

CALIBRATION AND KAPPA

Andrew Blance¹, Joanna Carvalho² and Mark S. Gilthorpe¹

¹University of Leeds, UK, ²Catholic University of Louvain, Belgium

Calibration exercises are common in oral health research. When the scale concerned is categorical, this usually involves the use of Kappa. Although well-documented, problems concerning Kappa are still not widely appreciated. In particular, confidence intervals associated with Kappa are usually very wide and this can have implication in the calibration exercises often undertaken by clinical researchers. This paper aims to investigate the lack of robustness that might arise when using Kappa in calibration exercises. We show how the concept of successful calibration might be concluded erroneously.

Caries diagnosis at tooth surface level was observed using a 10-point ordinal scale. 128 observations were made for each of 26 subjects, by 26 dentists (observers). Thus, 3328 observations were made per observer. Pairwise Kappa's for all 3328 observations ranged from 0.49 to 0.82, indicating that agreement was "*moderate*" to "*very good*". A hypothetical calibration study based on this dataset was conducted. A pair of observers was randomly selected and then ten pairs of observations were randomly selected and calibration was deemed to have occurred if Kappa exceeded 0.7. If calibration was deemed unsuccessful, another ten pairs were selected (with replacement). This was repeated until calibration occurred, recording the number of iterations taken. A total of 10,000 observer pair selections were performed.

It is important to note that nothing changed among the observers within this study; no more training had taken place and the true level of calibration remained constant under these hypothetical circumstances throughout the simulated calibration. It is intended that the hypothetical simulation mirrors the clinical process of calibration; i.e. if calibration is deemed unsuccessful, further training takes place and calibration is re-assessed. In the hypothetical study, successful calibration was deemed to have occurred half the time within only four iterations, with 75% successful calibration concluded within seven iterations.

This study shows that under the assumption that no change/ improvement occurred, an acceptable value of Kappa can often be obtained within a relatively small number of repeat samples. Thus, under the real circumstances of a genuine calibration (with possible improvement in agreement through repeated iterations of training), it would be impossible to distinguish true successful calibration from erroneous "success" due to the imprecision of Kappa. Conversely, attaining a high percentage of "calibrated" clinicians by means of such an exercise could be very misleading. The training exercises may have had no effect and yet within only a small number of iterations, one would eventually conclude success for the vast majority of those undertaking calibration.

The clinical implication of using Kappa in calibration studies is that calibration may be erroneously deemed successful if the criterion of success is for the point estimate of Kappa to exceed a pre-specified threshold. A more appropriate criterion would be to seek non-inferiority to one.