Reviewer's report

Title: Pitfalls in the statistical examination and interpretation of the correspondence between physician and patient satisfaction ratings: an example from shared decision making research

Version: 1 Date: 17 February 2011

Reviewer: Peter T Sawicki

Reviewer's report:

Review of the manuscript:
„Pitfalls in the statistical examination and interpretation of the correspondence between physician and patient satisfaction ratings: an example from shared decision making research”. Hirsch O. et al.

Review by Anna Drabik and Peter T. Sawicki, MD, PhD.
Institute of Health Economics and Clinical Epidemiology, University of Cologne, Germany.

Major comments:
In this manuscript Hirsch et al. analyse different ways to describe the quality and magnitude of satisfaction with shared decision making processes using the arribaTM tool. Statistical evaluation suggests large differences between patients’ and physicians’ satisfaction ratings. The authors, however, conclude that statistical measures of association are unable to sufficiently describe dimensions of interobserver agreement.

One of the aims of this study seems to be to evaluate the correspondence between satisfaction ratings of physicians and patients with a shared decision making process in primary care. If this was the objective of the study, it remains unexplained why a control group without shared decision making intervention was used. Was it also the author’s aim to compare the correspondence between patient and physician satisfaction with and without shared decision making? If this is the case, why are these results then not reported in full? And how did the authors prevent the control group from using other tools or unstructured approaches to shared decision making? How have the authors ensured that the intervention group and the control group had similar baseline characteristics?

The initial hypothesis of this study remains unclear. Did the authors assume that the use of shared decision making will change the agreement in satisfaction ratings between patients and physicians? Or was the hypothesis that shared decision making increases satisfaction rates both in physicians and patients?

At the end of the “Background” section the authors state: “The aim of our study was to evaluate the correspondence between patient and physician satisfaction ratings with the shared decision making process in primary care.” At the beginning of the “Discussion” section, however, the authors most interestingly
mention a different aim: “The aim of the study was to evaluate the correspondence between patient and physician satisfaction with the shared decision making process by different statistical procedures. “ For a reviewer, it is therefore difficult to decide what the primary aim of this study was and whether appropriate methods were used to reach this aim.

The authors do not mention how the statistical methods were selected and evaluated. What was used as a gold standard and why? How did the authors decide whether a statistical method is appropriate or not?

In addition and independent of the goal of the study, we do not think that the methods used and results obtained are entirely valid. On page 5 the authors write: „Since the data supplied by the PPS can be regarded as being on the borderline between an ordinal and a metric scale of measurement, we used procedures appropriate for both.“

The data used were derived from a 5-point-Likert scale. For several reasons, such a scale can clearly be seen as ordinal. The most important argument is, that a discrete variable takes on a finite or denumerable set of values, whereas a continuous random variable takes on a non-denumerable set of values [1]. A 5-point-Likert scale surely doesn’t consist of a non-denumerable set of values and for this reason it isn’t a metric scale. Further arguments are presented in the literature [2].

As pointed out above, it is not appropriate to use such statistical methods for continuous variables. In the article, this applies to the use of the Bland-Altman method, as this method includes statistics that are only valid for continuous variables.

Cohen’s Kappa is the most popular measure for determining the degree of agreement between two raters and although there are some drawbacks, it may also be appropriate for this analysis. The authors do not sufficiently explain why they are not satisfied with this approach.

The percentage of agreement can additionally be given, as well as a model for agreement between the ratings. It can also be fitted as proposed by Agresti (1988) [3]. We suggest that the authors consider these further possibilities for data evaluation.

Furthermore, the Wilcoxon signed-rank test for matched pairs might be the right choice to determine whether there is a difference between both ratings but its null hypothesis is that the median difference between pairs of observations is zero and not the mean difference between pairs. Due to this fact and because means and standard deviations are presented in table 3 as well as in the text, the presented results are not entirely correct. It remains unclear to us, whether the Wilcoxon signed-rank test has been applied in this case.

Minor comments:
Page 8, ll. 5-6: “This is counterintuitive, considering the small mean differences on the PPS.” This does not belong in the results section, but should be stated in the discussion section. Moreover, considering the large sample size, this result isn’t counterintuitive.
Page 8, ll. 10-11: “…which again is counterintuitive because of the relatively small mean differences.” See comment above.

Literature

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
We declare that we have no competing interests.