Author's response to reviews

Title: How Confidence Intervals become Confusion Intervals

Authors:

   James McCormack (james.mccormack@ubc.ca)
   Ben Vandermeer (bv1@ualberta.ca)
   Gary M Allan (michael.allan@ualberta.ca)

Version: 2 Date: 6 October 2013

Author's response to reviews: see over
October 5, 2013

Dear BMC Medical Research Methodology Editorial Board,

We are submitting our revised paper “How Confidence Intervals become Confusion Intervals” to the BMC Medical Research Methodology. We have responded to the reviewers and revised the content as required. The response to the reviewers is detailed below.

We certify that all authors made significant contribution to the design, writing and editing. G. Michael Allan is the corresponding author (contact information above). We also certify that the authors have no conflict of interest in this submission and that we have not published or submitted any related papers from the same study.

We look forward to hearing back from you.

Best Regards,

G. Michael Allan MD,
James McCormack Pharm D,
Ben Vandermeer MSc
Response to reviewers and authors revisions

Referee 1

1. What I missed in the paper is that there is no attention for the role of editors / reviewers in the way results are written down. Or even rewriting by journal editorial teams (as is common in the general medicine papers). Obviously, some journals have taken a more nuanced point of view more recently and that should be mentioned in the paper, I guess. Another issue that needs some attention is the role of regulators (e.g., FDA, EMA) in this, and how they stick to the magical statistical significance level of 0.05.

*Added comments about editors and the FDA into the conclusion*

Referee 2

Major comments

1. For the Connoly study in example 2, only the results are mentioned and not the conclusion of the authors, while for the other studies in this example and the other examples the conclusion rather than the results are stated.

*Added in a statement about this*

2. In the pragmatic interpretation that the authors give, the focus is more on presenting absolute risks instead of relative risks, rather than on focusing on the CI/p-value problem. This is also an important issue but is not discussed anywhere else in the paper, and therefore it comes out of the blue in the 'pragmatic interpretations'.

*As this is a pragmatic interpretation we felt the need to describe how we would interpret the results and to do that completely one needs to include absolute numbers. No changes made.*

3. In the discussion, the authors suggest that there are researcher that deliberately look for studies/meta-analyses where no difference is found and that they repeat the analysis in a slightly different way to get a significant result and therefore publish it in a high impact journal. I would say that is quite a conviction for science as a whole, and I was wondering whether the authors have evidence that these thing really happen.

*Agree – we have removed this statement*

Minor comments

1. The structure of the manuscript is a bit strange. First two paragraphs without heading and then a few paragraphs with 'Background' as the heading. What do the first two paragraphs represent?

*We have labeled the first 2 paragraphs as an introduction*
2. Page 4, 'A statistical test helps...'. Sentence is not clear.

Reworded

3. Page 4, 'A randomized controlled trial ... to test a hypothesis'. An RCT is not the best design for any hypothesis. It is usually for studying the effect of an intervention. Please revise the sentence.

Reworded

4. Page 4, 'A CI...'. Sentence is not clear.

5. Page 5, 'Statins have become...'. Please add a reference for this statement.

Reworded

6. Example 1, please present the five studies in the same order in the text and the figure.

Rearranged

7. Page 7, 'The Granger paper... that cross the magical line of 1.0..' should be 'do not cross the magical line of 1.0.

Reworded

8. It is more logical to put the pragmatic interpretation of the included studies directly after the description of the studies.

We felt that a synopsis of all the conclusions works best at the end so no changes made

Referee 3
Thank you for inviting me to review the manuscript “How Confidence Intervals become Confusion Intervals” by James McCormack, Ben Vandermeer and Gary M Allan. In this paper, the authors present several examples from the medical literature to support their argument that confidence intervals (CI’s) are often interpreted by authors of medical articles in a way that causes confusion among “consumers” of the results (clinicians, the media, the general public). CI’s, they state, should not be used strictly for hypothesis testing at the 0.05 significance level, to “make dichotomous conclusions”. They conclude with several recommendations: authors should refrain from making black-and-white conclusions about their results based on thresholds - especially when one end of a CI is near said threshold; journals should challenge authors who make such statements; and readers should examine the actual results and not just read the authors’ conclusions. The manuscript is well-written and the examples presented clearly. The “pragmatic interpretations” of the results are an excellent example of how researchers can aid clinicians in interpreting results of medical studies. The authors’ argument, with these or similar examples, should probably be the basis of a mandatory lecture in every beginning course on epidemiology and/or biostatistics.
Having said that, I do not believe that the argument of McCormack et al. is particularly novel. Upon reading the manuscript, my first thought was, “don’t we already know this?” (Though given the examples presented in the manuscript, the answer is evidently: no, not everyone knows this.)

What I miss in this manuscript is a sense of historical perspective. The authors approach the subject of confidence intervals as if the statistical testing vs. confidence interval debate had not been raging in medical literature since the 1970’s. Starting with Wulff in 1973 and Rothman in 1978, attention has been paid in the medical literature to the importance of giving point estimates and CI’s instead of simply concluding that a test is “significant” or “not significant”. Gardner and Altman (Br Med J 1986 March 15; 292(6522): 746–750) deplore the practice of dichotomizing results, and state in their discussion “Confidence intervals, if appropriate to the type of study, should be used for major findings in both the main text of a paper and its abstract.” Others followed suit in the 1980’s. In the 1980’s and 1990’s many medical and epidemiologic journals adopted the position that CI’s were to be preferred over p-values. The current guidelines of the ICMJE (http://www.icmje.org/manuscript_a.html) state: “When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as P values, which fail to convey important information about effect size and precision of estimates.” More recently, Cummings and Koepsell (Arch Pediatr Adolesc Med 2010;164(2):193-196) weighed in on the subject, and state as a precaution in their recommendation to use CI’s: “First, confidence intervals should not be used to simply judge estimates as statistically significant or not.”

The failure to sketch the forty-year history of the recommendations in the medical literature for presenting CI’s detracts from the manuscript’s potential for furthering the discussion on presenting and interpreting results from medical studies. By placing their examples in the larger context, their argument would be more persuasive: this is exactly why those recommendations on reporting point estimates and CI’s have been made. Authors, journals, and “consumers” of medical study results need to be alert in the reporting and interpretation of CI’s.

Below are my recommendations for revisions.

Discretionary Revisions
1. Background section, second paragraph, last sentence: “Once we believe a particular treatment has an effect, we also need an estimate of the size of that effect.” The rest of the paragraph addresses the fact that failing to reject a null hypothesis does not mean there is no effect. It seems strange to then recommend that estimation of effect size should only be done when we believe there is an effect.

\textit{Agree - removed}

2. Example 1, second paragraph: the meta-analyses are discussed in a different order than in Figure 1. In the paragraph, Mills et al. is presented before Thavendiranathan et al., but in the figure these two studies are reversed.
Fixed

3. Example 2, fifth paragraph, first sentence: “The Granger paper illustrates the importance that authors attach to results that cross that magical line....” From the rest of the paragraph, it seems that the sentence should read “to results that do not cross that magical line”.

Fixed

Minor Essential Revisions

4. Background section, second paragraph, last sentence: “A CI, while it has other interpretations, can be thought of as a range of that we are....” I believe a word or phrase is missing between “of” and “that” (perhaps “values”?).

Corrected

5. Example 1, first paragraph, penultimate sentence: “no single studies have been powered...” should read: “no single study has been powered”.

Corrected

6. Discussion, fourth paragraph, third sentence: “Investigators simply identify a study or meta-analysis where the CI approaches no difference, and repeat the analysis with subtle differences in trial/patient inclusion or statistically testing....” I believe the last two words should be “statistical testing”.

Removed this sentence

7. Summary, first sentence: “Dogmatic adherence to statistical significant thresholds...” should read “statistical significance thresholds”.

Corrected

8. Figure legend, Figure 2: “with 3 novel oral anticoagulants” should read “with 2 novel oral anticoagulants”

Corrected

9. The labels at the bottom of the figures read “igure 1”, “igre 2” and “igr 3”, but presumably this was a problem in the conversion to a .PDF and not a problem in the original figures.

It is not clear how this occurred, as the figures don’t have this writing on them. We will check on final submitted version.

Major Compulsory Revisions
10. My main criticism is that the arguments and recommendations presented in the manuscript are not new, see general comments above.

Agree that the arguments are not new BUT because of the examples mentioned that appear in high level journals it appears that this problem still needs to be highlighted.

11. In addition, the literature on presenting and interpreting confidence intervals is ignored, again see above.

Added comment and references

12. I have objections to Example 1, or at least to the presentation and “pragmatic interpretation” thereof. The authors clearly state that the meta-analyses are not statistically independent of one another: “Although the different meta-analyses included some different studies, overall the investigators used similar data and, not surprisingly, found similar results.” (Example 1, third paragraph, third sentence.) The approach used to present the three examples is similar to a meta-analysis forest plot (without the combined effect from the meta-analysis), which is appropriate for Examples 2 and 3. The “pragmatic interpretations” of Examples 2 and 3 are also appropriate, since the studies included in the figures were (presumably) independent of one another. My objection to Example 1 is that the interpretation of the graph is based on an assumption that the studies all independently measured the same effects, and that we can then use the “average” of the 5 studies to infer reduced risk in mortality among statin users. Since the meta-analyses often included the same studies, it is not possible to see them as independent estimates of the same effect. I do agree with the conclusion that the “difference” in conclusions from these five meta-analyses is an artifact of dichotomizing the resulting CI’s; in this sense, it is a good example of the point the authors wish to make. However, I feel that presenting the CI’s in a “forest plot” type of graph is misleading, and that caution should be used in interpreting how much evidence there is for statins in Example 1.

The figures are in no way meant to be used as a forest plot nor are they meant to be used to average the results. The figures are solely provided as a visual aid to how similar the point estimates and the confidence intervals are for these meta-analyses.