Reviewer's report

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

Version: 3  Date: 16 March 2014

Reviewer: Angela Evans

Reviewer's report:

Thank you for this further manuscript version, and whilst I appreciate the efforts the authors have made to improve aspects, I am still concerned regarding the rationale for this study, as previously stated in the first review. The main areas of concern which remain are:

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?

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 standardised and non-manipulated attributes eg AJ rom - which has been well tested previously using the standard weight-bearing lunge method. 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 standardised and non-manipulated attributes eg AJ rom - which has been well tested previously using the standard weight-bearing lunge method (Bennell).

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

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.

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 standardised test protocols is advised, and I would suggest that NNH (normalised 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.

Minor points:
Table 1 - length - I suggest this is 'height'

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).

**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