University of York Department of Health Sciences

Measurement in Health and Disease

Exercise: Kappa statistics

1. The following is the abstract of a paper.

Study Design. Multicenter, prospective equivalency trial with each patient serving as his/her own control.

Objectives. To compare the effectiveness of a Grafton(R) DBM gel composite with iliac crest autograft in posterolateral spine fusion.

Summary of Background Data. While autograft remains the preferred graft material to facilitate spine fusion, the supply is limited and harvesting produces undesirable clinical consequences.

Methods. A total of 120 patients underwent posterolateral spine fusion with pedicle screw fixation and bone grafting. Iliac crest autograft was implanted on one side of the spine and a Grafton(R) DBM/autograft composite was implanted on the contralateral side in the same patient. An independent, blinded reviewer evaluated anteroposterior and lateral flexion-extension radiographs. The fusion mass lateral to the instrumentation on each side was judged fused or not, and the mineralization of the graft was rated absent, mild, moderate, or extensive. The degree of correspondence in outcomes between sides was estimated by computing the percentage agreement and kappa statistic.

Results. Nearly 70% of patients (81 of 120) provided complete 24-month radiographic studies. The bone graft mass was fused in 42 cases (52%) on the Grafton(R) DBM side and in 44 cases (54%) on the autograft side. The overall percentage agreement for fusion status between sides was approximately 75% (61 of 81), indicating moderately strong statistical correspondence (kappa = 0.51, P<0.0001). Bone mineralization ratings also were similar between treated sides. Perfect agreement was realized in almost 60% of patients (48 of 81) with moderate statistical correspondence (weighted kappa = 0.54, P<0.0001).

Conclusions. Grafton(R) DBM can extend a smaller quantity of autograft than is normally required to achieve a solid spinal arthrodesis. Consequently, a reduced amount of harvested autograft may be required, potentially diminishing the risk and severity of donor site complications.

(Cammisa FP, Lowery G, Garfin SR, Geisler FH, Klara PM, McGuire RA, Sassard WR, Stubbs H, Block JE. Two-year fusion rate equivalency between Grafton (R) DBM gel and autograft in posterolateral spine fusion. *Spine* 2004; **29**: 660-666.)

QUESTIONS

a) For fusion status, kappa = 0.51. What does this mean and what conclusions could we draw?

b) Why is kappa less than the proportional (percentage) agreement for fusion status?

c) What hypothesis is the 'P<0.0001' testing? What does it tell us?

d) Why was weighted kappa used for bone mineralization ratings?

2. The following is the abstract of a paper.

Aim: The aim of this preliminary study was to test the reliability of radiographic evaluation of features of endodontic interest using a newly devised data collection system

Methodology: Twelve endodontic MSc postgraduate students and one specialist endodontist examined sample radiographs derived from a random selection of 42 patients seen previously on an Endodontic New Patient Clinic (EDI). Each student examined a random selection of 8-9 roots on periapical radiographs of single- and multirooted teeth, with and without previous root canal therapy and 3-4 dental panoramic tomograms (DPTs). A total of 100 roots were examined. A proforma was used to record observations on 67 radiographic features using predefined criteria. Intra-observer agreement was tested by asking the students to re-examine the radiographs. The principle investigator and the specialist endodontist examined the same radiographs and devised a Gold Standard using the same criteria. This was compared with the student assessments to determine inter-observer variation. The postgraduates then attended a revision session on the use of the form. Each student subsequently examined 8-9 different roots from the pool of radiographs. A further assessment of inter-observer variation was made by comparing these observations with the Gold Standard.

Results: Of the 67 radiographic features, only 25 had sufficient response to allow statistical analysis. Kappa values for intra- and inter-observer variation were estimated. These varied depending on the particular radiographic feature being assessed. Fifteen out of 25 intra-observer recordings showed 'good' or 'very good' Kappa agreement, but only three out of 25 inter-observer observations achieved 'good' or 'very good' values. Inter-observer variation was improved following the revision session with 16 out of 25 observations achieving 'good' or 'very good' Kappa agreement.

Conclusions: modification to the proforma. the criteria used, and training for radiographic assessment were considered necessary to improve the accuracy and reproducibility of the observations entered.

(Saunders MB, Gulabivala K, Holt R, Kahan RS. Reliability of radiographic observations recorded on a proforma measured using inter- and intra-observer variation: a preliminary study. *International Endodontic Journal* 2000; **33**: 272-278.)

NOTE

It is not very clear what they did here. What they have done is to get each student to observe 8 or 9 teeth only. The 12 students thus produced one obervation between them of each of the 100 teeth, not one observation by each student on each tooth. The students repeated their observations on the same teeth. This is what was used for intra-observer variation. The principle investigator and the specialist endodontist agreed for each of the 100 teeth what the 'correct' answers should be, which they called the Gold Standard. This is what they used for inter-observer variation. They then held a training session. They repeated the student observations, but this time each student looked at a different set of the 100 teeth.

QUESTIONS

e) What is meant by "good" or "very good" Kappa agreement'?

f) What do the kappas tell us about initial intra-observer and inter-observer agreement?

g) Why did inter-observer agreement improve after training?

h) How would you change the design to estimate inter-observer agreement between the students?