Journal of Biomechanics 46 (2013) 2757–2760

Contents lists available at ScienceDirect

Journal of Biomechanics

journal homepage: www.elsevier.com/locate/jbiomech
www.JBiomech.com

Testing agreement between a new method and the gold standard—How do we test?

Pat McLaughlin [a,b,*]

[a] College of Health and Biomedicine, Victoria University, PO Box 14428 MCMC, Melbourne 8001, Victoria, Australia
[b] Institute of Sport, Exercise and Active Living, Victoria University, Melbourne, Australia

ARTICLE INFO

Article history:
Accepted 25 August 2013

Keywords:
Agreement
Relationship
Statistical techniques
Bland and Altman

ABSTRACT

Data analysis can be the most challenging aspect of a research study. Having been taught statistical techniques that tend to be based on finding significant differences or significant relationships, difficulties arise when trying to determine if a newly developed method is equally as good as the established method (the gold standard).

Testing for significant differences is rigorous and it would be rare for researchers to report significant differences without using an appropriate statistical test. Testing for agreement is assessed with far less rigour. Analysis of papers in this journal suggests that testing for agreement is an area that could be improved by a better understanding of statistical methods by biomechanics researchers. This perspectives paper focusses on informing the reader about the assessment of agreement between two methods.

1. Introduction

Determining whether two methods are providing the same information is a common question in biomechanics research, especially when researchers want to assess their 'new' method against an established 'gold standard'. After much time and effort is spent in developing the technical expertise/software/hardware involved with the new method, a decision on agreement often comes down to statistics. This is a challenging problem.

The tendency for statistical textbooks and courses to focus on standard statistical techniques creates a reliance on these same techniques by researchers. Researchers mostly learn about techniques to test difference (e.g. T-tests, ANOVAs) or test relationship (eg. Pearson's *r*, ICC) and then apply these techniques to all manner of problems. Even when the question at hand relates to determining whether two measurement methods are in agreement, we have a tendency to fall back onto the tests we know and apply them – even though these techniques may not be the most appropriate. Agreement decisions are still often made even though incorrect statistical techniques are used.

Does the new (faster/cheaper/better) method developed provide the same output as the established method? As a snapshot of how this question is answered in the Journal of Biomechanics, the author surveyed the last four issues of the Journal of Biomechanics in 2010 (issues 13–16). Sixteen papers stated aims relating to this type of hypothesis. The majority of these papers conducted statistical tests that were designed to test for significant differences (t-tests, ANOVA models, effect size) (Cherukuri et al., 2010; de Oliveira and Menegaldo, 2010; DeWitt, 2010; Espy et al., 2010; Miller et al., 2010; Morrow et al., 2010; Segal et al., 2010; Weinhandl and O'Connor, 2010), one tested for significant relationship (ICC, Pearson's r) (Sereysky et al., 2010), another assessed raw mean square error (RMSE) (Quinn and Winkelstein, 2010), a couple tested for agreement using the accepted Bland–Altman method (amongst other less appropriate tests) (Mariani et al., 2010; Schepers et al. 2010), and Kramer et al. (2010) and Faber et al. (2010) used alternative, less well known, forms of agreement testing (Borman et al., 2009 and Lotters and Burdof, 2002 respectively). Many authors used tests of difference and tests of relationship in an omnibus approach to try to capture the correct answer (e.g. Gonzalez-Izal et al., 2010; Mariani et al., 2010).

Of note was the tendency for some authors to:

- State an aim that indicates a test of agreement is required (new vs. old) – are the methods producing the same output?
- Conduct a test of significant difference between methods ($p < 0.05$)
- Report non-significant differences ($p > 0.05$)
- Use the non-significant result ($p > 0.05$) as verification that the methods are in agreement
- Conclude that the methods are in agreement and that the new method is equivalent to the old.

A non-significant difference is not an indication of equality, yet "….statistically equivalent results…" are suggested because no significant differences were reported (Morrow et al., 2010). More

[*] Corresponding author at: College of Health and Biomedicine, Victoria University, PO Box 14428 MCMC, Melbourne 8001, Victoria, Australia.
Tel.: +61 3 9919 1131; fax: +61 3 9919 1030.
E-mail address: patrick.mclaughlin@vu.edu.au

0021-9290/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jbiomech.2013.08.015

correctly "… failure to reject the test hypothesis does not automatically permit acceptance of the "null hypothesis" as true, since the chance of failing to reject a false hypothesis (*beta* or Type II error probability) in this case may be (and typically is) rather large (i.e. β is much greater than .05)" (Londeree et al., 1990; p. 276). In tests of difference there is a greater likelihood of a non-significant result (often due to low power based on a small sample size), but it is not a binary decision making system – a non-significant difference ($p > 0.05$) does not equate to significant agreement – although some researchers (as indicated above) tend to use it as such.

Tests of relationship (correlation) are also often (incorrectly) used to assess whether data sets from different methods are in agreement. It is reasonable to assume that any two methods designed to measure the same output will have some level of relationship (just as they will most likely not produce a significant difference). Assessing whether this relationship is significant ($p < 0.05$) is not the same as measuring whether the two methods are in agreement. Even a perfect correlation value of $r = 1$ does not indicate agreement between methods. However, there is evidence that this type of test is used (and accepted) to determine agreement (for a more recent example see Huurnink et al., 2013).

The assessment of measurement error (reliability) in sports science has been addressed in relation to repeated trials from individuals on the same task using the same equipment or same method (Atkinson and Nevill, 1998; Hopkins, 2000). These authors also briefly discuss the relevance of these statistical techniques to the question of agreement between two methods, and suggest it "…warrant further discussion amongst sports science researchers" (Atkinson and Nevill, 1998; p. 235). In relation to the specific topic of this paper, there is a generally accepted statistical technique. Bland and Altman (1986) described a graphical statistical technique combined with the calculation of 95% confidence limits (limits of agreement or LOA) for the assessment of agreement between two methods of measurement. These authors addressed the issue of "If the new method agrees sufficiently well with the old, the old, may be replaced" (p. 307). This method has been cited over 21,000 times throughout scientific literature but its use in the biomechanics literature is still minimal (ISI Web of Knowledge accessed 3rd April 2013). In all Journal of Biomechanics publications, this paper has been cited 43 times.

The present authors conducted an "Abstract, Title, Keyword" search (ISI Web of Knowledge) of all publications in the Journal of Biomechanics using the following search terms: accura*; agree*; compat*; equal*; equiv*; precis*; relat*; repeat*; same*; simil*; valid*. These search terms were used to determine whether authors who aimed to identify agreement as a key part of their study were using the accepted method of data analysis.

The data in Table 1 indicates the number of papers in the Journal of Biomechanics that used one of these terms in the "Abstract, Title, Keyword" and also indicates how many of these papers referenced the Bland and Altman (1986) method.

It is possible that the published papers' use of the above terms was not directly related to the method employed, and there is overlap of publications, but the trend is for biomechanists to make decisions on the agreement between systems using methods other than the accepted standard.

To illustrate the use of Bland and Altman and the risk of using the incorrect statistical technique, the following dataset is provided as an example of two different measurement methods being used to assess whether the new method is in agreement with the gold standard.

As part of a larger study, the author wanted to determine whether the centre of pressure (COP) output from a portable pressure mat system was equivalent to the output from a lab based force platform. Data were recorded synchronously by placing the pressure mat on top of the force plate and having participants stand on the mat. For the purposes of this exercise, the parameter of interest was maximum excursion of centre of pressure in the medio-lateral (COPx) direction. The research hypothesis was that the pressure mat output were statistically significantly equivalent to the force platform output (or the null hypothesis is that the systems are not significantly in agreement). The researcher's interest was to justify the use of the portable mat in field based testing.

Table 1
Journal of Biomechanics papers citing Bland and Altman (1986) by search terms.

Search term	# of papers	# that used Bland and Altman (1986)	Percentage that used Bland and Altman (1986)
Accura*	1139	26	2.3
Agree*	544	15	2.8
Compat*	33	0	0
Equal*	210	1	1
Equiv*	196	1	0.5
Precis*	254	11	4.3
Relat*	2942	21	0.7
Repeat*	342	14	4.1
Same*	705	3	0.4
Simil*	937	6	0.6
Valid*	899	18	2

* indicates search term that is open to all possibilities of spelling of key term after the initial letters - for e.g., "accura*" will capture "accuracy", "accurate".

Table 2
Descriptive output of datasets and statistical analyses of the output for COPx ($n = 14$).

	COPx	
	Pressure mat	Force plate
Mean ± SD (mm)	11.5 ± 6.5	11.4 ± 6.6
p_A		0.974
ICC		0.99*
Pearson's r		0.99*

SD = standard deviation, p_A = significance of one-way ANOVA, ICC = Intra-class correlation co-efficient.

* Significant at $p \leq 0.001$.

The measurement systems provided the following data sets ($n = 14$ samples for each). Table 2 presents the mean and standard deviation output from each system for COPx, and provides commonly reported statistical output used to answer the question of agreement (though the author will argue that these methods are not the most appropriate). The data are not significantly different ($p > 0.05$) and are highly correlated ($p < 0.001$).

The data relating the two methods in a scatter plot highlight the trend for highly correlated data for COPx ($r = 0.99$) (Fig. 1a). Using the Bland and Altman (1986) method, the differences between methods and the mean of methods for each pair of data can be calculated and plotted (Fig. 1b). The data lie around a line of zero difference.

In order to assess the agreement numerically, the mean and standard deviation of the differences (d) and 95% limits of agreement (LOA) are calculated. For these data then:

$$95\%\ \textit{limits of agreement}\ (95\%\ \text{LOA}) = \overline{d} \pm 1.96\ SD_d$$

$$\textit{where}\ \overline{d} = 0.08,\ \ SD_d = 0.34$$

This results in 95% LOA values of −0.59 to 0.76 mm which are acceptable LOA with no trend towards over or under estimation by either system. Given these data, and the output presented, the decision about agreement would be accurate using the commonly reported measures in Table 2, but it would not be valid.