Author's response to reviews

Title: Reliability, construct and discriminative validity of clinical testing in subjects with and without chronic neck pain

Authors:

René Jørgensen RJ (rejo@ucsyd.dk)
Inge Hansen IR (iris@health.sdu.dk)
Deborah Falla DF (deborah.falla@bccn.uni-goettingen.uni.de)
Birgit J Kristensen BJK (bjuul-kristensen@health.sdu.dk)

Version:3 Date:18 September 2014

Author's response to reviews: see over
Cover letter

Thank you very much for the opportunity to submit our revised manuscript.

Reviewer 1

Reviewer's report
Title: Reliability, construct and discriminative validity of clinical testing in subjects with and without chronic neck pain
Version: 2 Date: 7 July 2014
Reviewer: David Walton

Reviewer's report:

Thank you very much for these comments, which certainly will improve the manuscript. In the following we have addressed the comments point-by-point

Intro
1. Line 72 re: gaze stability widely described – requires a citation?

Answer: Citation has now been added as suggested by the reviewer

Action: ‘Gaze stability and the smooth pursuit neck torsion test, tests of oculomotor control, have been widely described and applied in the assessment of people with neck pain (Kongsted 2007, Treleaven 2008).

Design
2. Line 104: Why was examiner C not blinded to patient status?

Answer: Paragraph has been rephrased in line 122.

Action: ‘Examiners were blinded to the status of the subjects, except for examiner C, since this was involved in the recruitment of cases and controls. Although examiner C was aware of the subject’s status, this examiner only performed the PPT and SWAY tests, which are two highly objective tests, thus limiting potential bias. All examiners were mutually blinded to the results of other examiners.’

Procedure
3. Line 150: Subjects performed SWAY and PPT tests? This sounds like they did it on themselves.

Answer: Sentence has been rephrased, line 205

Action: ‘Examiner C then screened for in- and exclusion criteria, after which SWAY and PPT tests were conducted.’

Sample Size
4. Line 160: In a two one-sided test analysis… Is this the correct term?
Answer: Thank you for this comment, but we believe this is the right expression, as described by Hopkins, 2000. A sentence has been added to the paragraph in line 217, and a citation has been added.

Action: ‘Sample size was estimated based on the 95\% confidence interval according to the recommendation from Hopkins (Hopkins, 2000).’

Statistical analysis
5. Line 173-174: I’m not sure that the formula for SEM is correct. Should it not be SD\text{diff} \times \sqrt{1-ICC)}? As written, the \sqrt{2} factors cancel each other out in your MDC formula. Perhaps your formula is simply one I’m not familiar with.

Answer: Both methods are frequently used. We have used the formula described by de Vet et al., 2006.

Action: A citation has been added in line 239

This is a quotation from the de Vet. article

3. The value of SEM can also be derived by dividing the SD of the mean differences between two measurements (SD_{\text{diff}}) by $\sqrt{2}$. The factor $\sqrt{2}$ is included because it concerns the difference between two measurements and errors occur in both measurements. Note that the SEM obtained by this formula is again $SEM_{\text{consistency}}$ because systematic error is not included in the SD. Thus, $SEM_{\text{agreement, AC}}$ cannot be calculated in this way either.


Results
6. Line 195: A purely semantic comment, but PCS makes me intuitively think of the Pain Catastrophizing Scale. Perhaps others don’t have my same hang up, but the authors may consider calling it something like SF36PCS, or follow the same convention as used for the Mental Component Score immediately after. Not as elegant, but might prevent confusion?

Answer: Thank you very much for this point. This has now been changed to SF36-PCS throughout the manuscript and Mental Component score to SF36-MCS.

7. Line 194: This is the first time the BDI acronym is used. In the pain field this would be assumed to be the Beck Depression Inventory. If that’s the case, you
may need to justify the use of the scale for this study.

Answer: Beck Depression Inventory was used to discriminate between depressed and non-depressed adolescents and was used as part of the exclusion criteria as described from line 139-139, in order to secure stable results. In the present study it is only used to describe the patients. The abbreviation is now explained.

Action: The abbreviation BDI is listed after Beck Depression Inventory (BDI) > 29, line 139.

8. Line 197: Just a comment, but strange that 5 of 21 controls stated their neck pain condition changed over the course of a week. What happened there?

Answer: The change in GPE could be due to the relatively long testing procedure of approximately 1.5 hours, which might have even affected the healthy controls. We expected that this might be an issue, which was why GPE was included for both patients and controls – and only subjects answering 0, representing no change, were included for data analysis.

9. A general comment: when reporting comments, unless there’s a compelling reason for reporting p values to less than 1/100th of a point, round them to 2 decimal places (e.g. <0.01)

Answer: Thank you for this point. This has now been changed throughout the manuscript as suggested.

10. I tend to follow Paul Stratford’s lead on the GPE and assume that a range of -1 to +1 (admittedly on a 15 rather than 11 point scale) is adequately similar to consider the condition stable. I’m not overly prescriptive here, no question that restricting to a zero is a rigorous approach to test-retest reliability. The only reason I make this comment is that by doing so you’ve reduced your sample to arguably underpowered and no doubt your CI’s will be large. Just a comment, do with it what you like.

Answer: Thank you for this point. We have kept a conservative approach and only included subjects answering 0.

11. When reporting your ICCs, consider using the word ‘to’ rather than a dash ‘-‘ which can look like a ‘minus’ sign to those not paying attention.

Answer: Thank you very much. This has now been changed throughout the manuscript as suggested.

Discussion
12. As I’m reading the Discussion on the different tests I’m realizing that I have no idea at this point how the different tests were conducted (e.g. ROM, JPE). I know the description of the tests are in the Appendix, and I suspect the conservation of words in the test may have been due to word limitations. This may also be a comment for the editor to consider if word count is an issue.
However, I would encourage the authors to consider adding in even some cursory descriptions of, for example, ROM and the tool used (‘Cervical Range of Motion was captured using a …’). Without this information, and unless I looked at the Appendix first, I am unable to adequately interpret the results and discussion.

Answer: We appreciate this important point, and have added one to two more descriptive sentences to each of the tests for clarifying this issue in the section ‘Clinical tests’ from line 173 to line 200

Action:

*Cranio Cervical Flexion Test (CCFT)* was performed, using a Pressure Biofeedback Unit (Stabilizer; Chattanooga Group, South Pacific), as described by Jull et al. [38]. The subject was asked to perform cranio-cervical flexion in five incremental stages guided by the pressure sensor. The activation score has six scoring options; 20, 22, 24, 26, 28 and 30 mmHg.

*Deep Cervical Extensor (DCE)* test was performed in prone with their head over the edge of the bed. A laser was fixed to the top of the subject’s head and was projected to a target. The duration of time the laser beam was kept within the centre of the target was measured in seconds (sec.).

*Range of movement (ROM)* was examined using a bubble inclinometer (Baseline Bubble Inclinometer, Fabrication Enterprises Inc, USA) for flexion/extension and lateral flexion, and custom-made equipment for neck rotation (Figure 2). All scores were registered to the nearest degrees, except for rotation which was registered to the nearest 5 degrees.

*Joint Position Error (JPE)* was examined following return from active rotation, flexion, and extension movements by measuring the location of a laser beam positioned behind the subject. Data was registered in millimetres (mm).

*Gaze stability (GS)* was registered during rotation, flexion and extension movements as positive/negative based on the patients report of symptoms such as nausea, dizziness, disturbed vision.

*Smooth Pursuit Neck Torsion Test (SPNTT)* was tested in both a neutral head position and with the trunk rotated 45 degrees and was registered as positive/negative based on the patients report of symptoms such as nausea, dizziness, disturbed vision.

*Postural control* was measured during one-legged stance (eyes open and eyes closed) using a Wii balance board (Nintendo, Kyoto, Japan) and quantified with the SwayWithWii software program. Data was registered in millimetres (mm).

*Pressure Pain Threshold (PPT)* was examined at three sites (neck, m. infraspinatus and m. tibialis anterior) using a hand-held algometer (Wagner, FPX algometer, USA) and was registered in kilogram-force (kgf).

13. The JPE discussion is a little disjointed, in that one sentence states that previous ICCs were reported higher, and then two sentences later they are reported to be lower. In other words, it’s hard to know if your results are in keeping with previous work or not. Can this be somehow reoriented to be a bit more clear?

Answer: We have now rephrased the sentences in line 361 and line 367 according to the reviewer’s suggestion

Action: ‘Bland Altman plots revealed … ‘Previous studies have reported varying results, however, most studies report ICC above 0.75 [23,46,53-55].’
Other explanations could be…‘This is further supported by the fact that other studies using three repetitions have reported ICC similar/lower than the present study [24, 51]’.

14. In your gaze stability discussion, I would urge caution in switching between ICC and % agreement, as not all readers will understand the difference.

Answer: We agree with the reviewers point. The sentences including data on overall agreement has now been deleted from the discussion of both GS and SPNTT as suggested by reviewer.

15. As I continue to read through your discussion (SWAY), I’m even more convinced of my previous comment, that the readers need at least some cursory information on how these tests were performed, captured and analyzed. At least one of the authors has surely forgotten more than I currently know about postural sway analysis, and while I don’t think a full dissertation on the subject is required here, providing some basic information on the nature of data collection and analysis would certainly be helpful in the text.

Answer: As described above, we agree on this point. Please see the answer and action taken to comment no. 12.

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

References


Reviewer 2

Thank you very much for these comments, which certainly will improve the manuscript. In the following we have addressed the comments point-by-point

Dear authors,
The manuscript is very informative and adds to the evidence-based practice of physiotherapy. However, the manuscript would benefit from some more structuring; there is a lot of information and a lot of abbreviations are used. All feedback concerns minor essential revisions and discretionary revisions. From the beginning it seems that you have divided the specific tests in clinical ((CCFT, ROM, JPE, GS, SPNTT, DCE) and performance (SWAY, PPT) test, however later on in the manuscript you mix those; you use table titles with clinical test but performance tests are also displayed.

Answer: The table title in table 3 has now been changed.

Action: Changes are as follows “Table 3. Summary statistics of tests included in the reproducibility and test-retest study”

BACKGROUND:
1. Paragraph 3 explains known information about the tests, however it is not always clear whether results were from patients or controls.

Answer: We agree, and would very much like to explain this further. The results for the CCFT are from chronic neck pain patients, cervicogenic headache patients and asymptomatic controls. For JPE, results are from WAD and asymptomatic controls. For Gaze stability data are from asymptomatic controls only.

Action: This has now been added to the paragraph in line 79-90.

2. Paragraph 4: I was just wondering; why would you assess reliability in people without neck pain when those tests are widely used clinically? Throughout the whole manuscript I kept wondering whether it is legible to use whole-group results, since reliability depends on the target population. Reading the discussion-consideration 1st paragraph I understood why you choose to do so. It might be useful to refer to this information earlier in the manuscript.

Answer: The fact that these tests are widely used, does not mean that their psychometric properties have been satisfactorily evaluated, on the contrary, it seems that these tests are widely used in spite of the fact that they have not been properly evaluated. The information from the discussion will be referred to earlier in the manuscript.

Action:
The changes are as follows in the methods section (design), line 114: “This standardized protocol inherently included a case as well as a control group, to confirm that both groups could be tested reliably”.

And with rephrased in the discussion, line 449: “Further, by including a case as well as a control group,
we demonstrated that both groups could be tested in a clinical manner reliably.”

3. You extensively explain the first aim of the study (reliability), however there is very little information on the second aim of the study (construct and discriminative validity). Why is this so important?

Answer: We appreciate the Reviewers comment. The reason we focused more on the first aim of the study was because the study was powered to test for reliability.

METHODS

4. Design: are examiner A, B and C all authors of the manuscript? Could this have influenced the results of the clinical tests?

Answer: Only examiner C is an author of the manuscript. This of course could have influenced the results, but given that the tests examiner C performed were only PPT and SWAY, which are two highly objective tests, bias due to this is very unlikely.

Action:

In line 123-125 a sentence has been added: “Although examiner C was aware of the subjects status, this examiner only performed the PPT and SWAY tests, which are two highly objective tests, thus limiting potential bias.”

5. Questionnaires: it would help if you add range to the numbers after NDI, SF-36.

Answer: Thank you very much for this suggestion. The changes have now been performed as suggested by reviewer in line 156-157.

6. Clinical tests: table 1 and additional file 1 are very informative, however they contain a lot of information. Is it somehow possible to combine those files or to integrate some information in the text?

Answer: We have now added one to two more descriptive sentences to each of the tests in the section ‘Clinical tests’ from line 173 to line 200

Action:

Cranio Cervical Flexion Test (CCFT) was performed, using a Pressure Biofeedback Unit (Stabilizer; Chattanooga Group, South Pacific), as described by Jull et al. [38]. The subject was asked to perform cranio-cervical flexion in five incremental stages guided by the pressure sensor. The activation score has six scoring options; 20,22,24,26,28 and 30 mmHg.

Deep Cervical Extensor (DCE) test was performed in prone with their head over the edge of the bed. A laser was fixed to the top of the subject’s head and was projected to a target. The duration of time the laser beam was kept within the centre of the target was measured in seconds (sec.).

Range of movement (ROM) was examined using a bubble inclinometer (Baseline Bubble Inclinometer, Fabrication Enterprises Inc, USA) for flexion/extension and lateral flexion, and custom-made equipment for neck rotation (Figure 2). All scores were registered to the nearest degrees, except for rotation which was registered to the nearest 5 degrees.

Joint Position Error (JPE) was examined following return from active rotation, flexion, and extension movements by measuring the location of a laser beam positioned behind the subject. Data was registered in millimetres (mm).
Gaze stability (GS) was registered during rotation, flexion and extension movements as positive/negative based on the patients report of symptoms such as nausea, dizziness, disturbed vision.

Smooth Pursuit Neck Torsion Test (SPNTT) was tested in both a neutral head position and with the trunk rotated 45 degrees and was registered as positive/negative based on the patients report of symptoms such as nausea, dizziness, disturbed vision.

Postural control was measured during one-legged stance (eyes open and eyes closed) using a Wii balance board (Nintendo, Kyoto, Japan) and quantified with the SwayWithWii software program. Data was registered in millimetres (mm).

Pressure Pain Threshold (PPT) was examined at three sites (neck, m. infraspinatus and m. tibialis anterior) using a hand-held algometer (Wagner, FPX algometer, USA) and was registered in kilogram-force (kgf).

7. Procedure: would benefit from a flowchart.

Answer: A flowchart describing the procedure has now been inserted as figure no 3, and referral to this appendix has now been made in the procedure section, line 210.

8. Procedure: the procedure for the controls is missing.

Answer: The procedures for the controls were exactly the same as that of the cases, except that inclusion criteria were different for controls. A sentence has been added to clarify this under the procedure section line 212.
Action: The changes are as follows “Cases and controls followed the same procedure throughout the testing session”.

9. Sample size: even though the power analysis is presented, I think a sample size of 19 pairs is very small. See also the COSMIN checklist for measurement properties, box on reliability: http://www.cosmin.nl/images/upload/files/COSMIN%20checklist%20with%204-point%20scale%2022%20juni%202011.pdf

Answer:
Thank you very much for this relevant point. We are aware that the COSMIN checklist state that a much bigger sample size is required. However, the COSMIN checklist is developed primarily for questionnaires and not for clinical testing as presented in this study and some of the measurement properties may not be relevant/applicable for clinical assessment methods. To be given an “excellent” score in the COSMIN checklist a sample size of at least 100 is required, which we consider quite large for this type of study. This is supported by the FIMM protocol for clinical reliability and validity studies which state that the “training phase”, “overall agreement phase” and “study phase” requires about 10, 20 and 40 subjects respectively.

10. Statistical analysis: it is not always clear which data you use for different analyses.
Answer: We have used ICC and MDC analyses for ratio interval data, and kappa and overall agreement analyses for ordinal data. This has now been clarified in the statistical analysis section

Action:

- in line 233: “For calculation of intra- and inter- examiner reproducibility for ratio-interval data, ICC (2,1) and Bland and Altman’s with 95% limits of agreement (LOA) were used.”
- In line 240: “For ordinal data, Cohen’s κ statistics with 95% confidence interval were calculated.”

DISCUSSION:
11. Nicely organized.

TABLES
12. Make sure that tables are clear; short titles, and comments, notes and abbreviations under the table. Also make sure that all abbreviations are listed.
13. Table 3: point out that the second variable in the first column concerns ROM, and the last variable concerns the SWAY test.
14. Table 3: make sure that all abbreviations are written in full (CEA, A/P, M/L)
15. Table 3: last sentence in note under title should be in methods-statistical analyses section.
16. Table 4: unit of measurement is missing for JPE.

Answer: Thank you very much for these suggestions. All suggestions have now been implemented as suggested by the reviewer.

FIGURE 1:
17. To me it seems that there were 3 steps where cases dropped out, while theoretically they probably dropped out right after screening?
Answer: Three cases dropped out when they received the questionnaires, as they did not want to complete the questions. Another three cases were not included, since they did not meet one of the inclusion criteria (NDI<10), therefore they were not tested further for exclusion criteria. A further three were excluded due to signs of neuropathy and one due to an unstable working situation. Since these dropouts occurred as part of the normal screening procedure for in- and exclusion it was decided to be of smaller relevance in a clinical study for reliability and validity. We agree that it is of importance when testing psychometric properties of questionnaires.