Author's response to reviews

Title: The Null Hypothesis Significance Test in Health Sciences Research (1995-2006): Statistical Analysis and Interpretation

Authors:

    Luis Carlos Silva-Ayçaguer (lcsilva@infomed.sld.cu)
    Patricio Suárez-Gil (patricio.suarez@sespa.princast.es)
    Ana Fernández-Somoano (afsomo@yahoo.es)

Version: 3 Date: 13 April 2010

Author's response to reviews: see over
Dear Colleagues of The BioMed Central Editorial Team:

We have received your note dated 2010/4/1 concerning our paper The Null Hypothesis Significance Test in Health Sciences Research (1995-2006): Statistical Analysis and Interpretation (MS: 287874907306160) and the comments made by one of the reviewers.

After our answers and corrections to 47 issues raised by the two initial reviewers, only 6 of them from one of the reviewers still demand corrections from us.

Following is a point-by-point response to the concerns expressed in these 6 comments. A revised version of the manuscript has been uploaded, as requested.

**COMMENT 1**

**Original comment:** Why did you decide to take 60 papers for each biennium and journal?

**Original Response:** It seems to be only a question, not a claim for us to introduce any change. Our answer is that we had no especial reason to take 60 papers. Since confidence intervals were to be computed for each estimation (as we actually did), the way to determine the sample size is not so important. We thought that around 1000 papers was good enough. We preferred to use our common sense, as recommended by Kenneth Rothman (Rothman JK. Modern Epidemiology., 1982, Boston: Little, Brown and Col.), to take into account the available time for us to devote to this task and avoid the “sample size samba” and retrofitting of the parameter estimates denounced by Schulz and Grimes (Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet. 2005; 365:1348-53).

**New Comment:** This is very reasonable. It is desirable, however, to provide a brief justification for your decision in the manuscript.

**Present Response:** Ok, we have inserted the following text:

We decided to take around 60 papers for each biennium and journal, which means a total of around 1000 papers. As recently suggested [36] [37], it was not a requirement to use conventional sample size planning, but to choose the maximum sample size that was feasible by means of a pragmatic perspective.

Although the reviewer fully agrees with our point of view (and to some extent, due to this fact), we especially suggest her to read the second paper, where we have found very clever ideas like the following ones: “The idea that a carefully chosen “adequate” sample size can ensure that a study will be definitive, reflected in the common phrase ‘enough to answer the question’, is certainly appealing, but it is just wishful thinking.”…. “When constraints imposed by funders determine feasibility, doing the maximum possible within those constraints is a sensible choice.”…. “Reviewers should usually refrain from criticizing sample size and should challenge fellow reviewers who do.”

The references above mentioned (and now added) are:


**COMMENT 2**

**Original comment:** The operational definition used to label a paper as having the problem of
significance fallacy is not clearly stated in the methods section. Please, explain how did you operationalize its assessment.

**Original Response:** The significance fallacy was defined in the paper as “the deficiency that consists of attributing substantive significance to the found differences simply because statistical significance was obtained”. It is added that “The use of the term ‘significant effect’ when only reference to ‘statistical significant difference’ is appropriate, is a direct consequence of the fallacy”. It constitutes an implicit way to operationalize its assessment. However, the reviewer’s point of view when demanding an explicit norm for us to appraise this condition is appropriate. Consequently the last part of the text cited has been changed to: The use of the term "significant effect" when only a reference to "statistical significant difference" is appropriate, is a direct expression of the fallacy and, as such, constitutes a way to operative detection of the problem in a specific paper.

**New Comment:** Now it is clearer how the presence of significance fallacy was operationalized. However, this is mentioned in the introduction of the manuscript. The lack of a clear operational definition for significance fallacy in the methods section is likely to confuse the reader. Please, provide this information in the methods section.

**Present Response:** Correct. It is a good suggestion and, as such, the text has been relocated and can be found in the proper place now (6th paragraph of the Methods section).

**COMMENT 3**

**Original comment:** Who did the data abstraction? What were their qualifications? Was it done in duplicate? How was the reliability of the data abstraction process assessed?

**Original Response:** Data collection was performed by four observers. Three of them have the qualifications of graduates in mathematics (statistics and operative research) and training in epidemiology, the fourth one is a medical epidemiologist and specialist in preventive medicine and public health. As described in the paragraph, it was conducted a prior training and a pilot study to assess reliability. Every paper was allocated to two reviewers; any disagreements were discussed and a third reviewer was consulted to break the tie when necessary. Agreement indices were not computed in the collection process after the pilot study. This information has been now included in one of the final paragraphs of Material and Methods.

**New Comment:** If the revision of all articles was done in duplicate, the best measurement of reliability of the study is the kappa values for each item evaluated in the whole sample. Would be possible for the authors to compute this data and provide them in the results section?

**Present Response:** Please, see COMMENT 5, which is closely related with the present one.

**COMMENT 4**

**Original comment:** It is not clear from the methods section if descriptive and theoretical papers were excluded from the analysis. Please, provide specific inclusion and exclusion criteria. It would be very helpful to add a table or a flow chart like the one recommended by the CONSORT statement showing the selection/ sampling process of all articles, number excluded with the reasons for exclusion and number of papers analyzed.

**Original Response:** We basically agree in making some modifications. In the Methods section it is stated that 49 items were eliminated for not being truly original papers, despite being framed in that section of the magazine (not all journals operate with the same sections or similarly classified items). The remaining exclusions are described at the beginning of the Results section, but now a comment was included in Methods to better clarify the exclusion of 169 items for being merely descriptive or theoretical and methodological. On the other hand, we believe that is not appropriate here the suggested CONSORT Flow Chart (http://www.consort-statement.org/index.aspx?o=1077) since it is a flow chart, proper to describe the selection and
randomization process corresponding to clinical trials (Consolidated Standards of Reporting Trials).

**New Comment:** The use of a flow chart to display the process of selection of manuscripts, indicating the number and reasons for exclusions and number finally used in the analysis is a very useful way to present the data because it makes it very clear for the reader. The CONSORT-like strategy is useful for any review as recommended by QUOROM, MOOSE, etc. I suggest the authors to add a chart like this in the manuscript.

**Present Response:** We should confess that we have no idea of what MOOSE is (not tracks of it can be easily found in Google, for instance). The QUOROM recommendation, however, is well known. In spite of the fact that in QUOROM the use of flow charts has been suggested to display the progress through the stages of a metaanalysis, we think that we can include this kind of resource to help the reader understand the sampling scheme. See the Figure 1, now inserted in the Methods section.

**COMMENT 5**

**Original comment:** At the end of the methods section, you mention a pilot study of your data abstraction. It would be very informative if you provide information on the number of papers evaluated in the pilot study. Please, provide also the kappa values for each of the items you evaluated in duplicate. In the description, it is not clear what the kappa value of 0.78 corresponds to. Please, also provide CIs for the kappa values.

**Original Response:** Both data (number of papers in the pilot study and 95%CI for kappa) were inserted. Unfortunately, the original database of the pilot study to compute the kappa values for each of the items we evaluated in duplicate is not in our hands at the present moment. If it were strictly necessary, please, let us know. We think, however, that another table with this information is rather superfluous.

**New Comment:** A kappa value of 0.78 is reported, however, it is not stated to which item this value corresponds. Therefore, it is not possible to interpret the value of this information. It is not necessary to add a table, but you should provide at least the interval of kappa values obtained either for the pilot study sample or (better) for the whole sample.

**Present Response:** The reviewer is absolutely right. We missed to mention the item for which this kappa value was computed. We have written the following text (partially present before in the Methods section) in the Results section, as requested:

The kappa coefficient for measuring agreement between observers concerning the presence of the “significance fallacy” was 0.78 (CI95%: 0.62 to 0.93), which is considered acceptable in the scale of Landis and Koch [39].

Note that the other items (“reference to numerical results or statistical significance in conclusions” and “use of NHST and confidence intervals”) do not offer doubts due to their absolutely objective nature. Note also that this kappa value corresponds to the final data (not for the pilot study), as stated in our Original Response to COMMENT 3.

**COMMENT 6**

**Original comment:** Regarding the English language, the paper needs thorough editing using shorter sentences and the active voice. Please, also correct some outstanding grammatical errors.

**Original Response:** We have done our best in this direction. As far as we are able to ponder this issue, no evident grammatical errors are present.
**New Comment:** Unfortunately, there are still language problems such as inconsistent use of the tenses and some outstanding grammatical errors. Please, revise the manuscript for language. The 3rd paragraph of the Background section reads: “if the null is rejected” and it probably should read “if the null hypothesis is rejected...”. Please, check this and correct it.

**Present Response:** In fact, a deeper reading of the text has let us detect some pitfalls. We hope no “outstanding grammatical errors” can be found now. Please, if they are still present, we would be very grateful if you could mention some examples. Concerning the expression “the null is rejected”, it has been changed to “the null hypothesis is rejected”. However, it is interesting to note that “the null is rejected” appears around 140,000 times in Google. It does not prove that it is a good choice, but this fact shows that it is widely used indeed.