer for his insightful criticism.

The intent of our paper was not to provide evidence that the NNT is as a measure is superior to any other measures, but to assess the utility of this measure as a supportive tool in clinical decision-making. Similar to the paper referenced by the Reviewer (Katz et al 2015; Journal of Pain), we echo the following: "*Moreover, the NNT taken alone does not summarize all necessary information for the clinician to make informed decisions regarding treatment*". This sentiment is also emphasized in the CONSORT 2010 statement, "*For both binary and survival time data, expressing the results also as the number needed to treat for benefit or harm can be helpful…Measures that incorporate baseline risk when calculating therapeutic effects, such as the number needed to treat to obtain one additional favourable outcome and the number needed to treat to produce one adverse effect, are helpful in assessing the benefit-to-risk balance in an individual patient or group with characteristics that differ from the typical trial participant*."

Although, we acknowledge the arguments described by Hutton (2002), we concur with the rebuttal of Altman & Deeks – selected excerpts: "*The NNT was proposed not for computation but for the translation of research results to patients. Hence, many of Hutton's criticism do not relate to the purpose which the measure was conceived and proposed... The NNT is a way of presenting results, not of analysing data, and this is how it is used in general. Arguments about the distributional properties, bias and the like largely miss the point. We see the NNT simply as a way of re-expressing the ARR, and likewise the CI for the NNT as an alternative way of depicting the CI for the ARR (Altman, 1988). The CI for the NNT can be obtained from any CI fro the ARR, and we agree that the best methods for constructing a CI for the ARR should be used (Newcombe and Altman, 2000)".*

The method of using the NNT in combination with the confidence interval has been highlighted by Guyat et al. 1995[5]. Additionally, a method to integrate clinical significance with statistical significance has been detailed by Man-Son-Hing et al. 2002[6]. Both papers were used to propose the method we described in our systematic review. Our method, similar to the aforementioned authors mentioned, combines an assessment of clinical significance with statistical significance as statistical significance can be discerned through visualization using a forest plot or assessing whether the CI of the ARR includes zero. Both complement each other and are by no means replacements for each other; however an assessment of clinical significance can provide information that would be gleaned from statistical significance while the vice-versa is not true. We have commented on the application of this method specifically in the excerpt provided the response to Reviewer #1- Major Question #1.

The NNT provides an inherent advantage that other summary measures in that it acts as supportive tool to assign relative values to outcomes and relate them to costs, which other measures cannot.

Altman has provided an effective method to report sensible confidence intervals, which we have adapted in the forest plot we report. Our forest plot aims to ensure that the emphasis lies on the absolute risk reduction in terms of scale and therefore both the numbers needed to benefit and harm are reported as well as the absolute risk reduction, "*We need to remember the absolute risk reduction scale when trying to interpret the number needed to treat and its confidence interval – Altman 1998*". In instances where the ARR CI crosses 0, the NNT will be infinity, which is one of the criticisms of reporting the NNT with a CI. Altman has recommended a simple way to aid interpretation where for example a NNT of 10 (95% CI -20, 4) would be reported as NNTB 10 (NNTH 20 to ∞ and NNTB 4 to ∞, which can be shorten to NNTH 20 to ∞ to NNTB), which emphasizes continuity. We have adopted this approach and believe it provides a simple method to report a sensible confidence interval.

We have expanded our discussion to include a comment on this controversial topic. Please see below.

[5] Guyatt GH, Sackett DL, Sinclair JC, et al. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. Jama 1995;274(22):1800-4. [published Online First: 1995/12/13]
[6] Man-Son-Hing M, Laupacis A, O'Rourke K, et al. Determination of the clinical importance of study results. Journal of general internal medicine 2002;17(6):469-76. [published Online First: 2002/07/23]

"*Scenarios where the NNT results in inconclusive evidence is an evident limitation in the utility of NNT, which has been discussed by Altman[7]. To illustrate, a RCT conducted by Lange et al.[8] assessing 5-year disease free survival in pediatric AML patients in first remission after intensive chemotherapy found a 7.0% (95% CI, -19.8% to 5.8%) absolute decrease associated with the experimental treatment (interleukin-2 infused on days 0-3 and 8-17) compared to the control treatment (no further therapy). The study was powered to detect a 10% difference in 5-year disease free survival, which although not explicitly stated was assumed to be the minimal clinical importance difference and hence corresponds to a threshold NNTB of 10. The resulting NNT of the RCT was -14 (95% CI, -5 to 17) or a NNTH 14 (95% CI, NNTH 5 to NNTB 17). Off first glance, it appears as though the point estimate does not fall within the 95% CI given the disjointed confidence limits. In other studies wherein the CI traverses both harm and benefit the NNT is reported without the CI[9]. In reality, the CI encompasses values from a NNTH of 5 to ∞ and NNTB of 17 to ∞. Plotting the NNT and CI on a forest plot (Figure 3) demonstrates that a NNTH of 14 does fall within the interval and in fact, the interval is continuous. Altman therefore recommended presenting the CI of the NNT as the following (using results from Lange et al. as an example): NNTH 14 (NNTH 5 to ∞ to NNTB 17), which emphasizes continuity.*

We strongly encourage plotting the ARR and the NNT on a forest plot simultaneously because the NNT is simply a method of re-expressing the ARR and supports the interpretation of the ARR. As the NNT is a relative measure it should always be accompanied by the absolute measure, the ARR[10]. Additionally, the utility of the NNT is inherently reliant on three major areas: baseline risk, the outcome and the time point. In order for the NNT from an RCT demonstrating a NNTB to have utility, the patient population of interest should share a similar baseline risk because the desired treatment effect may be overestimated and thus the NNTB may by underestimated. Outcomes related to event free survival often differ in what is considered an event and thus it is critical to ensure that the NNTB being applied to the population of interest is identical in terms of the outcome in question. Numerous studies have demonstrated how the NNT varies with time and thus, comparability in time points is critical to ensure accurate interpretation of the NNT to a population of interest[11]. Lastly, criticisms of the statistical properties of the NNT have been highlighted by Hutton et al. and Katz et al[12]. We agree with Altman & Deeks response to these criticisms in that the NNT was designed for translation of research results and therefore arguments related to computation and its distribution properties are of less relevance. The NNT is simply a metric to re-express the ARR and therefore should be viewed as a measure to support the interpretation of the ARR."

Some minor points

I found the use of the term "randomised questions" confusing – is this a pediatric oncology terminology,

[7] Altman DG. Confidence intervals for the number needed to treat. BMJ 1998;317 doi: 10.1136/bmj.317.7168.1309

[8] Lange BJ, Yang RK, Gan J, et al. Soluble interleukin-2 receptor alpha activation in a Children's Oncology Group randomized trial of interleukin-2 therapy for pediatric acute myeloid leukemia. Pediatric blood & cancer 2011;57(3):398-405. doi: 10.1002/pbc.22966

[9] McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. Annals of internal medicine 1997;126(9):712-20.

[10] Moher D, Hopewell S, Schulz KF. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340 doi: 10.1136/bmj.c869

[11] Suissa S. The number needed to treat: 25 years of trials and tribulations in clinical research. Rambam Maimonides Med J 2015;30

Suissa D, Brassard P, Smiechowski B, et al. Number needed to treat is incorrect without proper time-related considerations. J Clin Epidemiol 2012;65 doi: 10.1016/j.jclinepi.2011.04.009

McAlister FA. The "number needed to treat" turns 20—and continues to be used and misused. CMAJ 2008;179 doi: 10.1503/cmaj.080484

Citrome L, Ketter TA. When does a difference make a difference? Interpretation of number needed to treat, number needed to harm, and likelihood to be helped or harmed. Int J Clin Pract 2013;67 doi: 10.1111/ijcp.12142

[12] Hutton JL. Number needed to treat: properties and problems. J R Stat Soc A Stat Soc 2000;163 doi: 10.1111/1467-985x.00175

Hutton JL. Number needed to treat and number needed to harm are not the best way to report and assess the results of randomised clinical trials. British journal of haematology 2009;146(1):27-30. doi: 10.1111/j.1365-2141.2009.07707.

Katz N, Paillard FC, Van Inwegen R. A review of the use of the number needed to treat to evaluate the efficacy of analgesics. The journal of pain : official journal of the American Pain Society 2015;16(2):116-23. doi: 10.1016/j.jpain.2014.08.005

I thought these trials compared cancer therapies.

We use the term randomized questions to account for the fact that some superiority parallel RCTs do have more than one parallel group with corresponding sample size calculations and hence more than one randomized question. Therefore, it is appropriate to conduct the analysis based on all randomized questions with a sample size calculation. The methods section has been revised to provide clarity by providing a definition of a randomized question. Please see below.

"A randomized question is defined as an intervention comparison assessing a primary outcome for which a sample size calculation is reported."

Also, not sure if this is the convention but it is confusing to refer to the lower confidence limit as the higher number e.g.

"Although the NNTB is 15, the lower confidence interval is 33 and the upper confidence interval is 10. "

I understand why this is done (because higher numbers are associated with weaker effects but we don't do this with OR's in treatment trials.

We thank the reviewer for this comment. We have revised the order of the confidence limits to match traditional reporting in the literature.

VERSION 2 – REVIEW

REVIEWER	Diogo Mendes AIBILI - Association for Innovation and Biomedical Research on Light and Image, CHAD - Centre for Health Technology Assessment and Drug Research, Coimbra, Portugal
REVIEW RETURNED	28-Aug-2018

GENERAL COMMENTS	Limitations section: Second topic - First sentence: delete one "that".

REVIEWER	Prof. Simon Skene University of Surrey, UK
REVIEW RETURNED	03-Sep-2018

GENERAL COMMENTS	The authors have addressed each of the points I made, I hope resulting in a paper which now better argues the case for NNT as a tool for aiding the interpretation of results from RCTs whilst recognising the inherent limitations of the approach. The case is now well balanced with supporting evidence from a systematic review from trials in paediatric oncology.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Diogo Mendes

Institution and Country: AIBILI - Association for Innovation and Biomedical Research on Light and Image, CHAD - Centre for Health Technology Assessment and Drug Research, Coimbra, Portugal

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

Please leave your comments for the authors below

Limitations section: Second topic - First sentence: delete one "that".

We deleted the redundant “that” in this sentence.

Reviewer: 3

Reviewer Name: Prof. Simon Skene

Institution and Country: University of Surrey, UK

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

Please leave your comments for the authors below

The authors have addressed each of the points I made, I hope resulting in a paper which now better argues the case for NNT as a tool for aiding the interpretation of results from RCTs whilst recognising the inherent limitations of the approach. The case is now well balanced with supporting evidence from a systematic review from trials in paediatric oncology.

Thank you for your feedback.