Author’s response to reviews

Title: Spontaneous improvement in randomised clinical trials: meta-analysis of three-armed trials comparing no-treatment, placebo and active intervention

Authors:

Lasse T Krogsbøll (lassetk@hotmail.com)
Asbjørn Hróbjartsson (ah@cochrane.dk)
Peter C Gøtzsche (pcg@cochrane.dk)

Version: 2 Date: 22 September 2008

Author’s response to reviews:

Our replies are in italics

Reviewer’s report
Title: Spontaneous improvement in randomised clinical trials: meta-analysis of three-armed trials comparing no-treatment, placebo and active intervention
Version: 1 Date: 13 August 2008
Reviewer: David J Torgerson
Reviewer’s report:
I found this paper very interesting and the authors make the point, which is always worth making, that there is spontaneous improvement in many, if not most, clinical areas that are completely unrelated to the treatment effects.

MAJOR
My main criticism, which is a major one, but easily dealt with: is that the authors make no mention of regression to the mean. When patients appear to get better there are three, not two, alternative explanations in addition to a true treatment effect. Placebo and temporal changes as the authors correctly state but also regression effects.

We agree and had originally discussed regression to the mean (RTM) in our manuscript, but decided that it made the paper less complicated to incorporate this into the concept "spontaneous improvement". We have now made this clear and say in the Introduction: "For the purpose of this paper, we regarded regression to the mean effects as being part of the spontaneous improvement. Regression to the mean occurs, for example, when a patient can only be included in a trial if the symptoms are worse than some threshold value; for purely statistical reasons, the value will then likely be lower at a later time (1,2)".
I am surprised that they do not mention it at all as some of their findings are almost certainly due to this phenomenon. For example, on page 9 the say – “It seems likely that spontaneous improvement is more important in trials that include patients with high symptom scores and that do not implement a placebo run-in period.” NO this is regression to the mean – high symptom scores are most likely to have a higher error term and regress more than those near the mean

We agree, but only partly. What we had in mind here was actually not RTM but the difference between trials that study people with little symptoms and those that include people who are more sick. We have now made it clear that both considerations are relevant, as we have added: "particularly as the regression to the mean is likely to be more pronounced in such settings".

also having a run-in deals with regression to the mean as you are having repeated measures taken.

We agree, but an improvement could also be due to the fact that patients tend to see a doctor when symptoms are worst (truly worst, i.e. not a statistical phenomenon, as RTM is), in which case there would be less improvement if there was a placebo run-in. We have clarified this.

The reason you didn’t see an effect on hypertension trials is that physicians, unless they are incompetent, normally take 3 BP measurements and take the average and only diagnosis hypertension if this is above a certain threshold. This mainly deals with regression effects so if you then enrol these people into a trial those with high values to due to test error have been removed.

Thanks for the comment; we have made this clear; 3 of the 4 studies had run-ins..

Psychological tests or any quality of life tests have high error components and are classically subject to regression effects.

There are lots of references to this phenomenon here are some of mine but there are lots of others Stephen Senn as written quite a bit on this:

Torgerson DJ and Torgerson CJ Designing randomised trials in health, education and the social sciences Palgrave macmillan, Basingstoke, 2008 - Chapter 2

Morton and Torgerson BMJ 2003,326:1083-4


So what I recommend is that you include text about regression to the mean and perhaps discuss those areas where is less likely to occur due to either tests with
little error or sequential testing (e.g., hypertension) and interpret your findings in this way.

We have done as suggested.

Unfortunately, I don’t think it is possible to disentangle spontaneous improvement from regression to the mean – you can see where it is less likely but I don’t think you can ever disentangle this from spontaneous improvement.

We agree, which is why we decided to to incorporate RTM into the concept “spontaneous improvement”.

MINOR
Other less important issues is the type of tests – some will have a ceiling or floor problem. Therefore, those near the top or bottom of the test scores can only go one way, which assuming there is no true change whatsoever, just error you will tend to see a change downwards if it is a ceiling problem because the error can only go to one side – this exacerbates the regression problem (although it is a little different as even if you have a test with no ceiling or floor you will still get regression).

This is a very minor point, as the changes we observed were quite large, and as scorings were not close to ceilings or bottoms.

David Torgerson

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, and I have assessed the statistics in my report.
Declaration of competing interests:
I have no competing interests

Reviewer's report
Title: Spontaneous improvement in randomised clinical trials: meta-analysis of three-armed trials comparing no-treatment, placebo and active intervention
Version: 1 Date: 31 August 2008
Reviewer: Jesse Berlin
Reviewer's report:
1. Abstract: It’s not immediately obvious how the percentage contributions were calculated. I understand that you may not have room in the abstract to show calculations, but the same comment applies to the Results section of the paper.
At a minimum, you need to make it very clear in the abstract that the SMD given in the Results section of the abstract is for pre-post change?

We had already written this, in two places, but have now made it more clear.

2. In the Results section of the body of the paper, you need to show calculations of percent contribution. No treatment SMD = -0.44; Placebo = -0.54; and active = -1.27. You say the contribution of spontaneous improvement is 35%. Is that .44 / 1.27 = 0.35 (after rounding)? This is just a matter of clarification.

Yes; we have explained this now.

You might also provide a footnote for Table 2, giving the calculations? In particular, how did you define the contribution of placebo? Is that subtracting out the effect of spontaneous improvement?

Yes, this appears from the Figure we provided, and now also from the calculated examples, see your suggestion above.

3. Page 4: What was the basis for the selection of outcome measure?
Specifically, what do you mean by the outcome that you “found most relevant?” How was relevance defined and determined? Was there a consensus process?

Yes, and it was not difficult to agree on what was most relevant for patients. We have now explained this.

4. Page 5: With respect to the use of pre-post data, let’s assume we just accept the idea of ignoring the pairing in calculating effect sizes. First, you should probably acknowledge that ignoring the pairing implies that we are willing to assume that the effect of the ignored correlations between pre- and post-intervention measurements is the same in all situations.

We have added this explanation.

That may not be a realistic assumption, but I agree with the idea that various approaches to imputing missing correlations, correcting variance estimates, etc., would make little difference to the conclusions.

We agree, and it would also be problematic to try to guess the magnitude of missing correlations.

5. Correlations aside, your approach doesn’t seem to take the randomization into account, unless I’m misinterpreting your calculations. Wouldn’t it be possible to get a within-study measure of the relative contributions? These are all 3-arm
studies, so the effect of spontaneous improvement, within a study, could be defined as “change from baseline on no treatment / change from baseline on active treatment.” I think I’m advocating using a very similar approach to what you did, but if I’m interpreting correctly, your current calculations are based on the “ratio of the means” rather than the “mean of the ratios.” I hope that’s clear. The point is that you may be introducing potential confounding, if “study” is related to both outcome and treatment assignment? I don’t believe this will make a substantive difference in your conclusions, but it might, for example, if the allocation ratios are substantially different from 1:1 in some of the studies.

We thought carefully about these issues but could not come up with any better solution than the approach we have described. We have now added: "We considered other approaches and also consulted a PhD in statistics, but neither he, nor we could devise a better method than the one we used, partly because other methods would need to rely on unverifiable assumptions". We agree that it wouldn't make any substantive difference in our conclusions to use other methods.

6. Page 5: The difference in pre and post sample sizes could reflect bias in the effect sizes if, for example, the subjects who dropped out did so because of reduced efficacy or increased risk of adverse effects. It’s true that studies with loss to follow-up would be down-weighted, (and I know you’re aware of this potential bias), but you might want to note this issue as a potential limitation.

We do not think that this could be an important bias or limitation of our study; the drop-out rate were as expected for such trials.

7. General: Is the analysis here really a multivariate problem? You’ve focused separately on acute versus chronic conditions and observer-reported versus patient-reported outcomes. Could the nature of intervention (psychological versus physical versus drug) matter? I realize it wouldn’t be possible to separate out these effects, because not all types of interventions occur with all clinical conditions. I’m not sure you have enough studies to do real multivariable modeling, but please give this some thought.

We agree that there are too few trials to allow such explorations in the data.

Minor Essential
8. Introduction, paragraph 2: Needs a bit more elaboration on spontaneous improvement? Just add something that completes the thought. Patients seek treatment and enroll in trials when Sx are at worst, …, so natural variability
means they will be less severe over time (assuming the natural variability is “reversible”, not just increasing deterioration).

We have added more text on this, and have also described regression to the mean.

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, and I have assessed the statistics in my report.
Declaration of competing interests:
I am a full time employee of Johnson & Johnson Pharmaceutical Research and Development. I know of no specific conflict related to this methodologic paper.

Reviewer's report
Title: Spontaneous improvement in randomised clinical trials: meta-analysis of three-armed trials comparing no-treatment, placebo and active intervention
Version: 1 Date: 31 August 2008
Reviewer: Joseph Lau
Reviewer's report:

The authors seek to quantify the relative contributions of spontaneous improvement, effect of placebo, and effect of active interventions in change from baseline in randomized trials.

The title needs to be changed to reflect the types of trials analyzed. Even though the authors stated the limitations in the discussions, these are mostly trials with subjective measurement endpoints and clearly cannot be generalized to other settings. The nature of the treatments in these trials should be made clearer so that readers can judge what settings are the findings of this study applicable.

The title would become too long if we were to describe the type of interventions and outcomes.

It is unclear to me what is exactly the question addressed by the authors in this paper. They appear to ask the question of what is the spontaneous improvement in each arm of a 3-arm trial (no treatment, placebo, and active treatment) compared with their respective baseline.

We did not. We say at the end of the Introduction: We have not found any previous studies of the three main factors affecting the clinical course of patients included in randomised clinical trials: spontaneous improvement, effect of placebos and effect of active interventions. We aimed at quantifying the relative
contribution of these factors to change from baseline in randomised trials."

but they also performed indirect comparisons of the results among the 3 arms. While the question of change from baseline within each arm is valid, it has little usefulness in interpreting trial results.

We disagree and had described why. The improvement a clinician notes when treating patients is composed of all three factors. It is therefore relevant to study the contribution of each of them in order to understand why it is so common that clinicians think the treatments they use are much more effective than they really are.

The effect of a treatment is typically assessed as the difference of the mean difference between comparison groups. Thus whatever spontaneous change that occurs in the active treatment group is also cancelled out by change in the no treatment or placebo arm. There are methodological/statistical problems when the authors used the individual arms results to perform the indirect comparisons, which the authors acknowledged in the discussion. I would suggest that the authors create a simple diagram depicting the 3-arms of a trial and what question(s) they are asking.

The reviewer has not understood what the aim of our study was. We believe we have described our rationale quite clearly.

In this study, the authors updated a previous Cochrane review and combined 3 arms trials from their previous Cochrane review. The authors stated that it was not straightforward to do the analyses and ended up combining the 3 treatment arms separately comparing the post-treatment values with the values at baseline. This was not a meta-analysis of a specific treatment effect but an attempt at deriving an estimate of a study design effect.

Our aim was not to study design effects.

The method of combing individual arms across studies ignores the correlations of the treatment arms within the study and the principle of meta-analysis of combining the effect sizes (differences between arms and not just the measurements of the individual arms). Combining just the results of the same treatment arms across studies, rather than the difference of effect size, ignores potentially differences in the baseline values that may exist, and could be influenced by the well known Simpson’s paradox problem. The authors stated that they find it reassuring that the overall effect of placebo was -0.28, which was similar to the previous estimate of -0.24. Hover, having a number close to another does not necessarily mean that the answer is correct. In consultation with a statistician with in depth meta-analysis expertise, I was advised that this problem could be handled by analysis of variance or
regression methods.

We have now stated in our paper that we also consulted a statistician and that he could not come up with a solution to our problem that was better than the one we had worked out ourselves. Jesse Berlin, who also peer reviewed our paper, and who is a statistician, has accepted our methods, see above. We are not convinced that analysis of variance or regression methods would be the solution, as there would need to be assumptions that cannot be verified.

While I appreciate that in this data set “no treatment” has the smallest magnitude of spontaneous improvement, and placebo has larger effect, and active treatment the largest effect, I am not certain whether the actual estimates from the meta-analyses coming up with estimated has much generalizability beyond the data analyzed. The data set is from a relative small sample of 43 trials across 9 topics which were quite heterogeneous. Thus, sampling from another data set might easily yield different results.

The reviewer seems to overlook that our study is unique. There are no other data sets. We have assembled anything there is in the literature, and, to our knowledge, no-one has tried to quantify the contribution of spontaneous improvement and placebo separately to the observed treatment effect in an actively treated group before. We agree about the limitations, but it was the best we could do. We have now added, after our estimates of the contributions of the three factors: but with some uncertainty, as indicated by the confidence intervals for the individual SMDs and the dependency on the studied clinical conditions. The authors reports that “thus, on average, the relative contributions . . . were 20% and 25%, respectively.” It is unclear to me how these relative contributions were calculated. If I take -0.19 change for spontaneous improvement relative to -1.03 for active treatment, the relative contribution might be ~20%. For -0.47 change for placebo, shouldn’t the relative contribution be about ~50%?

We have now explained the calculations. To arrive at the placebo effect, one needs to subtract the spontaneous improvement, and the placebo effect is therefore 25%, not 50%.

Did the authors incorporated methodological quality of the studies in their analysis as a possible explanation of heterogeneity?

No.

There are many more (58) active treatment arms than no treatment or placebo arms (43 trials each). How were the extra treatment arms handled in the analysis?

We have already described this.
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:
I declare that I have no competing interests.