Author’s response to reviews

Title: What differences are detected by superiority trials or ruled out by noninferiority trials? A cross-sectional study on a random sample of two-hundred two-arms parallel group randomized clinical trials.

Authors:

Angèle Gayet-Ageron (angele.gayet-ageron@hcuge.ch)
Thomas Agoritsas (thomas.agoritsas@hcuge.ch)
Christophe Combescure (christophe.combescure@hcuge.ch)
Krisztina Bagamery (Krisztina.Bagamery@finddiagnostics.org)
Delphine S Courvoisier (delphine.courvoisier@hcuge.ch)
Thomas V Perneger (thomas.perneger@hcuge.ch)

Version: 3 Date: 27 September 2010

Author’s response to reviews:

Geneva, on 27 September 2010

Dear Dr Melissa Norton,

We would like to submit a revised version of our original article entitled “What differences are detected by superiority trials or ruled out by noninferiority trials? A cross-sectional study on a random sample of two-hundred two-arms parallel group randomized clinical trials”.

We took into account the suggestions from the two reviewers and we provided in the following pages a specific answer for each referees comments.

We thank you for having reviewed and considered our manuscript for a possible publication in BMC Medical research Methodology.

Yours sincerely

Angèle GAYET-AGERON et al.

Answers to reviewer’s questions and to the editorial board

Reviewer 1: Vance Berger

• First, in the Background Section, the difference to be detected does not by itself determine the sample size, but it does partially do so. This should be clarified. We agree. We have clarified and modified the sentence in consequence (page 3, line 3).

• Same paragraph, I would replace "meaningful" with "realistic". This is not mere semantics, and hopefully the distinction will be clear.
We wanted in fact to stress the importance of choosing a difference that is relevant in clinical or public health terms, and not a difference that appears realistic or achievable. We would like to keep the wording as is but we have clarified by adding “clinically meaningful”.

Next paragraph, it is not quite true that planning a trial needs a noninferiority margin as input. I wrote a paper with Valerie Durkalski that offers an alternative to this usual paradigm. Thank you for pointing out this integrated approach to superiority and non-inferiority; we have added the citation in the references (page 3, end para. 2, reference 4). However, most published trials do not use this approach, but rather the more conventional definition of noninferiority. We prefer to perform our analysis using the conventional definition.

Also, the point about one being smaller than the other is not true in any absolute sense, but rather only in the sense of all other things being equal. This should be clarified. We have clarified this sentence as suggested, thank you (page 3, line 14, and page 11, line 8).

But the manuscript suffers from another problem, and that is perhaps best expressed as "Who cares?". What decision is informed by knowing the margins currently used in practice? This seems to me an exercise without a purpose, or an answer without a question. Pages 9 and 11 strike me as the real heart of the manuscript, and the rest should probably be built around this. In fairness, I understand why the authors did what they did, and I once did the same thing. There is this perception, right or wrong, that journals want to see a study. So they did a study. A useless one at that, but still a study. I once did the same thing only to get the journal to look at the real point, which had nothing to do with the study itself. Likewise here, the study aspect of this strikes me as nothing more than a distraction. I cannot speak for the journal, of course, but I would think that the manuscript would become stronger with more discussion and less nonsense about which other studies use which noninferiority margins.

Whether this study was worth doing is obviously a matter of opinion. We see value in doing it, and believe that useful information was derived in the process. For starters, the absence of consensus on how to choose a meaningful difference is a real problem for researchers. We are often confronted to this difficulty at our Clinical Research Center where clinical researchers come to be helped with sample size estimation – their hypotheses are often weak and poorly argued, and many are more concerned about the feasibility of the study than about the interpretability of the results. We wanted to see if a similar variability existed for differences to be detected and ruled out used by researchers in published trials (the answer is yes). We also wanted to verify if on average the difference to be detected in superiority trials was larger than the difference to be ruled out in noninferiority trials (again, the answer is yes). Finally we wanted to identify factors that might explain the variability (we have identified a few). To sum up, we see a useful contribution in this work, a foundation upon which
further research or consensus building can be based.

Reviewer 2: Yuliya Lokhnygina

- This paper presents an interesting exploration of the variability in choices of differences to be detected in randomized superiority and noninferiority clinical trials.

Thank you.

- Major Compulsory Revisions

1) For the majority of the paper, the authors have chosen to treat standardized difference in means and standardized difference in proportions as the same standardized measure. However, there is no discussion about why this is a valid approach and whether standardized difference in proportions is an accepted measure of the effect size. For example, Cohen considers a difference in arcsine-transformed proportions for the effect size, while many trials with binary outcomes consider treatment differences expressed by odds ratios or by risk ratios. Were such trials included in this analysis (with effect size recalculated)?

Why did the authors chose this specific measure of the effect size for the binary outcomes?

We have added the reference of Burnand B. et al. (J Clin Epidemiol, 1990) for the use of standardized increment and have clarified this in the Methods’ section (page 6, under “Outcomes”, line 16).

Regarding studies which expressed treatment differences by odds ratio or by risk ratios, we have explained in the Methods’ section that proportions were “recalculated using the formulae for sample-size calculation adapted for a #2 test or Fisher’s exact test or those adapted for bioequivalence trials” (reference cited: Machin D, Campbell M, Fayers P, Pinol A: Sample size tables for clinical studies. 2nd edn. Oxford: Blackwell Science Ltd; 1997).

We chose to use standardized increment in order to allow comparisons with continuous outcomes. A similar approach was used by Lange S and Freitag G (Biom J, 2005) in their systematic review on noninferiority and equivalence studies. We have added a specific sentence in the limitations (Discussion, page 12) and we have cited this paper (reference no 15).

2) It is not clear why the p-values in Table 3 are presented separately by subgroup. It seems that an appropriate analysis approach would be to test both main effect of each factor and interaction between the factor and superiority/noninferiority type of the trial. If the interaction is not significant, a single p-value for the main effect of the factor should be presented.

We chose to make subgroups analysis in order to provide descriptive data before identifying the predictors of differences. We think the readers would be interested to have descriptive data on differences used in each type of trial and by study determinants. A specific sentence was added in the Methods’ section (page 7 para. 3 line 7). Nonetheless the method proposed by the reviewer is interesting and we have added in Table 3 a column presenting a p-value for each factor.
when the interaction term was not significant between the factor and the type of trial (P>0.10). We found a significant interaction between the type of trial and the sample size, which is why we did not provide a p-value for this factor.

3) Please clarify the last analysis of variance, results of which are described in the last paragraph of section "Trial characteristics associated with clinical differences used to estimate sample size". Was this a multivariable regression model with standardized difference as an outcome and predictors selected from factors listed in Table 3? In that case, what model selection method was used? Thank you. Yes this is right; we presented here a multivariate analysis of variance model with the determinants significantly associated with the standardized difference found in Table 3, except sample size and patients’ recruitment which were more a consequence of the specific difference used than a predictor of the difference. This was explained in the Methods’ section (page 8, line 2).

What was the R-squared coefficient for the model?
The R-square was 0.31 (see page 10 last lines)

For each factor, does "adjusted" in "adjusted mean" refer to adjustment for other predictors in the model?
Yes this is right. It is described in the Methods’ section of the article under “Statistical analysis” (page 8, line 4).

4) Have the authors considered the phase of the study (e.g. 2a, 2b, 3) as a potential predictor of the standardized difference? Typically, phase 3 trials aim to detect smaller treatment differences than phase 2 studies.
We did not record this information because an explicit statement of the phase of testing is often absent from the published articles. However, most published clinical trials seem to be phase 3 or comparison of therapeutic strategies.

- Minor Essential Revisions
1) In Table 1, it appears that for the analyses of categorical variables sometimes chi-square test and sometimes Fisher’s exact test were used. How was the statistical test selected in each case? Note that only chi-square test is mentioned in the "Statistical Analysis" section.
We agree and we have added in the Methods’ section that Chi-square and Fisher’s exact tests were used to compare categorical variables (page 7, line 19).

- Discretionary Revisions
1) It would be interesting to know how many (if any) of the trials considered in this study explained the choice of the treatment difference used to calculate the sample size of the trial.
This is an interesting question, thank you, and we have added this information in the results (page 9, line 9) and in Table 1. In fact among the 200 trial, only 49%
of the studies provided a justification for the choice of the treatment difference, either by a reference to published studies or by results from pilot studies. The justification was significantly higher in superiority compared to noninferiority trials (59% versus 39%, P=0.02).

Editorial board:
Please also do the following:
(1) Adhere to the PRISMA guidelines for reporting systematic reviews found at http://www.prisma-statement.org/

The design of our study was a survey not a systematic review. We based our study on a random selection of 200 published clinical trials. That is why we did not adhere to the PRISMA guidelines as suggested.

Please also highlight (with 'tracked changes'/coloured/underlines/highlighted text) all changes made when revising the manuscript to make it easier for the Editors to give you a prompt decision on your manuscript.
It was done as suggested.

Ensure that your revised manuscript conforms to the journal style (http://www.biomedcentral.com/info/ifora/medicine_journals ). It is important that your files are correctly formatted.
We have revised our manuscript following the journal style and we have highlighted the changes made.