Common errors in meta-analysis

Lessons from the Cochrane Review Screening Programme

November 2017

Kerry Dwan
Objectives

The objectives of this workshop are to highlight common statistical errors made in Cochrane Systematic Reviews, and to provide practical, hands on learning and guidance to help authors and editors address these errors.

❖ Slides with examples
❖ Practical Exercises
❖ General Discussion
Poll

What are your roles in Cochrane?

- Editor
- Author
- Statistician
- Other
- No role in Cochrane yet
Common Errors

- Funny Looking Results
- Analyses
- Errors we may not see
FLR (Funny Looking Results)

1. Data entry errors/ transposition errors
2. Study weight at odd with sample size
3. Outliers
4. Study ID appearing more than once in a forest plot
5. Reporting at odds with forest plot
Poll

Which study do you think is probably erroneous?

- Study 4
- Study 5
- Study 6
FLR #1 - Data Entry Error

Study 4 Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Baseline Mean (SD)</th>
<th>Placebo Mean (SD)</th>
<th>Clomipramine Mean (SD)</th>
<th>Haloperidol Mean (SD)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>CARS</td>
<td>41.8 (7.1)</td>
<td>39.4 (7.0)</td>
<td>37.8 (8.7)</td>
<td>36.7 (6.1)</td>
<td>0.05^b</td>
</tr>
<tr>
<td>ESRS</td>
<td>6.6 (6.7)</td>
<td>7.9 (7.1)</td>
<td>10.3 (7.3)</td>
<td>7.8 (5.8)</td>
<td>0.35^c</td>
</tr>
<tr>
<td>DOTES</td>
<td>0.6 (2.2)</td>
<td>0.8 (1.7)</td>
<td>2.0 (2.9)</td>
<td>2.3 (3.3)</td>
<td>0.07^d</td>
</tr>
</tbody>
</table>
FLR #2 – Study weight at odds with sample size

Poll
Which study do you think is probably erroneous?
• Study 1
• Study 2

Question: why? (type the answer in your question box)
FLR #3 – Outliers

Minus sign left off mean

SEMs used instead of SDs
**FLR #4 – Study ID appearing >1 in a forest plot**

Question: what is the problem with this? (type the answer in your question box)
Studies included multiple times
‘Higher proportions of participants were reported to experience side effects in the treatment group compared with placebo (100% vs 25%; RR 6.04, 95% CI 2.67 to 13.65)’.

Question: what is the issue here? (type the answer in your question box)
The confidence intervals for the estimated HR include large benefit and moderate harm of intervention (0.88; 95% CI 0.64 to 1.12), P = 0.43

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Hazard Ratio</th>
<th>SE</th>
<th>Intervention Total</th>
<th>Control Total</th>
<th>Weight</th>
<th>Hazard Ratio IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>0.88</td>
<td>0.12</td>
<td>5472</td>
<td>3940</td>
<td>100.0%</td>
<td>0.88 [0.64, 1.12]</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td></td>
<td></td>
<td>5472</td>
<td>3940</td>
<td>100.0%</td>
<td>0.88 [0.64, 1.12]</td>
</tr>
</tbody>
</table>

Heterogeneity: Not applicable
Test for overall effect: Z = 7.33 (P < 0.00001)

Question: what is the issue here? (type the answer in your question box)
Analysis

1. Unit of analysis
   • Crossover trials (Nolan et al. PLoS ONE 2016)
   • Cluster trials (Richardson et al. PLoS ONE 2016)

2. Subgroups
   • Post hoc, wrong analysis, incorrect interpretation
   • Adequate number of studies, 10?
   • Specify small number of characteristics in advance with rationale (Donegan et al. PLoS ONE 2016)

3. SMDs and MDs
   • Used incorrectly, not often back transformed

4. Random effects versus fixed effects
   • Inconsistently used
1. Unit of analysis

- **Unit of analysis issues**

We planned to take into account the level at which randomisation occurred, such as cross-over trials, cluster-randomised trials and multiple observations for the same outcome. In case of cross-over trials or cluster-randomised trials, we planned to extract estimates of effect that took into account the correlation of the measurements.

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Cognition-focused care</th>
<th>Enhanced standard care</th>
<th>Std. Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Total</td>
</tr>
<tr>
<td>Davies 2008 (1)</td>
<td>14.1</td>
<td>24.07276</td>
<td>437</td>
</tr>
<tr>
<td>Fisher 2011</td>
<td>1.78</td>
<td>0.96</td>
<td>256</td>
</tr>
<tr>
<td>Glasgow 2005</td>
<td>27.4</td>
<td>32.81</td>
<td>459</td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>1162</td>
<td>1162</td>
<td>1031</td>
</tr>
</tbody>
</table>

Heterogeneity: Tau² = 0.00; Chi² = 3.53, df = 2 (P = 0.17); I² = 43%
Test for overall effect: Z = 0.28 (P = 0.78)

Footnotes:
(1) Median values were reported, SDs values were calculated based on reported change in mean and P value

- Unadjusted data from study reports used in analysis
Practical Exercise 1
Practical Exercise 1 - Feedback
Practical Exercise 1 – Solutions

Figure 1: outcome 1

Figure 2: outcome 2
2. Comparing Subgroups

- use a formal statistical test to compare subgroups

**Abstract:** Our Review suggests that (INTERVENTION) may have more beneficial effects in (SUBGROUP).

**PLS:** In the further analyses, there is evidence indicated that the effects of (INTERVENTION) in reducing (OUTCOME) rate may be different between (SUBGROUP 1) and (SUBGROUP 2), with more benefits observed in (SUBGROUP 1).
Practical Exercise 2
Practical Exercise 2 - Feedback
## Practical Exercise 2 – Solutions

### Table 1.1: Risk Ratio for Intervention X

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Experimental Events</th>
<th>Control Events</th>
<th>Total Total</th>
<th>Weight</th>
<th>Risk Ratio M-H, Random, 95% CI</th>
<th>Risk Ratio M-H, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.1.1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 3</td>
<td>31</td>
<td>26</td>
<td>115</td>
<td>31.7%</td>
<td>1.23 [0.70, 1.94]</td>
<td></td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>31</td>
<td>26</td>
<td>115</td>
<td>31.7%</td>
<td>1.23 [0.78, 1.94]</td>
<td></td>
</tr>
<tr>
<td><strong>1.1.2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 1</td>
<td>20</td>
<td>34</td>
<td>54</td>
<td>31.5%</td>
<td>0.62 [0.39, 0.99]</td>
<td></td>
</tr>
<tr>
<td>Study 2</td>
<td>4</td>
<td>12</td>
<td>16</td>
<td>16.5%</td>
<td>0.33 [0.11, 0.98]</td>
<td></td>
</tr>
<tr>
<td>Study 4</td>
<td>7</td>
<td>4</td>
<td>11</td>
<td>21.2%</td>
<td>0.49 [0.22, 1.10]</td>
<td></td>
</tr>
<tr>
<td>Subtotal (95% CI)</td>
<td>31</td>
<td>60</td>
<td>184</td>
<td>68.3%</td>
<td>0.55 [0.38, 0.79]</td>
<td></td>
</tr>
<tr>
<td>Total events</td>
<td>31</td>
<td>60</td>
<td>299</td>
<td>100%</td>
<td>0.67 [0.39, 1.14]</td>
<td></td>
</tr>
</tbody>
</table>

**Heterogeneity:** Not applicable
**Test for overall effect:** $Z = 0.91$ ($P = 0.36$)

### Graph 1.2: Forest Plot

#### 1.2.1
- Study 3
- Subtotal
- Study 1
- Study 2
- Study 4
- Subtotal

#### 1.2.2
- Study 1
- Study 2
- Study 4
- Subtotal

**Total (95% CI)**

**Heterogeneity:** $I^2 = 0.00$, $I^2 = 1.20$, $I^2 = 0.55$, $I^2 = 0.00$
**Test for overall effect:** $Z = 3.13$ ($P = 0.001$)

**Main results**
The effect of intervention X on reducing outcome A was uncertain due to the low quality of the evidence (RR 0.67, 95% CI 0.39 to 1.14; 605 participants; 4 studies). Subgroup analysis by type of intervention X provided limited evidence that X (b) may lower the risk of outcome A.
“We will convert continuous outcome data into standardised mean differences (SMDs) and present with 95% CIs, as it is assumed that study authors will use different measurement scales. If continuous outcome data is recorded using the same measurement scale, data will be converted into mean differences (MDs) and presented with 95% CIs”.

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Experimental Mean</th>
<th>SD</th>
<th>Total</th>
<th>Control Mean</th>
<th>SD</th>
<th>Total</th>
<th>Weight</th>
<th>Std. Mean Difference IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 4</td>
<td>23</td>
<td>12.9</td>
<td>62</td>
<td>46</td>
<td>15.2</td>
<td>51</td>
<td>100.0%</td>
<td>-1.63 [-2.06, -1.20]</td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td></td>
<td></td>
<td>62</td>
<td></td>
<td></td>
<td>51</td>
<td>100.0%</td>
<td>-1.63 [-2.06, -1.20]</td>
</tr>
</tbody>
</table>

Heterogeneity: Not applicable
Test for overall effect: Z = 7.46 (P < 0.00001)
We considered statistical heterogeneity between trials to be substantial if, following meta-analysis, $I^2$ was greater than 30% and either $T^2$ is greater than zero, or there was a low $P$-value (< 0.10) in the Chi² test for heterogeneity. If substantial heterogeneity was identified used the random-effects (RE) model instead of the fixed-effects (FE) model to pool data.

4. Fixed Effect versus Random Effects

**Question:** what is the problem here? (type the answer in your question box)
Errors we may not see

• Have any papers been missed?

• Have the right results been copied from the papers?

• Have the standard deviations been confused with standard errors?

Question: Are there any other errors we may not see? (type suggestions in your question box)
Test drive

training.cochrane.org/common-errors
Final Tips
Tips for spotting errors

- Numbers that stand out (perfect homogeneity, single outlying results, sample size does not match with precision relative to other studies)

- For non-standard RCT designs - evidence of how SEs were adjusted (check methods against plots).

- For primary outcomes select the biggest study or the one that has most weight and check the analysis results against the paper.

- For other outcomes pick a study entirely at random and check numbers used against what is available in published trial report or elsewhere. If authors have stated that they got unpublished data then move on to next study.
Discussion
References and resources


MECIR http://methods.cochrane.org/mecir
training.cochrane.org/common-errors
Practical Exercises 3 and 4 - Feedback
# Practical Exercise 3 – Solutions

## TABLE 1: Total diary score over the first 10 days: mean (SD)

<table>
<thead>
<tr>
<th>Group</th>
<th>Bed rest and exercise and education (n = 50)</th>
<th>Exercise and education (n = 41)</th>
<th>Bed rest (n = 47)</th>
<th>Control (n = 48)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>22.7 (6.14)</td>
<td>23.30 (6.92)</td>
<td>21.66 (6.54)</td>
<td>21.54 (6.31)</td>
</tr>
<tr>
<td>Activities</td>
<td>24.36 (6.75)</td>
<td>21.34 (9.22)</td>
<td>24.34 (10.04)</td>
<td>20.99 (8.48)</td>
</tr>
<tr>
<td>Pain</td>
<td>23.7 (5.22)</td>
<td>25.94 (7.47)</td>
<td>24.15 (7.12)</td>
<td>22.68 (5.88)</td>
</tr>
</tbody>
</table>

*Note: Lower total scores indicate a better clinical result.*

## Study or Subgroup

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Experimental</th>
<th>Control</th>
<th>Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>Total</td>
<td>Mean</td>
</tr>
<tr>
<td>Brown 2003</td>
<td>19.1</td>
<td>21</td>
<td>145</td>
</tr>
<tr>
<td>Gilbert 1895</td>
<td>23.77</td>
<td>5.88</td>
<td>65</td>
</tr>
<tr>
<td>Smith 2015</td>
<td>31</td>
<td>17.21</td>
<td>42</td>
</tr>
</tbody>
</table>

**Total (95% CI):** 252 273 100.0% 1.56 [-0.17, 3.28]

*Heterogeneity: Chi^2 = 3.35, df = 2 (P = 0.19), I^2 = 40%*

*Test for overall effect: Z = 1.77 (P = 0.08)*
## Practical Exercise 4 – Solutions

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Experimental</th>
<th>Control</th>
<th>Std. Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Total</td>
</tr>
<tr>
<td>Study 1</td>
<td>0.07</td>
<td>0.3</td>
<td>69</td>
</tr>
<tr>
<td>Study 2</td>
<td>-1.02</td>
<td>4.77</td>
<td>50</td>
</tr>
<tr>
<td>Study 3</td>
<td>3.02</td>
<td>0.08</td>
<td>62</td>
</tr>
<tr>
<td>Study 4</td>
<td>-0.41</td>
<td>1.29</td>
<td>11</td>
</tr>
<tr>
<td>Study 5</td>
<td>-7.07</td>
<td>5.05</td>
<td>28</td>
</tr>
<tr>
<td>Study 6</td>
<td>-0.5</td>
<td>4.3</td>
<td>37</td>
</tr>
<tr>
<td>Study 7</td>
<td>0.31</td>
<td>3.08</td>
<td>31</td>
</tr>
<tr>
<td><strong>Total (95% CI)</strong></td>
<td>288</td>
<td></td>
<td>242</td>
</tr>
</tbody>
</table>

**Test for overall effect:** $Z = 2.03$ ($P = 0.04$)

**Risk of bias legend**

(A) Random sequence generation (selection bias)

(B) Allocation concealment (selection bias)

(C) Blinding of participants and personnel (performance bias)

(D) Blinding of outcome assessment (detection bias)

(E) Incomplete outcome data (attrition bias)

(F) Selective reporting (reporting bias)

(G) Sample Size

(H) Other bias

---

**Heterogeneity:**

- $\tau^2 = 1.77$
- $\chi^2 = 137.94$, df = 6 ($P < 0.00001$)
- $I^2 = 96%$

---

**Favours [control] Favours [experiment]**
### Practical Exercise 4 – Solutions

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Experimental Mean</th>
<th>SD</th>
<th>Total</th>
<th>Control Mean</th>
<th>SD</th>
<th>Total</th>
<th>Weight</th>
<th>IV, Random, 95% CI</th>
<th>Std. Mean Difference</th>
<th>IV, Random, 95% CI</th>
<th>Risk of Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>0.07</td>
<td>0.3</td>
<td>69</td>
<td>-0.04</td>
<td>0.3</td>
<td>5</td>
<td>7.8%</td>
<td>0.36 [-0.55, 1.27]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 2</td>
<td>-1.02</td>
<td>4.77</td>
<td>50</td>
<td>-5.93</td>
<td>4.57</td>
<td>52</td>
<td>17.6%</td>
<td>1.04 [0.63, 1.46]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 3</td>
<td>3.02</td>
<td>0.8</td>
<td>62</td>
<td>2.59</td>
<td>0.8</td>
<td>40</td>
<td>17.9%</td>
<td>0.53 [0.13, 0.94]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 4</td>
<td>-0.45</td>
<td>1.29</td>
<td>11</td>
<td>-1.11</td>
<td>1.45</td>
<td>9</td>
<td>8.0%</td>
<td>0.46 [-0.43, 1.36]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 5</td>
<td>-7.07</td>
<td>5.05</td>
<td>28</td>
<td>-7.85</td>
<td>3.59</td>
<td>26</td>
<td>14.5%</td>
<td>0.17 [-0.36, 0.71]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 6</td>
<td>-0.5</td>
<td>4.3</td>
<td>37</td>
<td>-0.9</td>
<td>4.2</td>
<td>69</td>
<td>18.0%</td>
<td>0.09 [-0.31, 0.49]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 7</td>
<td>0.31</td>
<td>3.08</td>
<td>31</td>
<td>0.06</td>
<td>2.85</td>
<td>41</td>
<td>16.2%</td>
<td>0.08 [-0.38, 0.55]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>288</td>
<td></td>
<td>242</td>
<td></td>
<td></td>
<td>100.0%</td>
<td></td>
<td>0.40 [0.10, 0.70]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Heterogeneity: Tau² = 0.09; Chi² = 14.37, df = 6 (P = 0.03); I² = 58%
Test for overall effect: Z = 2.58 (P = 0.010)

Risk of bias legend
(A) Random sequence generation (selection bias)
(B) Allocation concealment (selection bias)
(C) Blinding of participants and personnel (performance bias)
(D) Blinding of outcome assessment (detection bias)
(E) Incomplete outcome data (attrition bias)
(F) Selective reporting (reporting bias)
(G) Sample Size
(H) Other bias