Responses to Reviewer's comment

Reviewer #2: The authors have satisfactorily addressed most of my comments.

We would like to thank Reviewer 2 for the honesty of his/her statement. For the sake of clarity, we have fragmented and numbered comments, one of them raising a deep epistemological problem.

1. I disagree with the approach of only highlighting the better quality studies. The authors justify their approach by citing a reference from 2001, but the methodological thinking of 15 years ago has changed over time. When I did my first systematic review in 2000, poor-quality studies could be excluded entirely from review. Such thinking no longer exists.

We are not sure we understand your comment. Based on our understanding, you say that 15 years ago poor-quality studies were simply not-included in reviews and that such thinking no longer exists, meaning that you consider this evolution as a better approach, Nonetheless, you should recognize that our study did not exclude poor-quality studies. Moreover, close attention was paid to determining *a priori* our methodological inclusion criteria. We believe that the difference in our respective viewpoints could stem from what you mean by "highlighting".
If you mean that we only included better quality studies, this was not the case. We think there is some confusion here between high-risk of bias and quality, a distinction made by the Cochrane collaboration recognized leaders in systematic review methodology. However, if you mean that we should discuss the results of studies with a high-risk of bias studies (note that, in line with the Cochrane collaboration we make a distinction between risk of bias and quality) on an equal footing with the results of low-risk of bias studies, then we disagree. This would be contrary to the recommendations of the Cochrane collaboration and the high methodological standards set by PLoSOne.
Here we attempt to explain our position on epistemological grounds:
First, to be sure we are talking about the same thing; we would like confirm that the main issue raised here is whether it is possible to consider biased studies as evidence (discussed below). In the first version of our reponse to Reviewers we explained why we consider that taking into account the results of high-risk of bias studies was not appropriate (citing among others a reference from 2001 [Sterne et al. 2001]).
Even if we consider this reference as obsolete, we failed to find studies demonstrating the advantages of considering and discussing the results of highly biased studies. In fact, recent publications still recommend that only "high-quality" studies [Guyatt et al. 2011] when available should be considered, particularly in the field of health care and in medico-economics [Al-Shahi Salman et al. 2014; Chalmers et al. 2009; Chalmers et al. 2014; Chan et al. 2014; Glasziou et al. 2014; Ioannidis et al. 2014 Katikireddi et al. 2014; Macleod et al. 2014; Scott et al. 2012, Yordanov et al. 2015].
Finally, we would like to point out that our position on "the status of evidence" is not dependent of the results we found. If we had been able to include studies with low-risk of bias in favor of cranial osteopathy and many highly biased studies demonstrating no effect of cranial osteopathy, we would have discussed the studies in favor of the therapy (and would have been happy to do so). Unless there are studies to support your view, asking us to highlight poor-quality studies seems like a step back-in-time and contrary to what we teach our students.

Al-Shahi Salman R, Beller E, Kagan J, Hemminki E, Phillips RS, Savulescu J, et al. Increasing value and reducing waste in biomedical research regulation and management. Lancet 2014;383:176-85. [PMCID: PMC3952153] [PubMed: 24411646]
Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. Lancet 2014;383:156-65. [PubMed: 24411644]
Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009;374:86-9. [PubMed: 19525005]
Chan AW, Song F, Vickers A, Jefferson T, Dickersin K, Gøtzsche PC, et al. Increasing value and reducing waste: addressing inaccessible research. Lancet 2014;383:257-66. [PMCID: PMC4533904] [PubMed: 24411650]
Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet 2014;383:267-76. [PubMed: 24411647]
Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 2014;383:166-75. [PMCID: PMC4697939] [PubMed: 24411645]
Katikireddi SV, Egan M, Petticrew M. How do systematic reviews incorporate risk of bias assessments into the synthesis of evidence? A methodological study. J Epidemiol Community Health doi:10.1136/jech-2014-204711
Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JP, et al. Biomedical research: increasing value, reducing waste. Lancet 2014;383:101-4. [PubMed: 24411643]
Scott IA, Glasziou PP. Improving the effectiveness of clinical medicine: the need for better science. Med J Aust 2012;196:304-8. [PubMed: 22432658
Sterne JA, Egger M, Smith GD. Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. BMJ. 2001 Jul 14;323(7304):101-5.
Yordanov Y, Dechartres A, Porcher R, Boutron I, Altman DG, Ravaud P. Avoidable waste of research related to inadequate methods in clinical trials. BMJ. 2015;350:h809

2. To gain insight into the overall body of knowledge on a topic, the evidence from all studies should be summarized, with the low-quality studies discusses separately (but still discussed) from the better quality studies. The authors should strive to discuss all of the included studies in their paper. Otherwise, they put the spotlight on a subset of articles and do not report the totality of the evidence. Recent methods advances with respect to grading the strength of evidence are indicative of the need to include all of the evidence (good and bad) in systematic reviews.

In line with our comment above, we should define what we consider as “evidence” and “body of knowledge”.
Clearly, the evolution of a parameter after a treatment, as such, should not be considered as evidence of the treatment effect. In fact for us, “evidence” means scientific data obtained from a clearly planned methodical approach, limiting any bias as far as possible. We believe that our analysis and discussion were conducted within this framework and that this definition implies that low-quality studies should not be considered as evidence (even if mentioned in the analysis for our work). In our study, three articles reached this status of “evidence” (and the rules used to obtain them are described in the methodological section) and are discussed.
We included studies following clear criteria (not questioned by the reviewers); we methodically analyzed the weaknesses of all the included studies and, still *a priori*, proposed a general evaluation of risk of bias. We agree that this last point presents some limits and should be discussed. We propose to modify the manuscript as follows:

> For the studies of diagnosis reliability in cranial osteopathy, naturally we recommend that future researchers use the items proposed in our study and inspired from QAREL. We must be particularly vigilant about the personal expertise of the examiners and avoid those whose training

is not fully completed. **We should add that the tool we proposed was designed to specifically assess the risk of bias linked to the study methods but that reliability was not evaluated, representing one of the limitations of our study.** For inter-rater reliability studies, as much as possible must be done to ensure that exchange of information between examiners is not possible during the tests. Thus, procedures extending over several days are not recommended. This point leads us to consider strategies to avoid memorization of the results by the examiners. First, the order of assessments (subjects and examiners) has to be randomized and no information about subjects, outside of that necessary for the examination, should be communicated to the examiners. In addition, blinding of subjects and examiners has to be as strict as possible. On this last point, Halma et al. [32] proposed a quite outstanding plan to isolate the examiner from tactile, visual, auditory and olfactory cues. Note also that for studies involving simultaneous evaluation of a subject by two separate examiners, the method sections detailed in studies by Rogers et al. [28], Moran & Gibson [30] and Sommerfeld et al. [31], should serve as models for this methodological approach.

Not surprisingly, we advise future researchers to refer to the Cochrane risk of bias tool in order to build the ideal efficacy study**. However, we must mention that the reliability of this tool was only evaluated as fair for most of its items constituting another limitation of our study [51]. This tool or training in the use of this tool should be enhanced.**

The following reference has been added:

[51] Hartling L1, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, Tsertsvadze A, Hempel S, Shekelle P, Dryden DM. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol. 2013 Sep;66(9):973-81.

From our methodology, we considered (still *a priori* and with clear criteria) that some studies could not be considered as evidence (on either side). If the reviewers agree with our method (and we have had no comments dealing with these points) we do not understand why we should discuss results that our methodology led us not to consider as evidence (regardless of the meaning of the results). Detailed discussion of high risk of bias study results would artificially compensate with words the absence of evidence due to the lack of a rigorous methodology [see the paper by Horton, 1995 and the in depth debate around it] and would lead to a kind of spin in research publication. By “spin” we mean a way of reporting aimed at convincing the reader that the treatment used was effective whereas the methodology used and/or the analysis of data cannot support such a statement [see for instance Fletcher & Black, 2007; Boutron et al., 2010]. Note that this epistemological point is currently the subject of debate and PLoSOne has published an interesting study on this topic entitled “All That Glitters Isn't Gold” [ter Riet et al., 2013]. Moreover, a quick overview of the published systematic reviews in PLosOne leads us to think that discussing low-quality studies is not the rule.

In any event, we believe that this epistemological debate is crucial and that PLoSOne would be an excellent platform for it. Thus we propose to reproduce the reviewer’s comment and our response as additional material to the article.

Horton. The rhetoric of research. BMJ. 1995 Apr 15; 310(6985): 985–987.
Fletcher RH, Black B (2007) “Spin” in scientific writing: scientific mischief and legal jeopardy. Med Law 26: 511–525.
Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA. 2010 May 26;303(20):2058-64

Ter Riet G1, Chesley P, Gross AG, Siebeling L, Muggensturm P, Heller N, Umbehr M, Vollenweider D, Yu T, Akl EA, Brewster L, Dekkers OM, Mühlhauser I, Richter B, Singh S, Goodman S, Puhan MA. All that glitters isn't gold: a survey on acknowledgment of limitations in biomedical studies. PLoS One. 2013 Nov 20;8(11):e73623. doi: 10.1371/journal.pone.0073623. eCollection 2013.