Reviewer's report

Title: Determining Utility Values in Patients with Anterior Cruciate Ligament Tears using Clinical Scoring Systems

Version: 2 Date: 26 April 2011

Reviewer: Marek Brabec

Reviewer's report:

This is an interesting piece of research whose findings have strong potential for being useful for practical evaluation of different ACL treatment strategies. The statistical analysis is based on rather simple and classical tools. Testing for normality on such a small data in order to decide whether to use Pearson or Spearmann correlation coefficient is almost too much. But that is a rather minor issue related to style and taste.

The main problem lies in the fact that the data are not independent, so that they should not be treated by the standard tests assuming independence. From the general description around pages 6-7, it seems that each of the 27 orthopedic surgeons participating in the study was presented four vignettes. Since there will be random differences among surgeons (caused by their individually different psychology, perception, experience, mood etc.), there will be correlation among the 4 values of the same variable obtained from a particular surgeon, while different surgeons will be probably more or less uncorrelated. This is a situation that occurs in practice quite often and that calls for a statistical treatment. For decades, people use linear mixed models (LME) in such situation (with individual surgeon specific random intercept and fixed effect of vignette type), or at least the pragmatic correction in the GEE (generalized estimating equation) style, based on Huber-White sandwich estimate of the covariance matrix of the estimates. Both of these are nowadays widely available in many popular statistical software packages. The point is that without LME or GEE adjustments, the estimated residual variance tends to be too small, leading to p-values that tend to be too low (i.e. they tend to yield too many significant results). It will be necessary to re-compute the analyses properly to see whether the differences detected among the vignettes will remain even after the adjustments or whether the appropriate correlation treatment will dilute them. That will have to be done for all scores studied.

Further, the authors should explain more the rationale behind the particular weighting scheme they used (“The opinions of the orthopedic surgeons who had treated more than 100 ACL patients in their career (n=9) were counted as twofold and those from experts who had taken care of more than 300 ACL patients (n=2) were counted threefold”). While it seems to be reasonable to weight those that have more experience, it is not clear why to use the ratio 1:2:3 and not something else. If there is no strong or theoretical rationale for this particular choice, then, at the bare minimum, the authors should demonstrate that other more or less rational weighting choices will not change the conclusions.
The study works with a sample of local surgeons coming from one hospital. That is openly acknowledged at the beginning of the paper. It is OK to perform the study in this way as a sort of pilot study and publish its results. But then, the fact should also be stressed among the weaknesses of the study. In future, a similar research should be repeated more broadly, in a multicenter study to see whether the findings could not be somehow influenced by the fact that the surgeons here came from the same institution with the same culture, standards etc. Even if this institution would not be systematically different from other institutions, inter-institutional differences might still be viewed as random. How much variability is there between institutions, as compared to the within institutional variability? These and other important questions can be answered only from the multicenter study data when they would be analyzed properly (adding random hospital effect at another level of hierarchy to the effect of surgeon within hospital that should be already present in this study).