Author's response to reviews

Title: Development and evaluation of an instrument for the critical appraisal of randomized controlled trials of natural products.

Authors:

Tannis Jurgens (tannis.jurgens@dal.ca)
Anne Marie Whelan (anne.marie.whelan@dal.ca)
Melissa MacDonald (mmmacdon@dal.ca)
Lindsay Lord (lordl@dal.ca)

Version: 2 Date: 27 February 2009

Author's response to reviews: see over
Dr. Tannis Jurgens  
College of Pharmacy, Dalhousie University  
5968 College Street  
Halifax, Nova Scotia  
Canada, B3H 3J5  
Ph: (902) 494-3796  Fax: (902) 494-1396  
Email: tannis.jurgens@dal.ca  

Feb 27, 2009  

The Editor BioMed Central Complementary and Alternative Medicine  

Re: MS: 1648499584241248  
Development and evaluation of an instrument for the critical appraisal of randomized controlled trials of natural products.  
Tannis Jurgens, Anne Marie Whelan, Melissa MacDonald and Lindsay Lord  

Our manuscript was recently peer reviewed and we were asked to respond to the reviewers’ comments. In the pages that follow, we have addressed each specific comment from Reviewer 1 and Reviewer 2. In cases where editing of the manuscript was required, we have stated that we have made the required changes in the reply to comments and have made the changes in the revised manuscript that is being submitted.  

We thank the reviewers for their insightful comments and are confident that the revised manuscript is an improved document.  

We look forward to hearing from you regarding the suitability of our revised manuscript for publication in BioMed Central Complementary and Alternative Medicine.  

Sincerely,  

Tannis Jurgens  Ph D  
Associate Professor, Medicinal Chemistry and Natural Health Products
Response to Reviewer 1

Note: Original comments from Reviewer 1 are shown in regular text. Responses from the Authors are presented in bolded, italicized text.

Overall, the authors have completed a well done, interesting report outlining the development and evaluation of a critical appraisal instrument for evaluating RCTs of conventional medicines for use with single entity natural products. It is well known that critical appraisal tools offer a means for estimating potential biases and factors that may affect study findings; with CAM trials a burgeoning area of research the development of a tool specific to CAM is not only necessary, but timely. The authors use a traditional Delphi approach using 11 of a potential 14 experts for establishing and finalizing a list of items for use on the final evaluation tool. The paper is generally well written, although lengthy, its strength lies in the level of methodological detail for how the tool was developed and evaluated.

As a general comment, critical appraisal tools have often been criticized as being cumbersome to use when lengthy and thus are often not used by authors within many reports (particularly systematic reviews that include many studies). This tool is rather lengthy (28+ items) and may impede uptake of the tool. A suggestion would be to include this as a limitation of the tool to its potential use for authors. 

This has been added to the “Strengths and limitations” section

The authors have indicated that a tool has been developed for responding to questions within the tool – this should be included within the publication as well if available. 

We have added the “User’s Guide”, as Appendix B, to the end of the manuscript as requested.

I would urge the authors to also include information (if not already included) on interpreting findings and how to incorporate the information into reports. This is a common limitation for previous critical appraisal tools. 

We agree that this has been a limitation of some previous critical appraisal instruments. Our intent was to make the items contained in the instrument itself easy to understand for new users. The User’s Guide that we developed to accompany the instrument provides explanation for each item and, in some cases, examples. The User’s Instrument has been added to the manuscript.

- Major Compulsory Revisions

1. The strength of the Delphi process lies within the ability to choose the appropriate ‘experts’ to be involved within the process. It would be prudent for the authors to include a brief outline or descriptor for how they decided on experts and what the criteria was for determining candidates. This has been added to the Methods section under Participants p6.

2. The search for critical appraisal instruments for conventional medicines was taken from a previously published systematic review; however, the review search date was from 2006 (uncertain of specific month) which jeopardizes the currency of the information. It is suggested
that authors update the search and do a quick scan to ensure the information is up to date (even if only in the major databases)

In our manuscript we state that the database search was completed in June 2006. As a consequence of the time involved in taking the information gathered from the results of the June 2006 search through the steps of the instrument development and finally the instrument evaluation, there is a resulting 2 year time gap. While we did not repeat the databases search during this time, we consistently scanned the literature for articles describing new instruments etc.

3. I am deeply concerned with the methodology of using ‘frequency’ of items to determine relevance for inclusion.

We are assuming that this first sentence is referring to 1 of the 3 criteria we used to support the inclusion of the non-NP items into the new instrument. We initially thought that including those items that were in the majority of published instruments was logical, however, when we realized how many of the instruments were variations and/or adaptations of older instruments, we realized that this had potential to skew the frequency of some items. We noted this in the original discussion section of the manuscript. We also noted, that due to the limitations expressed, that we did not use frequency as a criterion. However, the way we originally wrote the paper appears to be confusing. Therefore, we thought it would be best to remove this criterion entirely as we did not end up using it. This has been revised in the Methods section. This does not affect any of the items in the new instrument as there were no items that were included exclusively based on frequency. A discussion of the frequency issue has been left in the Discussion section of the manuscript.

Additionally, the frequency column has been removed from Table 2.

Did the authors not calculate kappa’s for agreement between experts? They report in a way that is somewhat unclear – for example ‘there was general agreement’, or ‘no significant effect’ but do not provide the data (p values) or methods for assessing significance. These need to be reported clearly.

We think that this section of the reviewer’s comments is referring to the items selected for inclusion in the NP section. This was done by achieving consensus using the Delphi process. With the Delphi process, consensus does not have to be 100% but it should be stated as to what will be accepted as consensus. We had not clearly stated this in the original manuscript. We have added a statement declaring that consensus was considered to have been reached when at least 80% of participants were in agreement. In Round 1 we achieved 100% consensus on 9 of 16 items. Remaining items had only 1 or 2 participants who didn’t agree. These results are given in Table 1. At the end of Round 2 we reached consensus of a minimum of 80% on all items. As the goal was to reach consensus, no statistical analysis was conducted.

- Minor Essential Revisions

4. I’m uncertain as to what a ‘revised CONSORT’ statement is. If the authors revised it they should provide a description of what was changed.
The original CONSORT statement was published in 1996. The authors of the CONSORT revised it, based on feedback etc, and published it as the “Revised CONSORT statement” in 2001.

5. Comments and discussion between anonymous expert participants is a strength of the Delphi process, as such, the author may want to report how many of the participants in fact provided comments and descriptions of why they included certain items along with responses to the questions. This would help strengthen the transparency of the process. This information is provided in Table 1. Examples of some of the items that were suggested to be added etc are written in the text in the Results section.

6. Information was emailed to participants; the authors should include a statement as to how they maintained anonymity of participants in the process (e.g. BCC participants?). As it stands now, I’m uncertain as to how this was achieved.

All contact with participants was conducted by email, with each participant receiving an individual email everytime email contact was required i.e. no mass emails were sent so the identity of all participants in the group was concealed. This has been clarified in the manuscript in the Methods section of Phase 1, under the heading “Participants” and “Procedure” p6. This was also clarified in the Results section.

7. The authors include instruments that are ‘modified’ from their original versions within the inclusion of the report. Although the authors are unlikely to change their approach (and have indicated this as a limitation); they should have been excluded entirely. They are not validated in their new form and skew the data on frequency (which they relied on for determining criteria). It is true that that including instruments that were modified from earlier instruments has the potential to skew the data on frequency. As noted in # 3 above, we addressed this in the original Discussion section of the paper and did not use this as a criterion after we discovered the limitation that doing so would have on the methodology. Therefore, we have clarified this in the text. We do stand by our decision to include all instruments in the initial number of how many instruments were identified. The fact that other researchers thought their content was good enough to use as a basis to develop a new instrument tells us their content is important to include. Many of the original instruments were not validated in their original form so the fact that they are not validated in their modified form is not a major issue. In fact, of the 99 instruments that we identified, only 4 indicated that they had been validated. Therefore, we do not think that the lack of validation, in this context, is an issue.

8. Although they randomized participants to groups, the authors should provide a description of the experience/expertise in critical appraisal (match comparison?) of participants for the field test of the instrument. Were there differences between groups at baseline with such a small sample? Did they include a broad range of expertise? It would be interesting to see the findings if a range of expertise (novice vs. expert) existed. There is a brief comment on this in the discussion section; but should appear clearly and concisely within the methodology section.

The participants were pharmacists, pharmacy educators and pharmacy students. There was a range of experience with NPs and critical appraisal among all participants. Our original intent was to see if there would be obvious differences in responses to items, depending on the
expertise of the participant. However, we found that it was difficult to accurately quantify the range of expertise of each participant as there were actually 2 areas of expertise to consider for each participant. For example, a participant could know a lot about NPs but have little experience with critical appraisal and visa versa. Depending on their year of study in the pharmacy program, some students had more experience with critical appraisal than did some of the pharmacists who had been practicing for a while. With the small number of participants and the difficulty of having 2 areas of expertise to consider, no meaningful results were obtained from examining the effect of experience on results We were satisfied that all participants, regardless of experience, were able to successfully use the new instrument.

9. The criteria used for determine that an item was ‘essential’ or not is concerning. These criteria are neither validated and nor ‘standard’ as part of the Delphi process.

We assume the reviewer is referring to the NP items, however from comments later in this section, we are not sure if perhaps the reviewer is referring to the criteria that were used to determine the non-NP items that would be included in our new instrument.

NP items
The initial list of items was compiled from all published NP tools and several were added by the research group. The method that was used to determine whether an item is deemed “essential” was to have a number of people with a range of expertise in the field in question reach consensus, through the Delphi process. The Delphi process is a well documented method widely used to achieve consensus. The results produced using this method can, of course, be influenced by the expertise of the group. We included a range of expertise (which we have now more clearly spelled out in the manuscript) and we feel that this was sufficient to establish a list of items that this group of experts agreed were “essential”. Others, such as Gagnier et al (ref 3) have used a similar method. We are unclear as to what the reviewer means when they say the criteria is “neither validated nor standard as part of the Delphi process”. The Delphi technique we used has been used by others in similar research and was described for example in ref 4.

Are these arbitrary criteria created ad hoc by the authors for including items? If so, the authors should state this and provide a discussion for its limitations. These comments suggest that the reviewer is referring to the criteria for selecting the non-NHP items. After conducting the systematic review of all of the instruments that we could find and extracting and collating all of the items, we were left with a large number of items, some of which appeared many times. The criteria that we used to select items for inclusion were “ad hoc”, however, we felt that they were the most supported by published research and therefore defensible. Initially we had intended to include “frequency” (Included in > 50% of instruments) as 1 of the 3 criteria. However as noted in the original Discussion section of our paper we ended up not doing this because of the limitations identified upon completion of the data extraction from existing tools. As noted in # 3 and # 7 above, this appeared to be confusing the way we originally wrote the paper so it has now been clarified in the revised manuscript. Interestingly, removal of this criterion did not require removal of any non-NHP items as the items in our instrument were each supported by one or both of the other 2 criteria.
The question that follows is – why go through this entire process of using experts etc. to resort to using previous validated tools?

_We are assuming that the reviewer is still talking about the non-NHP items with regards to this question. We used experts ONLY for the NHP items. The selection of the non-NHP items was based on if they were contained in a validated instrument OR had empirical evidence to support their inclusion. We recognized that validation is an important step in ensuring that the item should be included. We compiled all items from the 4 validated tools, therefore putting the best together._

Why not simply update a validated tool and just add the section for single NP products?

_Initially, before starting this project, we thought that this would be possible. We did not find a critical appraisal instrument that was widely accepted as being the “gold standard”. Therefore, rather than chose one that may not prove to be the best, we decided to assemble our own, using the best from validated instruments._

_We have added a sentence to the Introduction to clarify this fact._

- Discretionary Revisions

10. A general comment – this report could be shortened by editing extraneous information within the text. Although stylistic, it would be less arduous to read through.

_We were not clear as to which sections the reviewer felt were extraneous. We are willing to edit to shorten the manuscript if some suggestions are given._

**Level of interest** An article of importance in its field  
**Quality of written English** Acceptable  
**Statistical review** Yes, but I do not feel adequately qualified to assess the statistics.  
**Declaration of competing interests** I declare that I have no competing interests

---

**Response to Reviewer 2**

**Note:** Original comments from Reviewer 1 are shown in regular text. Responses from the Authors are presented in bolded, italicized text.

This is an ambitious project attempting to develop a new instrument for the critical appraisal of RCT on natural products (NPs). The authors first performed systematic review of all existing instruments for critical appraisal of NP RCTs. (This in itself is the most important contribution). The authors then used Delphi method to identify items deemed essential in describing NPs. They then field tested their instrument against existing ones.  
While the overall approach is solid, I believe that the authors need to make some clarification in the paper (perhaps, under Limitation Section);

- The most instruments they identified pertain to the quality of REPORTING standards (e.g.
CONSORT authors themselves warned that the CONSORT checklist should not be used to assess the quality of research.

*It is true that the CONSORT statement is intended to be used as a checklist for what to report in an RCT and not as a critical appraisal tool itself. However, the content of checklists for what should be included when reporting on an RCT is useful to let us see what issues are important. This was the spirit in which the information was used-to inform as to which items were important.*

- The authors should discuss/acknowledge the difference between the quality of reporting and risk for bias (the latest Cochrane Handbook provides the useful discussion on this issue + identify the items that are empirically linked to bias)

*Risk of bias is, of course, extremely important to identify in RCTs if we want to determine if we can believe the results. Because of the complex nature of NPs, assessing the risk of bias alone would not be sufficient. It is also important to assess the information provided regarding the identity of the NP tested. Our instrument includes items important in assessing the risk of bias as well as other essential items regarding the identity of the NP.*

- Related to the same issue, the statement that “4 of 99 instruments have been validated” (page 13) is actually not correct (e.g. Jadad’s score has been abandoned as its scale did not show usefulness in several studies as it was case with general scale-based instruments etc)

*We understand the reviewer’s comments regarding the fall from favour of the Jadad scale. However, at the time of the original publication of the Jadad scale, the publication stated that the scale had been validated. That was our criteria for assessing whether instruments had been validated-whether the publication stated that they had been.*

- I found description of the field experiment lacking and ironically the authors did not follow their own advice in designing/reporting this trial (e.g. the primary outcome was not well defined, sample size was not provided etc)

  We have revised the description of the RCT.

  - The term “double-blind” is now discouraged (instead one should specify who was “blinded” to what and why; Gordon Guyatt’s group discussed this in JAMA article several years ago

    *The instrument has been revised to remove “double” from the blinding item.*

- The authors should distinguish evidence from decision-making; in other words, the quality of appraisal does and should not be in a straightforward fashion translated into application of evidence (items #28). Many factors play role in making recommendation (e.g. benefits/harms ratio, other relevant evidence, preferences etc). I think the item #28 should be dropped from their checklist.

  *We have clarified item #28 in both the instrument and user’s guide.*
- I am surprised that pharmacological/biological plausibility with respect to NP mechanism of action has not been included in the checklist. For example, most people found homeopathy trials not credible because of the lack of solid scientific theory.

In our systematic review of critical appraisal instruments for conventional and NPs we did not find any items referring to the pharmacological/biological plausibility of the intervention. The purpose of the instrument is to appraise the mechanics of the trial and the content of the intervention to determine if the results of the trial are solid. If the intervention is not credible/plausible then the results of a well designed trial should show this.

- Finally, please include definition of CAM and NP medicine.

The definitions to both of these abbreviations are in the first 2 paragraphs of the Background section of the manuscript.

Level of interest An article of limited interest
Quality of written English Acceptable
Statistical review No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests 'I declare that I have no competing interests'