Physiotherapy 100 (2014) 27–35

Pitfalls in the use of kappa when interpreting agreement between multiple raters in reliability studies

Shaun O'Leary [a,b], Marte Lund [a,c,1], Tore Johan Ytre-Hauge [a,d,1], Sigrid Reiersen Holm [a,e,1], Kaja Naess [a,f,1], Lars Nagelstad Dalland [a,g,1], Steven M. McPhail [h,i,*]

[a] *NHMRC Centre for Clinical Research Excellence in Spinal Pain, Injury and Health, University of Queensland, Brisbane, QLD 4072, Australia*
[b] *Physiotherapy Department, Royal Brisbane and Women's Hospital, Queensland Health, Herston, Brisbane, QLD 4029, Australia*
[c] *Norwegian Sports Medicine Clinic (NIMI), Oslo, Norway*
[d] *Medi 3 Clinic, Aalesund, Norway*
[e] *University Hospital of Northern Norway, Tromsø, Norway*
[f] *Hans & Olaf Physiotherapy Clinic, Oslo, Norway*
[g] *Eggedal Physiotherapy Clinic, Sigdal, Norway*
[h] *Centre for Functioning and Health Research, Queensland Health, Cnr of Ipswich Road and Cornwall Street, Brisbane, Australia*
[i] *School of Public Health and Institute of Health and Biomedical Innovation, Queensland University of Technology, Victoria Park Road, Brisbane, Australia*

Abstract

Objective To compare different reliability coefficients (exact agreement, and variations of the kappa (generalised, Cohen's and Prevalence Adjusted and Biased Adjusted (PABAK))) for four physiotherapists conducting visual assessments of scapulae.
Design Inter-therapist reliability study.
Setting Research laboratory.
Participants 30 individuals with no history of neck or shoulder pain were recruited with no obvious significant postural abnormalities.
Main outcome measures Ratings of scapular posture were recorded in multiple biomechanical planes under four test conditions (at rest, and while under three isometric conditions) by four physiotherapists.
Results The magnitude of discrepancy between the two therapist pairs was 0.04 to 0.76 for Cohen's kappa, and 0.00 to 0.86 for PABAK. In comparison, the generalised kappa provided a score between the two paired kappa coefficients. The difference between mean generalised kappa coefficients and mean Cohen's kappa (0.02) and between mean generalised kappa and PABAK (0.02) were negligible, but the magnitude of difference between the generalised kappa and paired kappa within each plane and condition was substantial; 0.02 to 0.57 for Cohen's kappa and 0.02 to 0.63 for PABAK, respectively.
Conclusions Calculating coefficients for therapist pairs alone may result in inconsistent findings. In contrast, the generalised kappa provided a coefficient close to the mean of the paired kappa coefficients. These findings support an assertion that generalised kappa may lead to a better representation of reliability between three or more raters and that reliability studies only calculating agreement between two raters should be interpreted with caution. However, generalised kappa may mask more extreme cases of agreement (or disagreement) that paired comparisons may reveal.

Keywords: Reliability; Kappa; Scapular; Posture; Inter-therapist; Agreement

* Correspondence: Centre for Functioning and Health Research, PO Box 6053, Buranda 4102, Brisbane, Australia. Tel.: +61 7 3406 2266; fax: +61 7 3406 2267.
E-mail addresses: shaun_oleary@health.qld.gov.au (S. O'Leary), marte.lund@nimi.no (M. Lund), toreyh@hotmail.com (T.J. Ytre-Hauge), sigrid.r.holm@gmail.com (S.R. Holm), kaja.nass@hof.nhn.no (K. Naess), larsdalland@hotmail.com (L.N. Dalland), steven_mcphail@health.qld.gov.au, steven.mcphail@qut.edu.au (S.M. McPhail).
[1] These authors contributed equally to this work.

0031-9406/$ – see front matter
http://dx.doi.org/10.1016/j.physio.2013.08.002

Introduction

Clinical decision making is often based on examination findings that are dependent on subjective nominal ratings of status such as that used to evaluate posture alignment [1–4]. Investigating the agreement between physiotherapists using these nominal clinical measures is challenging not only due to the subjective nature of the ratings, but also due to the limitations of reliability coefficients that are utilised to express agreement within and between therapists [5,6]. The most widely used reliability coefficient for nominal data include kappa coefficients [7–10].

The magnitude of the kappa coefficient represents the proportion of agreement between or within raters greater than that expected by chance; with a coefficient of 1.00 representing perfect agreement [7]. Perhaps the most frequently used type of kappa coefficient is Cohen's kappa [7–10]. One advantage of utilising Cohen's kappa to examine agreement between raters is the ability to use a weighting system to penalise disagreements of larger magnitude more than disagreements of a smaller magnitude. In this way, a disparate disagreement will contribute to a lower kappa coefficient more than a disagreement of smaller magnitude (in ordinal data sets with three or more levels) [7,9].

While guidelines have been proposed to interpret the magnitude of the kappa with respect to clinical utility [11], the magnitude of the kappa can be influenced by several factors including prevalence (of responses in each category) and bias within the data. These factors can result in seemingly paradoxical observations of high exact agreement between raters, but low kappa coefficients [12,13]. There have been suggested strategies to account for these issues with the kappa such as a methodology for adjusting kappa coefficients for prevalence and bias [14]. However, this approach has also been criticised as representing an artificial coefficient when the prevalence of ratings in each category and bias present in a dataset reflect real life occurrence [15]. Perhaps the most significant limitation to the use of Prevalence Adjusted Bias Adjusted Kappa (PABAK), and Cohen's kappa, is that they are only appropriate for examination of agreement between two raters; and not appropriate for three or more raters [5–7,15].

When evaluating agreement between multiple raters generalised kappa has been recommended [16]. However, the generalised kappa has at least two potential limitations. First, it does not permit adjustment for prevalence and bias within the kappa calculation. Second, it does not permit weighting to penalise disagreements of a larger magnitude. This inability to weight disparate disagreement is a limitation when a generalised kappa is derived from ordinal data with three or more levels. Due to the inherent limitations with the use of the general kappa there appears to be inconsistency with the statistical approached utilised in medical research assessing agreement between multiple raters. While some reliability studies with multiple raters have reported a general kappa statistic [17], others have only compared pairings within multiple raters using multiple Cohen's kappa's [18], or have provided both [19]. Intuitively it seems desirable to have knowledge of the reliability of a measure when multiple raters are evaluating as opposed to only two raters to support the ability to generalise findings from reliability studies.

The purpose of this study was to compare kappa coefficients calculated using a generalised versus multiple paired (when also weighted and prevalence/bias adjusted) kappa approaches when determining inter-therapist reliability for multiple raters rating scapular posture in healthy individuals. Typically an examination of the upper quadrant in the clinical setting includes visual inspection of scapular posture [20,21] with observers making judgements as to deviations from that considered to constitute normal scapular posture. We anticipate that this study will provide valuable empirical data to promote transparency as to the implications of utilising these different approaches to calculating agreement between multiple raters when utilising nominal clinical assessment tools in research and clinical settings.

Methods

Design

Inter-therapist reliability coefficients for four qualified and registered physiotherapists independently recording nominal ratings of scapular posture were calculated and compared using generalised and multiple paired rater approaches. These raters had one to four years of experience and were undertaking a post-graduate master's degree in musculoskeletal physiotherapy studies. Subjects were blinded to the intention of the study.

Participants and setting

A convenience sample of 15 subjects ($n = 30$ scapulae) with no history of neck or shoulder pain were recruited. A small convenience sample was utilised in order to replicate the sampling approach for many reliability studies that are conducted in this field of study [22–24]. These subjects included nine women and six men and had a mean (standard deviation) age of 28.8(4.4) years, and a mean body mass index of 21.8(2.1) kg/m^2. Participants were recruited from advertising within the university and community. Potential participants were excluded if they presented with obvious significant postural abnormalities such as a severe thoracic kyphosis or spinal scoliosis. The study was undertaken in a clinical research laboratory setting. Ethical approval was granted by the Institutional Human Research Ethics Committee and procedures were conducted according to the Declaration of Helsinki. Participants provided informed consent.

Outcomes

Scapula posture was rated in five different postural planes (Table 1 and Fig. 1) [25]. Therapists visualised the