Author’s response to reviews

Title: Visual categorisation of the Arch Index: a simplified measure of foot posture in older people

Authors:

Hylton B Menz (h.menz@latrobe.edu.au)
Mohammad R Fotoohabadi (m.fotoohabadi@latrobe.edu.au)
Elin Wee (e.wee@latrobe.edu.au)
Martin J Spink (m.spink@latrobe.edu.au)

Version: 2 Date: 18 March 2012

Author's response to reviews: see over
REFEREE 1: RAVINDRA GOonetilleke

Major Compulsory Revisions

1. Resizing and how it was performed. Is it the same scale in both directions? Visually, to me, Figure 4 appears to have some awkward scaling; they do not look like real feet. It appears that resizing was not the same in the x- and y-directions possibly because the scaling was different. This will give rise an error in the calculation of AI. The mathematics of the procedure and the comparison of the AI values of the two images should be given and shown in a table.

RESPONSE: The images in Figure 4 were scaled in a computer graphics package so that the footprints were the same length. This was to ensure that when using the tool, assessors were not distracted by varying dimensions of the footprints, i.e. all other elements of the footprints were standardised except for the relative areas of the middle third of the footprint and the entire footprint. Because the entire footprint and the middle third of the footprint were scaled to the same degree, the AI, being a ratio, remains constant irrespective of scaling. As such, there is no need to report any “comparison” of AI scores between the two images.

2. Was AI calculated on the resized image or on the original footprint. I would like to see the differences in magnitude of AI for the two cases to illustrate that they do give the same value. Sample images of the two cases should also be shown.

RESPONSE: The AI scores were calculated from the original footprints. However, as stated above, the scaling process does not change the AI.

3. What do the authors mean when they say, “To ensure that examiners using the technique focused on the midfoot contour of the footprint,…”. Why only midfoot?

RESPONSE: We acknowledge that this wording may be ambiguous, and have changed it to:

“To ensure that examiners using the technique focused on the contours of the footprint, the selected footprints were resized to standard dimensions and provided with identical toe prints”.

4. Please make the following easier to read. “There were significant differences in mean AI across the AI categories documented by examiner 1 in session 1 (F2 = 85.6, p<0.001), examiner 1 in session 2 (F2 = 62.7, p<0.001), examiner 2 in session 1 (F2 = 80.9, p<0.001) and examiner 2 in session 2 (F2 = 74.3, p<0.001). All Bonferroni post-hoc tests across AI categories were significant at the p<0.001 level.” What did you compare in this? Is it the AI of what each examiner categorizes? If so, what is the purpose of the test?

RESPONSE: This test compares the mean AI values for the footprints across the ordinal categories of AI determined by the examiners. If the examiners’ categorisations are valid representations of the “true” AI score, then one would expect to see large differences in the mean AI values across the categories, which is what is shown by the ANOVA. To make this clearer, we have reworded this to:

“There were significant differences in mean AI scores obtained using the graphics tablet across the AI categories documented by examiner 1 in session 1 (F2 = 85.6, p<0.001),
examiner 1 in session 2 (F2 = 62.7, p<0.001), examiner 2 in session 1 (F2 = 80.9, p<0.001) and examiner 2 in session 2 (F2 = 74.3, p<0.001). All Bonferroni post-hoc tests across AI categories were significant at the p<0.001 level.”

5. **Please report the misclassifications of false alarms or misses and discuss why each type occurs. This may give some insight into the possible reasons for the mismatches.**

**RESPONSE:** We have added a table which reports the n (%) of misclassifications, which shows that examiners had a tendency to underestimate the true AI (i.e. score it as a higher arch). However, the overall frequency of misclassification was low (8 to 18%).

6. **The authors claim that the tool can be used in a clinical setting to quickly assess the arch. How can one perform the necessary scaling to compare against the tool without proper equipment?**

**RESPONSE:** We feel that the referee has misunderstood how the tool is to be used, as there is no need for the assessor to perform any scaling. Simply, all that is required is for the assessor to obtain a carbon paper imprint, visually compare it to the images shown in Figure 4, and select which category best matches the imprint.

7. **Please include a more acceptable reliability measure, such as ICC: SHROUT PE, FLEISS JL, Intraclass correlations: use in assessing operator reliability. Psychol Bull 86: 420, 1979.**

**RESPONSE:** Intraclass correlation coefficients are appropriate for assessing the reliability of continuously-scored data that is normally distributed. In our study, the comparisons made are either between two sets of categorical data or two sets of ordinal data. The appropriate statistical tests for assessing associations for such data are weighted kappas and Spearman’s rho, respectively.
REFEREE 2: KEITH ROME

1. Overall. The manuscript requires proof reading as there are minor typo and grammar errors.

RESPONSE: We have re-read the manuscript carefully and cannot find any typographical or grammatical errors. We would appreciate if the referee could specify these.

2. One could argue that two standard from the mean accounts for 95 percent from the mean. Please can the authors debate the issues using 1 SD from the mean accounting for 68 percent from the mean.

RESPONSE: We are unclear as to what the referee is suggesting with this statement. In a normally distributed histogram, ± 1 standard deviation from the mean encompasses 68% of cases, whereas ± 2 standard deviations from the mean encompasses 95% of cases. In our study, we defined “normal” AI as those scores that fell within ± 1 standard deviation from the mean (i.e. 68% of cases), which is the most common statistical definition of “normal”. See below:

![Histogram](Source: WikiMedia Commons)

Had we decided to define “normal” AI as ± 2 standard deviations from the mean, we would be left with only 5% of people with “abnormal” AI scores: 2.5% with an abnormally high AI, and 2.5% with and abnormally low AI. We feel that such a definition would not be particularly useful and does not reflect the prevalence of abnormal foot posture in the general population.

3. Delete the median score from the AI scores

RESPONSE: We would prefer to keep the median AI score within the paper, as it provides further evidence that the AI scores were normally distributed. In a perfectly normal distribution, the mean should be equal to the median, which is true in our dataset (AI = 0.24).

4. Please can the authors describe how and why 20 footprints were needed for this study?

RESPONSE: We used 60 footprints for the reliability component of the study (i.e. 20 for each category of low, normal and high), not 20. To make this clearer, we have reworded this section of the methods to:
“All participant’s AI scores were categorised as described above, and 60 footprints (20 footprints from each of the three categories: normal, high and low) were randomly selected”

According to a computer simulation by Cicchetti [1], the minimal sample size requirement for a reliability study using the weighted kappa statistic with 3 categories (as is the case with our study) is n=20. We are therefore confident that our sample of n=60 is sufficient.

5. Move ethical approval sentence under participant’s information.

RESPONSE: The ethics approval specified in the manuscript relates to the reliability study, which involved randomly selecting 60 footprints from the larger dataset (n=305). It is therefore more appropriate to report this under the “Reliability evaluation” subheading rather than under the “Development of the visual AI tool”, as the other two datasets used to develop the tool were obtained under different ethics approvals.

6. No need for the equation to be included

RESPONSE: The equation has been removed.

7. The argument that AI technique is a "gold standard" can be quite contentious. Please can the authors review the term?

RESPONSE: In validity studies, a “gold standard” is defined as “any standardised clinical assessment, method, procedure, intervention or measurement of known validity and reliability which is generally taken to be the best available, against which new tests or results and protocols are compared” [2] (emphasis added). In the context of this study, the “gold standard” is the AI calculated from the graphics tablet, and the “new” protocol is the AI category determined from visual assessment. However, this does not mean that we consider the AI to be the best measure of foot posture available. To make this clearer, we have added quotation marks and have reworded the relevant sentence in the discussion to:

“The tool performed very well, with AI categories demonstrating almost perfect inter- and intra-examiner reliability and exhibiting strong associations with both continuous and categorical AI scores obtained with a computer graphics tablet (the “gold standard” for this measure)”

8. The authors describe the examiners to have a "slight tendency" to categorize....Please reword

RESPONSE: In response to a suggestion from Referee 1, we have added a table which reports the n (%) of misclassifications, which shows that examiners had a tendency to underestimate the true AI (i.e. score it as a higher arch). However, the overall frequency of misclassification was low (8 to 18%). We have reworded this section of the discussion as follows:

“Although the examiners’ AI categories correlated very strongly with the AI scores obtained with the computer graphics tablet, some degree of misclassification did occur (see Table 1). Specifically, there was a tendency for examiners to categorise participants
as having higher arches than their AI scores indicated, with between 80 and 90% of misclassifications being caused by the examiners documenting the AI category lower than the AI category determined from the graphics tablet”.

9. The last sentence before the conclusions describes "other foot assessments".... Please can the authors be more explicit and give examples.

RESPONSE: We have reworded this as follows:

“Finally, the AI tool only provides a simple three-group categorisation of foot posture, so other foot assessments (such as the Foot Posture Index [10, 43] and foot mobility [44, 45]) may be more appropriate in situations where a greater degree of discrimination is required”

10. Add "for older adults" to the first sentence

RESPONSE: We have added “in older people”.

11. Other comments: I am unclear of the need for Figure 1 and Figure 3.

RESPONSE: We respectfully disagree with the referee. Figure 1 is necessary as it shows what the “real” footprints look like to the examiner, and Figure 3 provides important information regarding the distribution of AI scores, which forms the basis of the three group categorisation. We would therefore prefer to retain this information.

References cited in this response