Does Cochrane reviews reasonably take into account risks of type I errors and type II errors - and what can we do about it?

Christian Gluud
Copenhagen Trial Unit
Centre for Clinical Intervention Research
Rigshospitalet, Denmark
What is the evidence-base?

<table>
<thead>
<tr>
<th>Study (year)</th>
<th>Key finding regarding overestimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ioannidis (2002)</td>
<td>&gt; 30% fluctuations before 2000 patients 10,000 patients to relieve uncertainty</td>
</tr>
<tr>
<td>Trikalinos (2005)</td>
<td>&gt; 50% fluctuations before 500 patients 2000 patients to relieve uncertainty</td>
</tr>
<tr>
<td>Thorlund (2009)</td>
<td>1/3 of ‘first statistically significance’ are clinically important overestimates</td>
</tr>
<tr>
<td>Pereira (2011)</td>
<td>Updated estimates 0.67-fold smaller when original MA has &lt; 300 events</td>
</tr>
</tbody>
</table>
What is the evidence-base?

Overestimation

A simulation study has quantified the impact of random error on overestimation of meta-analysed intervention effects in relation to the cumulative number of events and patients

Thorlund et al, PLoS ONE 2011
Simulated meta-analyses with more than 20% or 30% overestimation of a truly zero effect

Scenario: RRR=0%

RRR>20% (upper curve)
RRR>30% (lower curve)

Control group risk 5%-15%

Moderate heterogeneity

Risk of overestimation plotted as a function of the cumulated number of patients and events
What is the evidence-base?

Overestimation
- The likelihood of overestimation due to random error is profound at early stages (often 20% to 30%)
- The risk of overestimation decreases exponentially with number of participants and outcomes
- Reaching the required information size provides good protection against overestimation

Thorlund et al, PLoS ONE 2011
What is the evidence-base?

Lack of power in Cochrane reviews
Proportion of 77,237 trials in 14,886 meta-analyses of binary outcomes adequately powered to detect a relative risk reduction (RRR)

<table>
<thead>
<tr>
<th>RRR</th>
<th>≥ 80% power in all trials (%)</th>
<th>≥ 80% power in meta-analyses (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>2.0%</td>
<td>ND*</td>
</tr>
<tr>
<td>30%</td>
<td>17.0%</td>
<td>22.0%</td>
</tr>
</tbody>
</table>

Turner et al, PLoS ONE 2011

* Not determined
Houston, we have a problem!

Apollo 13
Available approaches

- The required information size
- Trial Sequential Analysis (TSA)
- Sequential Meta-Analysis (SMA)
- The Law of the Iterated Logarithm (LIL)
- Bayesian methods
The required information size (optimal information size)

The required information size (RIS) is the required meta-analysis ‘sample size’

Reaching the RIS and the corresponding number of trials ensures control of type I and type II errors

Reaching the RIS also provides good protection against overestimation of the intervention effect
Trial Sequential Analysis

- Number of patients
- Cumulated Z-curve
- P = 0.05
- Trial sequential monitoring boundary
- Futility area
- Z-score

(Number of events)
False-positive findings in Cochrane meta-analyses with and without TSA

TSA prevented 13 / 14 statistically significant meta-analyses (P<0.05) among cumulative meta-analyses becoming non-significant when RIS was reached

Imberger et al, BMJ OPEN 2016
Thank you
<table>
<thead>
<tr>
<th><strong>PROS</strong></th>
<th><strong>CONS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhanced reliability of inferences about effects</td>
<td>Conclusions from Cochrane systematic reviews will become more conservative</td>
</tr>
<tr>
<td>Avoidance of frequent reversed statistical significance diminishing credibility</td>
<td>Choice of anticipated intervention effect and heterogeneity may require Bayesian methods and be perceived complicated</td>
</tr>
<tr>
<td>Yardsticks for amount of evidence required for conclusive meta-analyses (aligned with GRADE)</td>
<td>May discourage authors from conducting Cochrane reviews due to reduced probability of statistically significant findings</td>
</tr>
<tr>
<td>Better direction of resource use</td>
<td></td>
</tr>
<tr>
<td>Approach</td>
<td>PROS</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
</tr>
<tr>
<td>TSA</td>
<td>Known from trials Adjusted CIs before RIS Information=patients Futility boundaries</td>
</tr>
<tr>
<td>SMA</td>
<td>Known from trials Adjusted CIs Semi-Bayesian (full distribution of heterogeneity) Futility boundaries</td>
</tr>
<tr>
<td>LIL</td>
<td>Can obtain adjusted P-values</td>
</tr>
</tbody>
</table>
Sequential Meta-Analysis

Efficient score vs. Statistical (Fischer) information
Law of the iterated logarithm

Penalised Z-score vs. Number of Patients

$\frac{Z}{\lambda}$
Outline

Random errors in meta-analysis
- how big a problem?

Approaches for dealing with random errors

Current practice versus change of practice
Before reaching RIS...

In the light of the available evidence, it seems sensible to employ methods that will provide more reliable inferences about the ‘statistical significance’, the magnitude of the intervention effect, and the associated confidence interval.
Available frequentist approaches

- Trial Sequential Analysis (TSA) O’Brien-Fleming type alpha-spending adjusted thresholds for significance (Z-score)

- Sequential meta-analysis (SMA) O’Brien-Fleming type Whitehead adjusted thresholds for significance (efficient score)

- Law of the iterated logarithm (LIL): penalisation (adjustments) of Z-score
Performance of the approaches

TSA and SMA have theoretically identical backgrounds

Simulation studies have demonstrated that SMA and LIL generally provide good control of the type I error

Empirical evidence suggests that TSA provide adequate protection against false positives and clinically important overestimates
Random errors in meta-analysis
- how big a problem?

Approaches for dealing with random errors

Current practice versus change of practice
Current practice vs change

What are the PROS and CONS of implementing either of these methods?

What impact will the more conservative nature of these methods have on Cochrane systematic reviews and The Cochrane Collaboration?
Recap

Random error, in concert with repeated testing, causes increased risks of false statistically significant findings (spurious $P < 0.05$ results)

Random error is likely to cause important intervention effect overestimates in meta-analyses with sparse data
What is the evidence-base?

Repeated testing
Three empirical studies have explored the probability of ‘early’ false positive results

In two, 20% to 25% of meta-analyses had at least one temporary instance of false statistical significance (P < 0.05)

In one study, the probability of reversed statistical significance after one update was 16% to 37%
What is the evidence-base?

Repeated testing
5 simulation studies have investigated increase of type I error under repeated testing

Depending on the simulated scenario and number of repeated tests, the overall type I error can be anywhere between 10% and 40%
What is the evidence-base?

Overestimation

4 empirical studies explored ‘early’ fluctuations in intervention effects estimates and the probability of ‘clinically important’ overestimation
Outline

Random errors in meta-analysis
- how big a problem?

Approaches for dealing with random errors
Outline

Random errors in meta-analysis - how big a problem?

Approaches for dealing with random errors