European Journal of Orthopaedic Surgery & Traumatology (2019) 29:1833–1834
https://doi.org/10.1007/s00590-019-02493-5

LETTER TO THE EDITOR • HIP - FRACTURES

Methodological issue on reliability of the commonly used classification systems for interprosthetic fractures

Mehdi Naderi[1] · Siamak Sabour[2,3]

Received: 2 July 2019 / Accepted: 2 July 2019 / Published online: 29 July 2019

Dear Editor,

We were interested to read an article that recently published by Jennison T and colleagues in the July 2019 issue of Eur J Orthop Surg Traumatol [1]. The purpose of the authors was to investigate the interobserver and intraobserver reliability of the most commonly used interprosthetic fracture classifications [1]. The intraobserver and interobserver errors were assessed with Cohen's *k* coefficient. Nineteen interprosthetic fractures were classified by four reviewers for inter- and intraobserver reliability. The interprosthetic fracture classifications used for this purpose were the Soenen classification, Platzer classification and Pires classification. Based on the authors' results, the kappa value and interobserver reliability for all the classification systems ((Platzer classification = 0.586), (Pires classification = 0.499) and (Soenen classification = 0.489)) were moderate. Also, the kappa value and intraobserver error for Platzer classification, Pires classification and Soenen classification were 0.767 (substantial agreement), 0.636 (substantial agreement) and 0.318 (fair agreement), respectively.

It's crucial to know that to assess agreement of a qualitative variable, applying Cohen's *k* coefficient is not always an appropriate test. First, the kappa value depends on the prevalence in each category. Second, it also depends on the number of categories [2–7]. We should mention that when a variable with more than two categories or an ordinal scale is used (with 3 or more ordered categories), then the weighted kappa would be a good choice. Finally, another important flaw is that the two raters have unequal marginal distributions of their responses [2–7]. Table 1 shows the agreement by applying kappa (0.43 as moderate) and weighted kappa (0.63 as good) which have different values and consequently different interpretations. In this table, the marginal distribution in the first category (grade 1) is different from the other categories, and also, the number of categories is more than two.

The authors concluded that there are a moderate interobserver reliability and significant intraobserver reliability for both Platzer and Pires classifications.

In this letter, we discussed important limitations of applying Cohen's *k* coefficient to assess reliability [2–7]. Any conclusion in reliability analyses needs to be supported by the methodological and statistical issues mentioned above. Otherwise, misinterpretation cannot be avoided.

Table 1 The kappa and weighted kappa values for calculating agreement between two observers for more than two categories depending on the prevalence

		Observer 1			Sum
	Grade	1	2	3	
Observer 2	1	60	20	1	81
	2	2	12	4	18
	3	3	11	11	25
Sum		65	43	16	124
	Estimate				
Kappa	0.43				
Weighted kappa	0.63				

✉ Siamak Sabour
s.sabour@sbmu.ac.ir

[1] Clinical Research Development Centre, Taleghani and Imam Ali Hospital, Kermanshah University of Medical Sciences, Kermanshah, Islamic Republic of Iran

[2] Department of Clinical Epidemiology, School of Health and Safety, Shahid Beheshti University of Medical Sciences, Chamran Highway, Velenjak, Daneshjoo Blvd, Tehran 198353-5511, Islamic Republic of Iran

[3] Safety Promotions and Injury Prevention Research Centre, Shahid Beheshti University of Medical Sciences, Tehran, Islamic Republic of Iran

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Jennison T, Jawed A, ElBakoury A, Hosny H, Yarlagadda R (2019) Reliability of the commonly used classification systems for interprosthetic fractures. Eur J Orthop Surg Traumatol 29(5):1069–1072
2. Szklo M, Nieto FJ (2014) Epidemiology Beyond the Basics, 3rd. Jones and Bartlett Publisher, Manhattan
3. Sabour S (2014) Methodologic concerns in reliability of noncalcified coronary artery plaque burden quantification. AJR Am J Roentgenol 203(3):W343
4. Naderi M, Sabour S (2019) Inter and intraobserver reliability and critical analysis of the FFP classification of osteoporotic pelvic ring injuries: methodological issue. Injury 50(6):1261–1262
5. Naderi M, Sabour S (2019) Methodological issue on reliability of 9.4T MRI in the assessment of degenerative disc disease compared to 3T MRI. Spine (Phila Pa 1976) 15 44(10):E629–E630
6. Sabour S (2016) Reliability of immunocytochemistry and fluorescence in situ hybridization on fine-needle aspiration cytology samples of breast cancers: methodological issues. Diagn Cytopathol 44(12):1128–1129
7. Sabour S, Dastjerdi EV (2013) Reliability of four different computerized cephalometric analysis programs: a methodological error. Eur J Orthod 35(6):848