Magnitude of bias in trials stopped early for benefit reviews

Dirk Bassler, Victor Montori, Paul Glasziou
Stephen Walter, Qi Zhou, Stephen Walter
Gordon Guyatt

clarity
Clinical Advances Through Research and Information Translation
Stopping early apparent benefit

- **ethical mandate**
  - unethical to randomize to control
  - priority to get effective treatment to patients

- **increasing proportion of trials stopping early**

- **danger**
  - arbitrary stopping violates statistical principles
  - statistically sound stopping rules

- **remaining danger**
  - rules may not be observed
  - simulations suggest still overestimate effect
  - systematic review suggests overestimate in real world:
    almost 50% of 143 trials RRR > 50%; 25% RRR > 70%
Addressing uncertainty

• survey didn’t prove overestimates

• survey suggested large less problems
  - OR 31 for RRR > 47% for events < 66
  - also not proved

• what is average overestimate?
  - what factors associated?
Study design

- obtain all trials stopped early for benefit

- obtain meta-analyses
  - same question (population, intervention, comparator)
  - outcome that drove early stopping
  - if tRCT non included, update meta-analysis

- compare effects
  - tRCTs versus non-tRCTs
  - predictors of difference
    - rigorous rule yes/no
    - sample size/number of events
    - methodologic quality
Details of methods

- search included MEDLINE, Embase, Current Contents
  - databases including full text of journals (OVID, ScienceDirect, Ingenta, and Highwire Press, Lancet, New England Journal of Medicine, JAMA, Annals of Internal Medicine, BMJ)

- duplicate assessment of eligibility
  - blind to results
  - reviewers content area expertise

- duplicate data abstraction
Analysis

• ratio of RRs of individual tRCTs to corresponding non-tRCTs:
  \[ \log(\text{ratio of RRs}) = \log(\text{RR of tRCT} / \text{RR of pooled non-tRCTs}) \]
  \[ = \log(\text{RR of tRCT}) - \log(\text{RR of pooled non-tRCTs}) \]

• overall estimate
  - \[ \log(\text{ratio of RRs}) \] inverse variance-weighted average of \[ \log(\text{ratio of RRs}) \]
  - back transformed to the overall ratio of RRs

• two meta-regressions

• first dependent variable log of difference in RRs of tRCTs and non-tRCTs
  - independent variables use of stopping rule, number of events

• second hierarchicial meta-regression
  - meta-analysis and individual study were levels in hierarchy
  - dependent variable log RR of each individual study
  - independent variables added concealment, blinding, stopping early
tRCTs identified in prior systematic review (n=143)

Additional tRCTs identified (n=52)
Total tRCTs as basis for literature search (n=195)

Relevant SRs identified (n=238)

SRs updated (n=32)  SRs not updated (n=206)

Potentially relevant RCTs retrieved and blinded (n=2488)

Included in analysis:
- 91 tRCTs
- 424 matching non-tRCTs
- 63 research questions

Excluded because insufficient similarity to the tRCT or not randomized (n=2012)
- RR not calculable (n=30)
- Truncated early for benefit (n=22)
Study Characteristics

- **area of study**
  - cardiology > 35%, no other concentration

- **publication in high impact journals**
  - 62 tRCTs (68%), 128 non-tRCTs (30%)

- **methods**
  - concealment 53% and 34%; blinding 60%

- **preplanned stopping rule**
  - 76% of tRCTs, 13% of non-tRCTs
55/63 “favor” tRCT
20/63 significantly “favor” tRCT
if RR non-tRCT 0.8
RR tRCT 0.57
more than double RRR
39/63 (62%) results of non-tRCTs > 0.05
16/63 (25%) non-tRCTs RR > 0.90
## Predictors of difference

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Parameter (95%CI)</th>
<th>p-value</th>
<th>R-square*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Univariable Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stopping rule</td>
<td>0.14 (0.02, 0.27)</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Univariable Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every 100 events in the tRCT</td>
<td>0.0169 (0.0088, 0.025)</td>
<td>&lt; 0.0001</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Multivariable Model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stopping rule</td>
<td>0.07 (-0.05, 0.19)</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>Every 100 events in the tRCT</td>
<td>0.0151 (0.0066, 0.0237)</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
</tbody>
</table>

**Concealment**  p = .96  
**Blinding**  p = 0.32
Conclusions

- trials stopped early for benefit overestimate magnitude of treatment effects
  - overestimates substantial, potentially effect treatment decisions
  - may sometime create completely spurious treatment effects

- overestimates less with large sample size
  - but overestimates still substantial
  - probably need > 500 events before safe from major overestimates
Editorial comments

• problem made worse by
  - publication in top journals
  - may obscure adverse effects

• ethics questionable
  - scientific value (overestimated compromise)
  - value to society (dissemination of overestimates)

• if really unethical to continue
  - should be no subsequent trials addressing question

• DMCs stop only when completely confident
  - our results suggest never that confident
Alternative comparison

• ideal comparator
  - no stopping rule, not stopped
  - unidentifiable, not feasible

• alternative
  - all trials including stopped early for benefit
  - rationale non-tRCTs will underestimate, simulations
  - RR 0.85 (95% CI: 0.81 - 0.91)
  - 16 of 63 (25%) p>0.05

• simulations suggest low weight tRCTs
  - 28% (interquartile range 12% to 40%)
  - 37 (60%) tRCTs more than 20% of weight
  - possibly stopped early phenomenon