Cochrane Scientific Committee

Recommendation statement/report

Date: July 2017

Relates to agenda item and meeting reference: 5ii 18th May 2017

Priority: Medium

Open access/restricted: Open

Review of approaches to cumulative meta-analyses for systematic reviews

Lead developers/investigators: Christian Gluud, Jørn Wetterslev, Julian Higgins, Mark Simmonds and many other colleagues

Abstract:

The problem

The CSC were asked to consider whether methods are required to manage the occurrence of both Type I and Type II errors in cumulative meta-analyses. If so, which of the proposed methods should Cochrane use.

Type I error: Repeatedly updating meta-analyses to incorporate more studies leads to the probability of type I error occurring, that is the false conclusion that an intervention has an effect when it does not (false positive). False positive results can occur due either to systematic errors, or random errors due to repeat testing.

Type II error: False negative results can occur when assuming there is no benefit before the meta-analysis has reached a sufficiently powered information size (sample size).

Summary

Julian Higgins introduced sequential approaches for meta-analyses to Cochrane at the Rome Colloquium in 1999, based on previous work by Anne Whitehead. This led to a publication in 2011 reporting a simulation study comparing six approaches and providing a worked example for **"Sequential methods for random-effects meta-analysis"**. The Higgins and colleagues' approach uses an approximate semi-bayes procedure to update evidence on the among study variance, starting with an informative prior distribution possibly based on findings from previous meta-analyses. Further work on these approaches is led by Jørn Wetterslev, Christian Gluud and colleagues (2005, 2008, 2013), referred to **"Trial Sequential Analysis in Systematic Reviews with meta-analysis" (TSA).** This work received the Thomas Chalmers award for a Cochrane Colloquium abstract. TSA is akin to the process for assessing interim analyses in trials to see whether a large enough effect (benefit) is achieved warranting trial discontinuation (stopping rules). They extend the method and test on six randomly selected meta-analyses. An important aspect to their work is the assumption that 'information size', the total number of participants across all included trials in a meta-analysis, is usually underpowered in systematic reviews and potentially increases with updates. So, they argue these MA's represent interim analyses rather than an endpoint. They suggest that this information size (when MA is underpowered), heterogeneity across studies, and bias assessment are used to provide an adjustment to the naïve 95% confidence intervals and 5% thresholds for statistical significance in meta-analysis. The Lan-DeMets' sequential monitoring boundaries in TSA provide the adjusted, expanded

confidence intervals and adjusted restrictive thresholds for statistical significance before the diversity-adjusted required information size is reached.

In 2012, *Cochrane Methods* published a discussion between Higgins on one hand and Jørn Wetterslev, Christian Gluud and colleagues on the other as to the issues raised by these methodological developments. See extract from Cochrane Methods (2012) attached.

Additional work under investigation is Shuster and Neu (2013) **"Pocock approach to sequential meta-analysis of clinical trials"** and Hu and colleagues (2007) **"Applying the law of iterated logarithm to control type I error in cumulative met-analysis of binary outcomes"**. These study reports are simulation studies with worked examples. These key approaches, are evaluated in a Cochrane funded (Methods Innovation Fund) research project led by Mark Simmonds, York University, UK. We expect this work to complete in 2018 and the CSC will receive an interim report on this work.

The documentation list provides references to these key studies and other relevant work. Methodologists do not yet agree on the approach, although they agree the principle problem of the increased probability of rejection of the null hypothesis on repeated meta-analysis and the problems with early results before the meta-analysis has reached a sufficiently powered information size. There is a mix of caution (methods not ready) and pragmatism (problem needs addressing now). Methodologists suggest Bayesian meta-analysis shows some promise (Spence et al, 2016), however, several issues need resolving, including access to software and methodological expertise.

The table below highlights some issues from key references.

Questions:

- Is the problem with too little power in most meta-analysis when a required information is not reached with false positive support for the null hypothesis a sufficient problem that undermines the evidence produced by Cochrane reviews?
- Is the problem of false positive meta-analytic conclusions due to random error introduced by underpowered meta-analysis and the probability of repeated analyses rejecting the null hypothesis a sufficient problem that undermines the evidence produced by Cochrane Reviews?
- Is the current state of development for adjustment in cumulative meta-analyses to address, specifically, type II and type I errors sufficient to recommend their implementation in Cochrane Reviews?
- If so, can the CSC recommend one or more techniques?
- If not, what further knowledge or development does the CSC need to reach a satisfactory point to decide?

Critique	By who	Reference
Sequential approaches encourage the use of significance tests and the inappropriate division of results as 'significant' or 'not significant' rather than the direct interpretation of intervention effect estimates and corresponding confidence intervals.	Higgins	*Cochrane Methods* (2012) P32-33
Problem of creating inappropriate 'stopping rules' in MA.	Higgins	*Cochrane Methods* (2012) P32-33
Measurement of accumulated information:	Higgins	*Cochrane Methods* (2012) P32-33

• The sum of the study weights in the meta-analysis. (Higgins) • Numbers of participants (Wetterslev et al.) is less sensible because the sample size needs to convert into statistical information for the analyses, and the conversion requires the additional prespecification not only of quantities such as the control group risk for dichotomous data but also of the anticipated amount of heterogeneity when a random effects meta-analysis is planned.		
Sequential methods should be applied prospectively with a full analysis plan in the protocol.	Higgins	*Cochrane Methods* (2012) P32-33
Assumptions underlying the sequential design are clearly conveyed and justified, including the parameters determining the design such as the clinically important effect size, assumptions about heterogeneity, and both the type I and type II error rates.	Higgins	*Cochrane Methods* (2012) P32-33
Major disagreement lies in whether the use of the traditional significance level of 0.05 and unadjusted 95% confidence interval is valid in MAs where the available information has not yet reached a required information size. MA results should be interpreted in the light of a realistic required information size and therefore adjustments made to ensure appropriate inference.	Wetterslev & colleagues	*Cochrane Methods* (2012) P33-35.
Response to critique for transferring TSA methods to sequential analysis in MAs – MAs impact on decisions to continue to update or not based on the level of significance. Also, the traditional unadjusted confidence interval will represent a too narrow confidence interval which by chance does not include the null effect, and so the observed effect of the intervention may be misleading and premature.	Wetterslev & colleagues	Trial Sequential Analysis in systematic reviews with meta-analysis *BMC Medical Research Methodology* (2017) 17:39. See paper for further discussion on calculating the required information size.
To overcome the type I error inflation problem Hu et al propose a way to estimate and penalize the Z statistic using the law of iterated logarithm. The penalty to the Z statistic accounts for multiple tests in a cumulative meta-analysis of binary outcomes and, in addition, accounts for estimation of heterogeneity in treatment effects across studies and the unpredictable nature of information from clinical trials. It does not require the pre-specification of the maximum information.	Hu and colleagues	Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes *Clinical Trials* (2007) 4:329-340.
In reference to methods developed by Wetterslev et al, Van der Tweel, and Bollen, and Higgins, Shuster & Neu state: None of these methods allow for the effect sizes to be dynamic. Random effects are drawn from the same conceptual urn from trial to trial. These competitors to our methods reweight the relative contributions of the included trials after each trial is added. This violates the critical independent increment property. A potential shortcoming of	Shuster & Neu	A Pocock approach to sequential meta-analysis of clinical trials. *Research synthesis Methods* (2013) 4 269-279.

all methods (including ours) lies in the lack of knowledge of the true information fraction (the ratio of the variance of the estimate at the final look presuming no stopping to that after the current look). 'Look' refers to the moment of meta-analysis in time – updating.		See paper for further explanation and methods proposed. Please see also further information in *Current controversies in data monitoring for clinical trials* (Pocock, 2006),

SUPPORTING DOCUMENTATION

[Extract] Wetterslev & colleagues and Higgins JP. Trial sequential analysis: Methods and software for cumulative meta-analyses. In Chandler J, Clarke M, Higgins JP, editors. Cochrane Methods, *Cochrane DB Syst Rev* 2012 Suppl 1:29-35.

Presentation providing an interim report on the evaluation of these methods by Mark Simmonds to the Methods symposium at the Seoul Colloquium 2016 on Living Systematic Reviews.

Relevant publications

Higgins, J P, A. Whitehead A, Simmonds M. (2011). Sequential methods for random-effects meta-analysis. *Stat Med* **30**(9): 903-921.

Hu M, Cappelleri JC, Lan KK (2007). Applying the law of iterated logarithm to control type I error in cumulative meta-analysis of binary outcomes. *Clin Trials* **4**(4): 329-340.

Imberger G, Gluud C, Boylan J, Wetterslev J.(2015). Systematic Reviews of Anesthesiologic Interventions Reported as Statistically Significant: Problems with Power, Precision, and Type 1 Error Protection. *Anesth Analg* **121**(6): 1611-1622.

Imberger G, Gluud , Boylan J, Wetterslev J. (2015). Systematic Reviews of Anesthesiologic Interventions Reported as Statistically Significant: Problems with Power, Precision, and Type 1 Error Protection. *Anesth Analg* **121**(6): 1611-1622.

Imberger G, Thorlund K, Gluud C, Wetterslev J. (2016). False-positive findings in Cochrane meta-analyses with and without application of trial sequential analysis: an empirical review. *BMJ Open* **6**(8): e011890.

Jackson D, Turner R. (2017). Power analysis for random-effects meta-analysis. *Res Synth Methods.*

Mascha, E J. (2015). Alpha, Beta, Meta: Guidelines for Assessing Power and Type I Error in Meta-Analyses. *Anesth Analg* **121**(6): 1430-1433.

Pereira, TV, Horwitz RI, Ioannidis JP.(2012). Empirical evaluation of very large treatment effects of medical interventions. *JAMA* **308**(16): 1676-1684.

Pocock SJ. (2006). Current controversies in data monitoring for clinical trials. *Clin Trials* **3**(6): 513-521.

Shuster, J. J. and J. Neu (2013). A Pocock approach to sequential meta-analysis of clinical trials. *Res Synth Methods* **4**(3): 269-279.

Spence GT, Steinsaltz D, Fanshawe TR. (2016). A Bayesian approach to sequential meta-analysis. *Stat Med* **35**(29): 5356-5375.

Thorlund, K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JP, Thabane L, Gluud LL, Als-Nielsen B,
Gluud C. (2009). Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? *Int J Epidemiol* **38**(1): 276-286.

Thorlund K, Imberger G, Walsh M, Chu R, Gluud C, Wetterslev J, Guyatt G, Devereaux PJ, Thabane L.(2011). The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis--a simulation study. *PLoS One* **6**(10): e25491.

Thorlund K, Engstrøm J, Wetterslev J, Brok J, Imberger G, Gluud C. (2011) User Manual for Trials Sequential Analysis (TSA), Copenhagen Trial Unit, Centre for Clinical Intervention Research.

Turner R, Bird M, Higgins JP. (2013). The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. *PLoS One* **8**(3): e59202.

Wetterslev J, Jakobsen JC, Gluud C. (2017). Trial Sequential Analysis in systematic reviews with meta-analysis. *BMC Med Res Methodol* **17**(1): 39.

Wetterslev J, Thorlund K, Brok J, Gluud C. (2008). Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *J Clin Epidemiol* **61**(1): 64-75.

CSC RECOMMENDATION

- [ ] *Highly recommended*
- [ ] *Recommended with provisions*
- [ ] *Optional/advisory (one among several options)*
- [ ] *Not recommended*
- [x] *Further evaluation required*

The CSC agreed that further technical examination of the key approaches was required to ascertain whether there is a preferred method, or whether the methods provide value to managing random error and are needed at all, or only in certain scenarios. An expert panel will be asked to consider the work completed by colleagues to date and will report to a future CSC.

CSC STATEMENT

Summary statement

Both presenters concurred over the problem of Type I and Type II random errors in repeated meta-analyses. These errors tend towards overestimation and the problem of leading to early

false conclusions in meta-analysis. An appropriate information or sample size protects against this and was illustrated by Christian Guuld's simulation work. He presented a specific approach, Trial Sequential Analysis, to address these errors. Mark Simmonds, as part of the Cochrane Methods Innovation Fund project, presented the findings of simulation studies evaluating four methods (above). In summary, any possible benefit is derived from either Trial Sequential Analysis, or Sequential Meta-analysis with or without using the approximate Bayesian Heterogeneity. These approaches are applied haphazardly in Cochrane and DT would like advice on the best approach.

Key points to note are:

- Both TSA and SMA control for Type I error well in Mark Simmonds's simulations, except for a few trials with high heterogeneity. Therefore, approximate Bayesian Heterogeneity is not required in most circumstances.
- Many analyses showing positive significant results at the 0.5 or lower level are based on too little evidence. And most Cochrane Reviews are underpowered.
- There is no difference in TSA or SMA in controlling for Type II error.
- Other approaches that control Type I error only, are considered overly conservative and are not recommended by Mark Simmonds's work. In addition, controlling for Type I error impacts and lessens power for Type II.
- JH (SMA) summarized his key points:
    - There are two candidates for these methods and it is problematic to suggest we select one over the other at this point.
    - He suggested we should abandon significant testing rather than create methods to correct errors that occur in their use.
    - In addition, repeated confidence intervals are an area of statistical debate, however, these methods could be converted to address repeated confidence intervals
    - Possibly not ban use of these methods, but not encourage them either and explain that they are a comparison of two hypotheses in the traditional Neyman-Pearson paradigm in statistics.
- CS stated there was other work not included here with the use of the random effects model between study variance changing overtime and how that can be accounted for, so these methods should not be used with less than five studies because you will not have a good estimate of all parameters unless you use prior information.
- Following on from whether P Values should be used at all it was noted (JS) that both confidence intervals and P Values are derived from the effect size and the standard error. The problem is that the P Value decides an arbitrary cut point (0.05) into whether a result is positive or negative. The American Association guidelines state this is not scientific and does not have utility. Confidence intervals are not used to accept or reject the null hypothesis.

The discussion indicated that these methods were driven by both technical and theoretical issues that warranted greater examination with other experts in the field.

Credibility & validity

Further work required to assess utility of the approaches.

Limitations/caveats

This issue resides within current theoretical debates amongst statisticians.

Areas of concern/uncertainty

Unclear at this point.

Impact on Cochrane

Training and guidance and utility of method to make a difference.

Cochrane resources needed

None currently.

Implementation

CSC members are not responsible for managing implementation of these recommendations which will require an implementation plan to ensure co-ordination for a smooth introduction. This will include launch, timescales and roll out strategy. Therefore, this statement does not signify immediate implementation.