Methodological issues on interobserver reliability of radiographic assessment after radial head arthroplasty

Mehdi Naderi¹ and Siamak Sabour^{2,3}

Dear Editor,

We were interested to read the paper by Bexkens et al.¹ published in the April 2018 issue of *Shoulder & Elbow*. Radiographs are part of routine clinical care after radial head arthroplasty (RHA). The authors aimed to assess the interobserver reliability of radiographic assessment following RHA. The anteroposterior (AP) and lateral radiographs of 24 patients on 14 parameters were reviewed by the 14 surgeons. Observer agreement measurement was assessed by kappa statistic.¹ They reported that nine parameters and nonbridging heterotopic ossification had poor interobserver agreement ($k=0.0-0.20$) and moderate agreement ($k=0.47$), respectively, and also subcollar bone resorption ($k=0.27$), capitellar erosion ($k=0.30$), ulnohumeral joint degeneration ($k=0.35$) and stem positioning in AP view ($k=0.24$) had fair agreement.

It is of crucial importance to know that no value of kappa can be universally known as a sign of good agreement. To assess agreement of a qualitative variable, the k value has two major weaknesses: (1) it depends on the prevalence in each class, i.e. we can get different kappa values of the same percentages for concordant and discordant cells! As can be seen in Table 1, the prevalence of concordant cells in both (a) and (b) situations is 90% while that of discordant cells is 10%. However, we get different values of kappa (0.44 and 0.80, respectively). (2) Kappa value also depends on the number of categories.²⁻⁹ To obtain unbiased results, it is preferable to use weighted kappa in such a situation.

They concluded that the overall interobserver reliability of radiographic assessment following press-fit bipolar RHA was poor among experienced elbow surgeons. Therefore, radiographic evaluation after RHA should be interpreted with caution when making treatment decisions.

Considering the above mentioned limitations of the kappa value to assess reliability, such a conclusion might be a misleading message.

Table 1. Limitation of k to assess reliability of two surgeons' diagnoses with different prevalence in the two categories.

		Positive	Surgeon 1 Negative	Total (%)
Situation (a)	Positive	85	5	90
Surgeon 2	Negative	5	5	10
$k = 0.44$	Total	90	10	100
Situation (b)	Positive	45	5	50
Surgeon 2	Negative	5	45	50
$k = 0.80$	Total	50	50	100

Bold indicates two different kappa value having the same percentages for both agreement and disagreement cells.

In this letter, we discussed two important limitations of the kappa value to assess reliability. Any conclusion in reliability analyses needs to be supported by the methodological and statistical issues mentioned above.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

¹Department of Operating Room, School of Paramedical, Kermanshah University of Medical Sciences, Kermanshah, Islamic Republic of Iran

²Department of Clinical Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, Islamic Republic of Iran

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

Corresponding author:

Siamak Sabour, Department of Clinical Epidemiology, School of Health, Safety Promotion and Injury Prevention Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Islamic Republic of Iran.
Email: s.sabour@sbmu.ac.ir

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Bexkens R, Claessen FMAP, Kodde IF, et al. Interobserver reliability of radiographic assessment after radial head arthroplasty. *Shoulder Elbow* 2018; 10: 121–127.
2. Szklo M and Nieto FJ. *Epidemiology beyond the basics*, 2nd ed. Manhattan: Jones and Bartlett Publisher, 2007.
3. Sabour S and Dastjerdi EV. Reliability of four different computerized cephalometric analysis programs: a methodological error. *Eur J Orthod* 2013; 35: 848.
4. Sabour S. Spinal instability neoplastic scale: methodologic issues to avoid misinterpretation. *AJR Am J Roentgenol* 2015; 204: W493.
5. Sabour S. Adherence to guidelines strongly improves reproducibility of brachial artery flow-mediated dilation. Common mistakes and methodological issue. *Atherosclerosis* 2016; 251: 490–491.
6. Sabour S. Reliability assurance of EML4-ALK rearrangement detection in non-small cell lung cancer: a methodological and statistical issue. *J Thorac Oncol* 2016; 11: e92–e93.
7. Sabour S. Reliability of immunocytochemistry and fluorescence in situ hybridization on fine-needle aspiration cytology samples of breast cancers: methodological issues. *Diagn Cytopathol* 2016; 44: 1128–1129.
8. Sabour S. Reliability of a new modified tear breakup time method: methodological and statistical issues. *Graefes Arch Clin Exp Ophthalmol* 2016; 254: 595–596.
9. Sabour S. Reproducibility of endometrial cytology by the Osaki Study Group Method: methodological issues. *Cytopathology* 2017; 28: 441–442.