False statistically significant findings in cumulative meta-analyses and the ability of Trial Sequential Analysis (TSA) to identify them.
Protocol

Georgina Imberger[1]
Kristian Thorlund[2,1]
Christian Gluud[1]
Jørn Wetterslev[1]

1. Copenhagen Trial Unit, Center for Clinical Intervention Research, Copenhagen University Hospital, Copenhagen, Denmark.
2. Department of Clinical Epidemiology and Biostatistics, McMaster University, Ontario, Canada.

Purpose

To quantify how many false statistically significant findings would have occurred in real-life cumulative meta-analyses if these meta-analyses had been performed sequentially over time.
To assess the ability of Trial Sequential Analysis (TSA) to identify false statistically significant findings in real-life cumulative meta-analyses that are performed sequentially over time.

Hypothesis

Conventional meta-analytic techniques cause false statistically significant findings when cumulative meta-analysis is performed sequentially over time. TSA can identify these false statistically significant findings.

Background

Conclusions of systematic reviews with cumulative meta-analyses, as with all medical research, always include a risk of error. Systematic error, or bias, can be limited by good design. Random error, or the result of the 'play of chance', cannot be totally avoided. Rather, the risk of random error must be acknowledged, calculated, and controlled. Usually, we consider it reasonable to accept a 5% risk of making a false positive error (type 1 error) and a 20% risk of making a false negative error (type 2 error). That is, when we decide that the risk of error for a given conclusion is less than these thresholds, then we accept that conclusion as robust.

When a meta-analysis is updated over time, as new trials are completed, the risk of random error is increased[1]. This increased error is analogous to the increased risk of error present when interim analyses are done in a single trial. In a single trial, it has long been accepted that adjustments are required for the multiplicity caused by repetitive testing[2]. Early stopping can be problematic and monitoring boundaries, incorporating the sample size calculation, are commonly used to control the risk of random error at the desired levels of 5% and 20%[3-5].

In meta-analyses, techniques for controlling type 1 and 2 errors at their desired level in the context of repeated updates have been used but there is no consensus within The Cochrane Collaboration, or between meta-analysts in general, of the necessity to use them[6,7]. Likewise, it is presently not explicitly recommended for a meta-analysis to include a consideration of whether the number of patients included represents an adequate sample size[7].

The assessment of a 'sample size' in meta-analysis is more complicated than it is in a single trial, and heterogeneity between trials needs to be incorporated[8]. The required information size (RIS) describes the number of patients needed in a meta-analysis to make a conclusion, based on a pre-specified control event proportion, intervention effect and risks of type 1 and type 2 errors, and also incorporating a measurement of heterogeneity in its calculation. The RIS may be a reasonable measure of the required 'sample size' for a meta-analysis[9].

TSA is a methodology which combines conventional meta-analysis techniques with thresholds for declaring significance in the context of sequential meta-analysis performed before the RIS has been reached[9-13]. The thresholds are based on methods already widely used for interim analysis in single trials, using an O'Brien Fleming alpha-spending function which varies the threshold for statistical significance, such that it is more conservative when the data is sparse and becomes progressively more lenient as the accrued information gets closer to the RIS[4].

There is empirical evidence that TSA can prevent the increased risk of false positives in repeated updates of meta-analyses[11]. This study aims to further investigate the false statistical significance rate caused by traditional cumulative sequential analysis of meta-analyses using a fixed threshold to declare statistical significance and then to test the ability of TSA to identify false statistical significance.

Methods

Selection of systematic reviews for inclusion

We will select meta-analyses from The Cochrane Database of Systematic Reviews that we consider to have demonstrated, beyond a reasonable level of doubt, that the given intervention does not cause a clinically relevant effect.

Starting with the most recently published systematic reviews on the database, we will select meta-analyses that fulfill the following criteria:

1. The data is dichotomous
2. The result of the meta-analysis is not statistically significant, having a conclusion of no relevant effect
3. A relevant RIS has been reached.

We will use the TSA software to calculate the RIS (http://www.ctu.dk/tsa/). We will consider that a meta-analysis has passed a relevant RIS if the number of patients accumulated passes a RIS using the following variables:

Boundary type – two-sided.
Type 1 error – 5%.
Power - 80%.
Incidence in the control arm – calculated as an un-weighted mean of the proportion with the outcome in the control groups of all the included trials.
Relative risk reduction – 10% RRR OR a RRR such that the Number Needed to Treat (NNT) is 100 (using the above definition of incidence in control arm).
Heterogeneity adjustment – model variance based (ie. allowing the TSA software to use a calculation of diversity[8] of the included trials to adjust the RIS.

In the case that a systematic review contains more than one meta-analysis fulfilling the above criteria, we will select the first one presented in the 'Data and Analysis' section.

We will aim to find 50 meta-analyses that surpass a RIS with a relative risk reduction of 10% and 50 meta-analyses that surpass a RIS with a relative risk that corresponds to a NNT of 100.

Retrospective cumulative meta-analyses

We will use the TSA software to conduct these analyses (http://www.ctu.dk/tsa/).

We will conduct retrospective sequential cumulative analyses on each included meta-analysis, repeating the analysis after the addition of each new trial in the order that they were published. Where more than one trial was published in the same year, we will add the trials according to the alphabetical order of the first authors.

We will use relative risk as the effect measure. We will use the Der-Simonian-Liard random-effects model to conduct the meta-analysis. In the case of zero events in one or both arms, we will make a constant adjustment of 0.001 in both arms.

First we will construct a linear boundary based on the conventional threshold of P=0.05 (two sided).

If a retrospective cumulative meta-analysis crosses the conventional boundary for significance, we will classify this is a 'false statistically significant finding'. (Using the information we have now, such a meta-analysis would have represented a false statistically significant finding had a conventional meta-analysis been done at that time.)

For the cumulative meta-analyses that do have false statistically significant findings at any point, we will assess whether TSA would have identified this finding as false. That is, we will perform TSA to check whether the Z-value would have crossed the TSA threshold for significance. We will construct a two-sided TSA boundary, using an O'Brien Fleming alpha spending function. This testing will be two-sided, using a type 1 error of 5% and 80% power. For the relative risk reduction, the incidence in control arm, and the heterogeneity correction, we will use three different models:

Model 1. Using the criteria for inclusion.

We will use the variables that we used to select the meta-analysis as one with an intervention with no effect. (Described above.)

Model 2. Using the point estimate at the time of the false statistically significant finding.

Relative risk reduction - the point estimate calculated by the conventional meta-analysis at the time of the false positive.

Incidence in the control arm - the unweighted mean of all included trials at the time of the false positive.

Heterogeneity correction - model based, allowing the TSA software to use a calculation of diversity[8] of the included trials to adjust the RIS.

Model 3. Using the border of the 95% confidence interval closest to null at the time of the false statistically significant finding.

Relative risk reduction - the border of the 95% confidence interval closest to null calculated by the conventional meta-analysis at the time of the false positive. (ie. the smallest effect size contained within teh 95% confidence interval.)

Incidence in the control arm - the unweighted mean of all included trials at the time of the false positive.

Heterogeneity correction - model based, allowing the TSA software to use a calculation of diversity[8] of the included trials to adjust the RIS.

Measurement of false statistically significant conclusion rate

We will measure the proportion of included meta-analyses that would have produced one or more false statistically significant conclusion using conventional meta-analysis. We will measure how many of these falsely significant results would have crossed the TSA threshold for significance if this analysis had have been performed at the same time. We will compare the proportion of meta-analyses that would have produced false statistically significant conclusions when using conventional meta-analysis with the proportion of meta-analyses that would have produced false statistically significant conclusions using TSA.

References

1. Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. *J Clin Epidemiol* 2009;62:825-830.
2. McPherson K. Statistics: The problem of examining accumulating data more than once. *N Engl J Med* 1974.
3. Pocock SJ. Clinical trials, a practical approach. First edition. Chichester: John Wiley & Sons; 1983.
4. DeMets D, Lan KK. Interim analysis: the alpha spending function approach. *Statistics in Medicine* 1994;12:1341-1352.
5. Bassler D., Montori V.M., Briel M., Glasziou P., Guyatt G. Early stopping of randomized clinical trials for overt efficacy is problematic. *J.Clin. Epidemiol*. 61 (3) (pp 241-246), 2008.
6. Bender R., Bunce C., Clarke M., et al. Attention should be given to multiplicity issues in systematic reviews. *J. Clin. Epidemiol*.. 2008;61(9):857-865.
7. Higgins JPT, Green S. *Cochrane Handbook for systematic reviews of interventions, version 5.0.0*. John Wiley & Sons, 2009.
8. Wetterslev J., Thorlund K., Brok J. & Gluud C.. (2009). Estimating required information size by quantifying diversity in random-effects model meta-analyses. *BMC medical research methodology*, 9, 86.
9. Brok J., Thorlund K., Gluud C. & Wetterslev J.. (2008). Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. *Journal of Clinical Epidemiology*, 61, 763-769.
10. Brok J., Thorlund K., Wetterslev J. & Gluud C.. (2009). Apparently conclusive meta-analyses may be inconclusive - Trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating

data in apparently conclusive neonatal meta-analyses. *International Journal of Epidemiology*, 38, 287-298.
11. Wetterslev J., Thorlund K., Brok J. & Gluud C.. (2008). Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. Journal of Clinical Epidemiology, 61, 64-75.
12. Thorlund K., Devereaux P.J., Wetterslev J., Guyatt G., Ioannidis J.P.A., Thabane L., et al. (2009). Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? International Journal of Epidemiology, 38, 276-286.
13. Thorlund K., Engstrøm J., Wetterslev J., Brok J., Imberger G., Gluud C. (2011). User manual for Trial Sequential Analysis (TSA). www.ctu.dk/tsa/.