Author's response to reviews

Title: Constraint-induced movement therapy: trial sequential analysis applied to Cochrane Collaboration systematic review results.

Authors:

Greta Castellini (gre.caste@gmail.com)
Silvia Gianola (silvia.gianola@gmail.com)
Rita Banzi (rita.banzi@marionegri.it)
Davide Corbetta (corbetta.davide@alice.it)
Roberto Gatti (gatti.roberto@hsr.it)
Valeria Sirtori (sirtori.valeria@hsr.it)
Christian Gluud (cgluud@ctu.rh.dk)
Lorenzo Moja (lorenzo.moja@unimi.it)

Version: 2
Date: 15 October 2014

Author's response to reviews: see over
Dear Editor,

We would like to thank the Trials Journal editor and the reviewers for their interest in our manuscript (Constraint-induced movement therapy: trial sequential analysis applied to Cochrane Collaboration systematic review results. MS: 1210174271166103) and for their helpful comments. We hope we have addressed all of their concerns.

With best wishes,
Greta Castellini.

On behalf of all authors

Reviewer's report 1

Title: Constraint-induced movement therapy: trial sequential analysis applied to Cochrane Collaboration systematic review results.

Version: 1 Date: 18 August 2014
Reviewer: Charlie Goldsmith

Reviewer's report:
The clinical examples provided here show a method that has been published already; however, is not common in Cochrane Reviews as yet and could be in the future. However, the choice of a single study to establish the MID could be enhanced by surveying more widely to see whether any other authors have used other values. The scoring done by the authors does not indicate whether the Risk of Bias rating and data extraction were done by multiple raters and what level of agreement they obtained as this would help the credibility of the data the authors use.

Here are some more specific recommendations.

1. P(age) 5, p(aragraph) 1, l(ine) 8. Suggest deleting the [s] to read [error]. Also P 6, p 2, l 3. Also P 7, p 1, l 1 and 2.
   Author answer
   Now –s is deleted.

2. P 5, p 2, l 1. It is common to use 8 components for specifying a sample size, and the authors have not justified why they chose fewer.
   Author answer
   It is common to use eight components; several ways exist to determine the sample size. I have updated the sentence based on the CONSORT Statement declaration. “Elements of the sample size calculation are (1) the estimated outcomes in each group (which implies the clinically important target difference between the intervention groups); (2) the alfa (type
I) error level; (3) the statistical power (or the beta (type II) error level); and (4), for continuous outcomes, the standard deviation of the measurements.”

3. P 5, p 2, l 2. In many locations the apparently Greek symbols such as alpha and beta have been replaced by square boxes in the text, so it is not possible to tell if they have been done properly. Others will not be noted. None of the equations on P 10 were printed for checking.

Author answer
We replaced the Greek symbols with “alpha” and “beta” in the text. Equations have also been converted into figures.

4. P 6, p 2, l 2. Suggest adding the text [, indeed a feature of Cochrane reviews] after [data].

Author answer
We added the text.

5. P 6, p 2, l 2. Suggest inserting [statistically] between [A] and [significant].

Author answer
We added the word.


Author answer
Done.

7. P 7, p 2, l 8. This would be a good place to mention whether the Risk of Bias as well as data extraction were done by more than one rater and a measure of their agreement. Also how was disagreement resolved.

Author answer
We mentioned the process of independent extraction in the Method section (page 8, paragraph 1, lines 2-3) and the independent risk of bias assessment in the Method section (page 8, paragraph 2, lines 4-5). Please refer to the correction in the text.


Author answer.
Done.

9. P 7, p 2, l 9. Suggest deleting [valuable]. This has not yet become a common tool, and the authors have not cited a source of its value.

Author answer.
We deleted the word.

10. P 7, p 3. For the search example being used, was there an estimate of how complete the searches were to obtain the RCTs used? If so, report it. A reference to the original article might suffice if the articles [4, 21] contain such a statement.

Author answer.
We included a sentence about searches and references to previous articles. Furthermore, the complete search strategies was added to the appendix.

11. P 8, p 2, l 1. Presumably the word [randomized] should be replaced by the

*Author answer*

Citation provided.

13. P 9, p 2. The articles were presumably ordered by year. What was done to ensure the order you used was indeed correct as there are certainly some duplicate years in your examples.

*Author answer*

The adjusted significance boundaries for the cumulative Zcurve were constructed under the assumption that significance testing may have been performed each time a new trial was added to the meta-analysis. For this reason, trials were added according to the year. If, in the meta-analysis data sets, there were some years when more than one trial was published, for these years, we analysed trials in alphabetical order, according to the last name of the first author.


*Author answer*

Done.

15. P 11, p 4, l 3. Did you consider a conversion from the Barthel to the FIM to use the trial that was excluded?

*Author answer*


FIM has 18 items (with 5 of them related to cognitive tasks) while the Barthel Index has only 10 items on daily activities: the conversion from Barthel to FIM could be challenging because of the absence of cognitive items in the Barthel Index. For this reason, it is more likely to convert FIM to Barthel because the same daily activities tasks. We should have done a statistical correlation between items until converting the two scales. We decided to not consider studies with Barthel Index measure scale in the analysis.

16. P 11, p 4, l 7. Were there other trials that used an MID? If so they might have been chosen as well and this might have led to other recommendations. Was a search of the literature for the FIM MIDs and discuss that impact of its choice? This should be discussed in the paper. Also for the second MID choice as well.

*Author answer*

We studied in depth if there were articles that used FIM MIDs, and we did not find other trials that used a MID. The choice of MID is usually correlated with the context and the population examined. According to some authors, who referred in the article, there are two methods to establish the MID. First is called anchor based approach, which is used when the MID derives from a study judged as interpretable, valid, and generalizable. The second is the distribution-
based approach: the MID can be estimated by multiplying the effect size by the pooled standard deviation between groups. Since we found a single trial that established MID for FIM and its population did not follow our criteria (Benito 2006), we decided to get the distribution-based approach. Otherwise, we found the MID for ARAT scale adapted for our population. The peer reviewer can find a clearer answer on MID choice in Box 1 and page 12, paragraph 2, line 6.

17. P 13, p 1, l 2. The issue of Power is not part of the Risk of Bias form. Why was it introduced here? This whole approach is one reason why Cochrane reviews are encouraged. It could rescue small sample size studies to make a proper recommendation for patients and clinicians. Indeed the examples suggest even the method you have shown leads to a cautious recommendation.

Author answer
We agree with the peer review that the issue of power is not part of the risk of bias. The concept of power in this line is linked to the fact that trials are usually underpowered with a small sample size: in rare occasions, small sample size studies could be sufficient to develop recommendations. Furthermore, results could have been influenced by systematic errors.
Page 13, paragraph 2, line 1: sentence clarified.

Our cautious recommendation is linked to the importance of considering that inference drawn about the conclusiveness of a meta-analysis, and so of a TSA, can only be generalized to the patient population for which the a priori minimally important difference apply.

18. P 13, p 3, l 2. Suggest a rewrite as [… the null effect, while a few …].

Author answer
Done.

19. P 13, p 2, l 9. Repeated trials on a topic may modify the Inclusion/Exclusion criteria for patients and this might induce a change in the effect sizes of the studies being combined. Was this checked? If so, what impact did it have? It could be a reason for the heterogeneity of the second example.

Author answer
Trials included in our meta-analysis included patients from the same population groups, similar regimens of the intervention, same study designs, but variable methodological qualities. It is natural to expect an additional degree of variation in meta-analysis data compared to data from a single trial, in particular when trials have a small sample size. Because increased variation can decrease the precision of results, information size considerations must incorporate all sources of variation in a meta-analysis, including heterogeneity. One approach for incorporating heterogeneity in information size considerations is to calculate the required information size adjustment for heterogeneity. We calculated the AF (adjustment factor= 1/ (H)) on the basis of heterogeneity; If major heterogeneity is expected, then H may become 75% and AF would be estimated to 4.00.

20. P 13, p 2, l 12. Suggest adding the text [and the later modest effects] after [impressive].

Author answer
21. P 14, p 1, l 7. Where is the literature search to support this claim? It might be better to state something like [As far as the authors know, this study is the first ...].
   
   **Author answer**
   
   Thank you for the advice, we added the statement.

22. P 14, p 2, l 6. Since all bias is systematic, suggest dropping the word.
   
   **Author answer**
   
   We dropped the word.

23. P 14, p 2, l 7. Suggest deleting [only] as it implies an unstated expectation.
   
   **Author answer**
   
   Done.

   
   **Author answer**
   
   Done.

   
   **Author answer**
   
   Done.

26. P 17, R(eference) 5, l 2. Add more to the location of where the Handbook can be found.
   
   **Author answer**
   
   Done.

27. P 17, R 8, l 2. Trials likes to publish the first 30 authors before using et al; so add some more authors. Also P 23, R 57, l 2.
   
   **Author answer**
   
   Done.

28. A random sample of 10 Rs was checked for citation accuracy. For the most part these Rs did not have an issue number that would help any reader trying to find the R. P 18, R 10, l 4. Insert [(1)] after [38].
   
   **Author answer**
   
   Done.

29. P 18, R 13, l 2. Add more to this R as to where it might be obtained in Copenhagen.
   
   **Author answer**
   
   Done, we added the website.

30. P 18, R 14 appears to be correct.
   
   **Author answer**
   
   Ok

31. P 18, R 16, l 1. The seventh author has initials [GH]. See other Rs. Also P 20, R 28, l 1.
Author answer
Done.

32. P 19, R 21, l 2. Add [(4)] after the year.
Author answer
Done.

33. P 19, R 24, l 3. Add more to where this can be found.
Author answer
Done.

34. P 19, R 27, l 2. Insert [(3)] after [35].
Author answer
Done.

35. P 20, R 32, l 3. Insert [1)] after [63].
Author answer
Done.

Author answer
Done.

37. P 22, R 51, l 3. Insert [(12) after [31].
Author answer
Done.

Author answer
Done.

Author answer
Done.

40. P 23, R 60, l 3. Insert [(6)] after [92].
Author answer
Done.

41. P 24, R 64, l 2. Insert [(1)] after the year.
Author answer
Done.

42. P 25, Fig 3. The colour did not come thru on the graphs when printed in black and white, suggest using different plotting symbols. On l 4, replace [P] by the Greek [alpha] and on l 6, replace [nhil] by [null]. Also P 26, Fig 5.
Author answer
We replaced the words suggested.
Figure 3 and 5 were changed in black and white.
43. P 25, Fig 4, l 4. Suggest adding the text [was high!] after [85%]. Most analysts would search for a reason for the heterogeneity. Did you? It should be reported as to what was tried for an explanation.

Author answer
We replaced the words suggested.
Statistical heterogeneity is a consequence of clinical or methodological diversity, or both, among the studies. In this second example, the heterogeneity was high due to methodological diversity because the majority of trials were of high risk according to the Risk of Bias. We have taken into consideration the suggestion of searching for a reason for the heterogeneity; it will be included in the update of the Cochrane Review (in progress).

44. P 26, l 4. Rewrite [nineth] as [ninth].

Author answer
We replaced the word suggested.

45. Fig 1. Was there a way to convert the different scale for the trial excluded for ARAT? See 15.

Author answer
The answer is in the point 15.

46. Fig 2 and 4. Suggest citing RevMan for these.

Author answer
Done.

47. Fig 3. The horizontal line uses [141] while the footnote to the graph uses [142]. Should these be the same?

Author answer
We checked this and the figure is correct. The software always defines the number of patients reached at the earliest study to the information size line.

48. Fig 4. The text below the graph for subgroups is a duplicate and could be left off.

Author answer
Done.

49. Fig 3 and 5. Cite how these were drawn.

Author answer
We used the TSA software that made the graph. The software is cited in the method section.
Reviewer’s report 2

Title: Constraint-induced movement therapy: trial sequential analysis applied to Cochrane Collaboration systematic review results.

Version: 1  Date: 25 July 2014

Reviewer: Graziella Filippini

Reviewer’s report:

Major revision

1) In this review the inconclusive conclusion on CIMT efficacy resulted from analyses of the cumulative Z-scores for both outcomes (FIM and ARAT) disregarding the risk of bias of the included studies. I suggest to use GRADE to define the quality of evidence (not risk of bias alone), conduct a sensitivity analysis including studies of high/moderate quality level according to GRADE and consider results of the sensitivity analysis to reach conclusion.

Author answer
We appreciated the suggestion and we are going to add a sensitivity analysis according to GRADE in the update (in progress) of the Cochrane Review, in order to define an overall quality of evidence on CIMT. In this paper, we intended to demonstrate trial sequential analysis on a sample of randomized controlled trials, performing a quality evaluation of each trials.

2) The Authors should include the complete search strategies in the method section.

Author answer
We decided to report the complete search strategies in the Appendix.

Minor revision

1) Risk of bias assessment
“We used all Cochrane Handbook domains for assessment of randomized”. This phrase is not completed.
“Any included trial was rated at low risk of bias in all dimensions”. This phrase is unclear. I suggest to delete it.
Did they use Cochrane criteria or Hemmingsen criteria for assessing risk of bias?
The authors should clarify.

Author answer
The sentence “We used all Cochrane Handbook domains for assessment of randomized” is now “We used all Cochrane Handbook domains for assessment of Risk of Bias”; we’ve deleted the unclear sentence “Any included trial was rated at low risk of bias in all dimensions”.
We used Cochrane criteria, from Handbook. We have clarified as: “we assessed the risk of bias as advised in the Cochrane Handbook of Systematic Reviews of Interventions”

2) Meta-analysis
Use heterogeneity estimated by I2 (no inconsistency)

*Author answer*

*Ok.*

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.
Reviewer's report 3

Title: Constraint-induced movement therapy: trial sequential analysis applied to Cochrane Collaboration systematic review results.

Version: 1 Date: 14 July 2014
Reviewer: Gordon Murray

Reviewer's report:

I have a number of major reservations with this paper.

1. The most serious is the way the authors raise the issue of multiple testing in cumulative meta-analysis, likening this issue to, for example, repeated interim analysis in a clinical trial. The crucial difference is that with interim analysis the trial continues as planned unless an efficacy or safety or futility boundary is crossed, at which point the trial stops and is reported. This process can clearly ring false warning bells, and some correction is needed to allow for multiple testing. However, in meta-analysis one wants to use the totality of available (randomised) evidence, and the most up-to-date meta-analysis (if robust, etc) will always trump a previous one. In particular, one is not 'chasing a p-value'.

Author answer
TSA methodology can be analogous to the concept of performing interim analyses in which a trial is stopped for efficacy or futility: we would stop the publication of trials if we obtain evidence on the intervention’s overall clinical benefit or harm. Furthermore, cumulative meta-analyses allows clinicians to get more conservative results, because it determines if the addition of a new trial to a series of existing trials results in the inflation or deflation of the overall effect size. We agree with the peer reviewer on the fact that the process of multiple testing can ring false warning bells and need some corrections. When meta-analyses are updated, they are repeatedly subjected to significance testing over time such that the likelihood of observing a false positive or false negative result becomes higher. TSA should address the problem of ‘multiplicity’ by adjusting the thresholds for statistical significance by using sequential monitoring boundaries as a threshold for the employed test statistic. We are aware that TSA has a limit: in contrast to an interim analysis in RCTs, it is not possible to know the previous number of statistical looks in meta-analysis. Given the adjustments, results could have a good consistency: the strength of the evidence should consider the impact of multiplicity.

2. I accept that cumulative meta-analysis is interesting if one wants to look back and ask at what point the evidence became strong, or to see evidence of early bias, or potentially as a means to power a new trial (but I have yet to encounter a funding body which is prepared to fund a trial on the basis that it could tip the balance of a meta-analysis, rather than being powered to stand alone). However, I am not convinced that one needs to take account of history to make a valid inference on the strength of the currently available totality of evidence.

Author answer
We thought that a cumulative meta-analysis is interesting not only to look back, but also to look forward and ask how much effort researchers should make to obtain strong evidence, combining even a GRADE approach. The firmness of a conclusion depends on the amount and quality of evidence provided. The great potential of TSA is to provide a tool to
assess these elements and prioritize the “new” research. We took into account the totality of the existing evidence because it reflects the time researchers spent to test their hypothesis looking for more effective and innovative treatments. We firmly think that history is crucial and necessary to stimulate researchers looking back and ahead to improve evidence-based practice.

3. The authors seem to be saying that early, small trials are typically biased, and that as more evidence accumulates then the treatment effect estimated from the meta-analysis tends to head towards the null. This is certainly true, but one should not attempt to solve a problem which relates to bias by using a conservative inference technique which is based on adjusting for multiplicity.

Author answer
Since the trials included in our review were small sample trials, which can underrate or overrate the meaning of their finding, meta-analysis results could be erroneous. Several studies indicated that the risk of overestimation of an intervention is usually high when the number of patients or events is small; the risk decreases over time when the number of patients or events increases. Little attention has been paid for the risk of random error in literature for which it is important to have a powered study. On the contrary, there is much literature on the risk of bias, such as methodological bias, outcome reporting bias, and ways to address them. TSA is an improvement of a meta-analysis, which combines a priori information size calculation (to handle the random error) with the adjustment for multiplicity. The interpretation of findings are even linked to study quality assessment (risk of bias). In our results, the evidence accumulated tends the treatment effect towards the null: interpretations are due not only to the small sample trials published but also to the low methodological quality over years.

4. The authors seem to get mixed up with the MID, and how it should impact on the estimated treatment effect. These two points are unrelated, although clearly the MID is crucial for a power calculation, or for interpreting an inference. In terms of their case study, this observation seems to circumvent their problem. The estimated effect size for the conventional meta-analysis for the primary endpoint is 2.88 units with a 95% CI of 0.08 to 5.68 and a p-value of 0.04. Although conventionally ‘significant’ it is obvious that the p-value is so borderline and the 95% CI comes so close to including a null effect, that it would be ridiculous to take this as evidence of clinical efficacy. It does not need a spurious injection of a multiplicity correction to add a degree of conservatism, just common sense!

Author answer
We agree with the peer reviewer that TSA methodology and MID calculation are two unrelated topics, but we consider this review a particular scenario to explore both aspects. We investigated an innovative methodology to refine a meta-analysis in which it is essential to fill in a minimal important clinical difference. Not every outcome has an established MID: this aspect is particularly compelling in rehabilitation. Our objective was to introduce a new statistical method to make the reader interested in calculating the MID.

We agree with the peer reviewer when he asked for common sense; however, we conducted TSA as a powerful and easy method to get over the random error and avoid misleading findings. Results of the meta-analysis could have been erroneous. To clearly define our findings, we have introduced in the discussion this sentence: “Although meta-analysis results of primary outcome were statistically significant, the clinical relevance was questionable because of the small effect size”
5. Not that it is that relevant (see point 4), I am concerned to see that authors advocating 0.5 SD as a way to establish the MID. The MID depends totally on the clinical context, with, for example, a cheap, safe intervention needing only a modest MID to justify its use in clinical practice whereas an expensive intervention with a worrying safety profile would need a much larger MID before its use could be considered.

Author answer
We agree with the peer reviewer that MID depends on the clinical context. MID is not always reported for every outcome and type of population. In our context, we found MID for the disability outcome (FIM) representative for a stroke population (Beninato 2006), which did not follow our inclusion criteria. Stroke patients included in our population had a higher initial FIM score, which was necessary to be able to perform CIMT technique, compared to the Beninato population. Beninato estimated a MID for stroke patients as 22 on total FIM (total is 126 points). However, such improvement on a patient that started with a high score is not possible: the higher the FIM score, the lesser the mean difference is clinically evident. For this reason, we have decided to do the 0.5 SD methods as a way to establish a MID adequate for our population.

6. A minor point, but given the authors’ focus on multiplicity, it would be good to see some discussion of how to handle the two outcome measures which they are exploring in their case study.

Author answer
We thank the peer reviewer for the suggestion. We explored two important outcome measures in stroke rehabilitation (disability and arm motor function) with no discussion on how to handle them because the purpose of our case study was to show an example of how to perform a trial sequential analysis. However, we have inserted a paragraph about this problem in the discussion.

“Reaching more evidence on what treatments could be more effective for patient recovery, should be imperative to decrease the social-economical burden of this pathological condition: the choice of disability and arm motor function as main outcomes is due to their importance not only for patients but also for clinicians, families, hospitals and society. An effective treatment that decreases disability level and increases arm motor function, can improve patient ability in daily life, his life satisfaction and quality of life and reduce family load, social assistance, hospitalization and economic social charge.”

Level of interest: An article of limited interest
Quality of written English: Needs some language corrections before being Published
Statistical review: Yes, and I have assessed the statistics in my report.
Declaration of competing interests: I declare that I have no competing interests.