Author's response to reviews

Title: Reproducibility of and sex differences in common orthopaedic ankle and foot tests in runners

Authors:

Maarten P. van der Worp (MvdWorp@gmail.com)
Anton de Wijer (Anton.deWijer@Radboudumc.nl)
J.Bart Staal (Bart.Staal@Radboudumc.nl)
Maria W.G. Nijhuis-van der Sanden
(Ria.Nijhuis-vanderSanden@Radboudumc.nl)

Version: 5 Date: 22 April 2014

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

Thank you for consideration of our manuscript for publication. We want to thank the reviewers for their (critical) comments and their suggestions to improve our manuscript.

We reviewed the manuscript according to the reviewer’s comments (please see below the point to point answers to the reviewer comments) and highlighted the changes in the revised manuscript.

We are submitting the revised version of, Reproducibility of and sex differences in common orthopaedic ankle and foot tests in runners, for publication as an original research article in BMC Musculoskeletal Disorders.

Yours sincerely on behalf of the co-authors,

Maarten van der Worp
**Reviewer: Takashi Nagai**

**Major Compulsory Revisions**

(Please check the values. - If it is just a typo (where to put a decimal, it is minor revision.)

We checked the values/units of both, tables and manuscript, and adjusted the typo faults.

Please check the values for NDT Stance and NDT Single Limb-Stance in the TABLE 2. Should they be in millimeter (mm)? If so, should NDT Stance be 6.0(sd2.7), 6.0(2.8), 6.0(2.8)? The same comment for NDT Single Limb-Stance.

Indeed, the values of the NDT stance and NDT Single Limb-stance are in millimeter. We followed your suggestions for correction and the NDT values, with the standard deviations, are now in mm.

SD values for AJD test and MTP 1 test should have one decimal. Please round them to the nearest 1/10 of a degree.

The descriptive values in the manuscript have now one decimal including the values in figures 7 & 8.

**Minor Essential Revisions**

**TABLE 1**

- Each table has different style. Please make all tables consistent. The TABLE 3 is probably the best table to follow.
- Total, males, females will be in the first row.
- The second row should contain "Mean +- SD" as seen in the TABLE 3.
- Then, each row should be the following: Age (yr), Weight (kg), Height (cm) - please change from meters to centimeters, BMI (kg/m2) - "2" should be superscript, Running experience (yr) - please be sure to keep "mean+-SD" instead of "median; range", Weekly training frequency (days) - the same comment on "mean+-SD" as above, Weekly running distance (km) - the same comment on "mean+-SD" as above.

We edited all the tables in compliance with Table 3 (lines, font sizes, etc.) and made the changes in the rows as you proposed for Table 1. In addition, all values were given in mean± SD.
TABLE 2

- "Total (n = 42)" instead of "Total population (n = 42).
- Headings should match TABLE 3. For example: NDT Stance (mm), NDT SL-S (mm), AJD-test (degree symbol), MTP1-test (degree symbol).
- As stated in the major essential revisions, please check the units on NDT Stance and NDT SL-S. AJD-test and MTP1-test should contain only one decimal.

The text of “Total population” was removed and we used “all participants” in the title of the table. Furthermore, headings were changed in accordance to Table 3 and we changed the units concerning NDT Stance and NDT SL-S as stated above.

TABLE 3

- Headings: Males (n = 22), Females (n = 20).
- You do not need to include mean difference and 95% CI. However, if you would like to keep them, please check for errors. For NDT Stance and NDT SL-S, the mean difference between males and females should be 1.8 and 1.9, respectively?
- P-Value contains two decimals on the first three variables while the MTP1-test has three decimals. It is the authors’ reference. But, please be consistent.
- Footnote: please change to "between males and females" instead of "between men and women"

We added the counts of males and females in the headings of Table 3. We kept the mean difference and 95% CI in Table 3, because the use of these values in the discussion regarding the findings of other studies.

“In the studies of Allen et al. [26] and McKeon et al. [19] no difference was found for the ND between males and females. Allen et al. [26] reported only the ND values for the ACL-injured group (mean ND of 10.2 and 10.7 mm for female and males, respectively) and did the NDT measurement with a metrecom. In the study of McKeon et al. [19] a sex difference of 0.1 was found with a 95% confidence interval (CI) of -0.01 to 0.24 mm in a cohort of 118 healthy adults and used the same protocol (sitting to stand) as in our study. However, McKeon et al. [19] did the seating measurements in subtalar neutral position and measured the navicular drop with a straight edge ruler.”(Page 17; Sentences 370-378)

Furthermore, the P-values are now consistently given with three decimals and we checked the values for errors.

FIGURE 1

- Step2: please change to "height" instead of "length".

We changed the “length” into “height”.

REFERENCE:

- Please be consistent on the format. One obvious inconsistency is journal names. Please keep their journal abbreviation. If you are not too certain, you can check on the PubMed website: 1. Curr Sports Med Rep, 3.Br

We apologize for this. We checked this and use the correct abbreviations now.
Reviewer: Angela Evans

1. The rationale for these participants

- Whilst I understand that runners are the future area of interest, given the many studies which have already examined the reproducibility of all of these clinical measures in adults, why have the authors sought to do this again; and adapting protocols has weakened any comparisons made.
- Why would a small recreational runner sample differ from the adults in previous studies?

We have chosen to adapt these three orthopaedic tests slightly, because we want to use them for a future prospective cohort study including over 400 participants of a running event. In this cohort study, measurement time was limited to a maximum of 10 minutes for the NDT, AJD-test and MTPI test.

In addition, the measurements had to be done on location and with minimal equipment. Therefore we chose to use skin markers and a ruler to optimize measurement time with a minimum of equipment.

“Moreover, in this study, the protocols in the literature of the NDT [8], AJD-test [11] and extension MTP1-test [10] were adapted for the use in our planned prospective cohort study of female runners (n= 433). This adaptation was necessary for practical reasons, which required that these orthopaedic tests are performed in maximal 10 minutes, on location and with a minimum of measurement tools and equipment.” (Page 5 and 6; Sentences 81-85)

We expected no big difference between the recreational runners in our study versus the adults in previous studies. However, because we slightly adapted the existing protocols a new reliability study was needed.

Furthermore, the reason for examining the reproducibility of these orthopaedic tests in runners was emphasised.

“In conclusion, above mentioned studies focused on reliability of the NDT, AJD-test and extension MTP1-test in healthy adults. However, these tests seem to be important to identify runners with higher injury risk and for prevention purpose.” (Page 5; Sentences 78-80)
2. The chosen clinical tests; and adapted protocols

- Please see above

- As raised in the first review: The use of clinical tests for both foot and lower limb which involve skin markings have long been shown to be less reliable, yet such methods are used for both 1st MPJ and NDT. Similarly the protocols for both of these tests (1st MPJ and NDT) were ill-defined (no standard step length; sit to stand vs ST palpation; measurement rods). It has also been shown in many previous studies that the reliability of measures which involves manual manipulation of foot posture (ie ND with ST position / sit-stand; 1st MPJ with variable step length) is reduced as opposed to those with standardized and non-manipulated attributes eg AJ rom - which has been well tested previously with the standard weight-bearing lunge method (Bennell).

We agree with the reviewer that it is important to use measurement tools with high reliability and that they should be performed in an optimal standardized way. As stated above, we choose to adapt these three orthopaedic tests slightly to improve the feasibility of test procedures, because we wanted to use them in a future prospective study (see also our previous comment). Furthermore, we rephrased the sentence about the transfer from sitting to standing position, so the standardization of the NDT Stance protocol is clearer.

“*The runner was then asked to stand, without moving the feet, equal weight bearing on both legs and the distance between the navicular marker and the floor was measured again (FIGURE 2).*” (Page 8 and 9: Sentences 155-157)

The other points of your concern (step length and using measurement rod/ruler vs. card method) were discussed later in the manuscript as stated in previous cover letter.

- “*In our protocol the knee was fully extended and the maximum stress on the MTP1 joint reached without balance problems. The possible difference in the size of the step length was not expected to influence the outcome of the MTP1 mobility.*” (Page 16; Sentences 345-348).
- “*Another explanation for the lower reliability in our study is that we measured the height of the navicular bony landmark with a ruler. The ruler was used so to observed the navicular drop directly and it is less time consuming than using a blank card [8], metrecom [26] or digital images [13]. However, as the ruler is placed at an angle, measurements might differ depending on the angle at which the examiner looks at the ruler. A 1.5 x 3-inch note card, as used by Sell et al. [27], might be better than a ruler. Using digital height gauge in measuring the navicular height for avoiding reading error could be most ideal.*” (Page 14 and15; Sentences 304-311)

We understand the reasoning of the reviewer, and some of our findings point out that our adaptations were not as adequate as expected, but this is what we did. We hope we explained which choices were made and why.
3. The lack of intra-rater reliability for both examiners (and the lack of time between the single intra-rater examinations)

- This is a very unusual and difficult omission to recover:
  - Lines 286 - 356: in many places the issue of only one set of intra-rater results present limitation to the ability to interpret the findings.
  - Line 286: it is stated that examiner experience is important in such reliability studies, yet the intrarater results (reflecting examiner consistency) are unavailable for one examiner.
  - Line 310: the lack of intra-rater results for both examiners disallows the effect of the NDT ruler vs card method as a factor in reduced inter-rater results (ie the examiner's individual consistencies are not able to be determined)
  - the methodology is scant regarding the role of the second examiner

We agree that it would be better to determine the intrarater reliability of examiner MW as well. However, we decided to provide only the intrarater reliability of examiner HD. We have chosen this option to limit the time involvement for the participating runners. The total time for the measurements of one runner was about one hour, including the breaks in between. If the other examiner had taken the tests twice for every runner, the randomization schedule had to be adapted and probably runners had to spend more than two hours while being measured. We already stated this point as a limitation in the last paragraph of the discussion on the MTP1-test; page 17, line 356-364. To further emphasis this shortcoming, we introduced this theme as a second point of limitations of this study.

“Secondly, by the possibility of using only one set of intra-rater results, the examiner consistency in our study was not optimal to determine. Future studies should include a minimal of two sets of intra-rater results so the degree of examiner consistency can be calculated and discussed.” (Page 19; Sentences 409-412)

Furthermore, we agree that it is more usual to separate sessions by hours rather than minutes to avoid bias. However, we choose a minimum of 10 minutes between sessions for practical reasons and did not expect test differences based on runners differences but only on measurement differences. Furthermore, we expected that recall bias was minimized with a minimum of 10 minutes between measurement sessions because of the number of measurements and the amount of runners: 3 tests, both feet and legs, with a minimum of 10 runners on every testing day.
So, we agree with the reviewer concerning the optimal design, however, we clarified the reasons for our choices and the resulting limitations.
4. **Additional concern is basing gender differences on a small convenience sample.**
   - This aim is unrealistic given the sampling, and sample size - suggest deleting all reference to gender differences throughout.

   Although, the convenience sample seems small (n=42) we found significant differences between males and females for the NDT Stance, NDT SL-S and MTPI-test (Table 3). Moreover, we aimed to provide descriptive information of our population in order to facilitate comparison with other populations including descriptive statistics of female and male runners. Therefore, we decided to keep the values on gender difference in our study.

5. **From here, I suggest that the author's review and re-think the wisdom of their study protocol - particularly if planning to use the same on a larger scale study.**
   - The use of standardized test protocols is advised, and I would suggest that NNH (normalized navicular height) be used in place of NDT. Additionally the authors could incorporate the reliability of their examiners using these tests, by repeating tests (having a second examiner do the same) for approx 40 of the larger study subjects. This larger study may also have sufficient power to explore gender differences. I strongly encourage the authors to re-evaluate the rational and design of any further investigations. Good luck with your future efforts.

   Thank you for your critical review and advice for our upcoming studies. We take your advice concerning standardization of the test protocol, the NNH and sample size into account for future research.

**Minor points**

Table 1 - length - I suggest this is 'height'

We changed “length” in “height” in Table 1 and Figure 1.

Table 3 - delete; as stated above, the sample and sample size does not provide a realistic basis for between gender mean difference (nor significance levels).

See above; point 4.