Stopping trials early for benefit – a problem?

Review of the empirical evidence

Matthias Briel MD MSc
Senior Scientist, Basel Institute for Clinical Epidemiology and Biostatistics, Switzerland
Assistant Professor, McMaster University, Hamilton, Canada
STOP-IT Team

- Victor Montori
- Dirk Bassler
- Paul Glasziou
- Elie Akl
- German Malaga
- Ignacio Ferreira-Gonzalez
- Pablo Alonso-Coello
- Gerard Urrutia
- Regina Kunz
- Heloisa Soares
- Nisrin Abu Elnour
- John You
- Paul Karanicolas
- Heiner Bucher
- Julianna Lampropulos
- Alain Nordmann
- Amit Sood
- Jagdeep Kaur
- Clare Bankhead
- Rebecca Mullan
- Kristina Lutz
- Christine Ribic
- Noah Vale
- Patricia Erwin
- Rafael Perera
- Qi Zhou

- Stephen Walter
- Melanie Lane
- Carolina Ruiz Culebro
- Suzana Alves da Silva
- David Flynn
- Mohamed Elamin
- Brigitte Strahm
- Hassan Murad
- Benjamin Djulbegovic
- Neill Adhikari
- Ed Mills
- Femida Gwadry-Sridhar
- Haresh Kirpalani
- Karen Burns
- Sohail Mulla
- Heike Raatz
- Kara Nerenberg
- Per Olav Vandvik
- Fernando Coto-Yglesias
- Holger Schünemann
- Fabio Tuche
- Pedro Chrispim
- Deborah Cook
- Diane Heels-Ansdell
- Tim Ramsay
- Gordon Guyatt
Hamilton Health Sciences, McMaster University
Why stop early?

- RCTs typically plan sample size
- why would investigators stop early?
  - harm
  - futility
  - benefit = truncated RCT (tRCT)
- harm: ethical imperative
- futility: optimal resource allocation
- benefit: resource allocation, ethical imperative
Agenda

- Why tRCTs tend to overestimate
- Epidemiology of tRCTs (STOP-IT 1)
- Systematic review of meta-analyses including tRCTs (STOP-IT 2)
- Implications of results
Five vs Four Courses of Therapy for Acute Myeloid Leukemia

<table>
<thead>
<tr>
<th>Timepoint</th>
<th>Deaths/Patients</th>
<th>HR &amp; 95% CI</th>
<th>Odds Redn. (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Five courses</td>
<td>Four courses</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(O-E) Var.</td>
<td>(Odds : )</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>7/102</td>
<td>15/100</td>
<td>57% (29); 2P = 0.05</td>
</tr>
</tbody>
</table>
## Five vs Four Courses of Therapy for Acute Myeloid Leukemia

<table>
<thead>
<tr>
<th>Timepoint</th>
<th>Deaths/Patients</th>
<th>Statistics</th>
<th>Odds Redn.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Five courses</strong></td>
<td><strong>Four courses</strong></td>
<td>(O-E)</td>
</tr>
<tr>
<td>1997</td>
<td>7/102</td>
<td>15/100</td>
<td>-4.6</td>
</tr>
<tr>
<td>1998 (1)</td>
<td>23/171</td>
<td>42/169</td>
<td>-12.0</td>
</tr>
</tbody>
</table>

Wheatley K, Clayton D. Controlled Clinical Trials 2003;24:66-70
<table>
<thead>
<tr>
<th>Timepoint</th>
<th>Deaths/Patients</th>
<th>Statistics</th>
<th>HR &amp; 95% CI</th>
<th>Odds Redn. (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Five courses</td>
<td>Four courses</td>
<td>(O–E) Var.</td>
<td>Five courses : Four courses</td>
</tr>
<tr>
<td>1997</td>
<td>7/102</td>
<td>15/100</td>
<td>-4.6</td>
<td>5.5</td>
</tr>
<tr>
<td>1998 (1)</td>
<td>23/171</td>
<td>42/169</td>
<td>-12.0</td>
<td>15.9</td>
</tr>
<tr>
<td>1998 (2)</td>
<td>41/240</td>
<td>66/240</td>
<td>-16.0</td>
<td>26.7</td>
</tr>
</tbody>
</table>

Wheatley K, Clayton D. Controlled Clinical Trials 2003;24:66-70
## Five vs Four Courses of Therapy for Acute Myeloid Leukemia

Wheatley K, Clayton D. Controlled Clinical Trials 2003;24:66-70

<table>
<thead>
<tr>
<th>Timepoint</th>
<th>Deaths/Patients</th>
<th>Statistics</th>
<th>HR &amp; 95% CI</th>
<th>Odds Redn. (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Five courses</td>
<td>Four courses</td>
<td>(O-E)</td>
<td>Var.</td>
</tr>
<tr>
<td>1997</td>
<td>7/102</td>
<td>15/100</td>
<td>-4.6</td>
<td>5.5</td>
</tr>
<tr>
<td>1998 (1)</td>
<td>23/171</td>
<td>42/169</td>
<td>-12.0</td>
<td>15.9</td>
</tr>
<tr>
<td>1998 (2)</td>
<td>41/240</td>
<td>66/240</td>
<td>-16.0</td>
<td>26.7</td>
</tr>
<tr>
<td>1999</td>
<td>51/312</td>
<td>69/309</td>
<td>-11.9</td>
<td>30.0</td>
</tr>
<tr>
<td>2000</td>
<td>79/349</td>
<td>91/345</td>
<td>-9.5</td>
<td>42.4</td>
</tr>
<tr>
<td>2001</td>
<td>106/431</td>
<td>113/432</td>
<td>-6.2</td>
<td>53.7</td>
</tr>
<tr>
<td>2002</td>
<td>157/537</td>
<td>140/541</td>
<td>6.7</td>
<td>74.0</td>
</tr>
</tbody>
</table>

Five courses better | Four courses better
Look after every patient or event
Interim analyses every $q$ patients or events
True beneficial effect
No effect
Stopping boundary
Poldermans, NEJM, 1999

• RCT elective vascular surgery
  – compared bisoprolol to placebo

• primary endpoint cardiac death or nonfatal MI

• prior planned single look at 100 pts
  – stop if exceeded O’Brien-Fleming boundary \( p < 0.001 \)
  – planned sample size 266, stopped at 112

• primary endpoint cardiac death or MI
  – 2 of 59 (3.4%) in bisoprolol group
  – 18 of 53 (34%) in placebo
  – RR 0.09, 95% CI 0.02 to 0.37, \( P < 0.001 \)
Composite – fixed effects

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Relative Risk (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jakobsen</td>
<td>1997</td>
<td>3.00 (0.13 to 69.09)</td>
</tr>
<tr>
<td>Wallace</td>
<td>1998</td>
<td>0.65 (0.17 to 2.41)</td>
</tr>
<tr>
<td>Bayliff</td>
<td>1999</td>
<td>0.73 (0.15 to 3.52)</td>
</tr>
<tr>
<td>Poldermans</td>
<td>1999</td>
<td>0.12 (0.03 to 0.43)</td>
</tr>
</tbody>
</table>

p=0.11 for heterogeneity, I²=50%

Relative Risk (95% CI): 0.40 (0.18 to 0.85)
ACC/AHA PRACTICE GUIDELINES

ACC/AHA Guideline Update for Perioperative Cardiovascular Evaluation for Noncardiac Surgery—Executive Summary


COMMITTEE MEMBERS

KIM A. EAGLE, MD, FACC, Chair

PETER B. BERGER, MD, FACC
HUGH CALKINS, MD, FACC
BERNARD R. CHAITMAN, MD, FACC
GORDON A. EWY, MD, FACC
KIRSTEN E. FLEISCHMANN, MD, MPH, FACC
LEE A. FLEISHER, MD, FACC

JAMES B. FROEHLICH, MD, FACC
RICHARD J. GUSBERG, MD, FACS
JEFFREY A. LEPPO, MD, FACC
THOMAS RYAN, MD, FACC
ROBERT C. SCHLANT, MD, FACC
WILLIAM L. WINTERS, Jr, MD, MACC

TASK FORCE MEMBERS

RAYMOND J. GIBBONS, MD, FACC, Chair
Subsequent best estimates

- > 10,000 randomized
  - myocardial infarction 0.71 (0.57 to 0.86)
  - death 1.23 (0.98 – 1.55)
  - stroke 2.21 (1.37 – 3.55)

- clinical expert community
  - still recommending beta blockers!

- same story with
  - low glucose target ICU

- other prominent overestimates
  - low tidal volume ventilation
  - activated protein C
The Epidemiology of Clinical Trials Stopped Early for Benefit

STOP-IT 1

A systematic review of trials stopped early for benefit

• eligibility
  – RCTs reported stopped early because of finding in favor of experimental intervention

• search
  – 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, and BMJ)

• duplicate assessment of eligibility, data extraction
STOP-IT 1

- 143 eligible trials

- increasing use
  - 1975-1979: 1/6574 (0.001%) 0/620 (0%)
  - 1980-1984: 1/12653 (0.008%) 1/1175 (0.1%)
  - 1985-1989: 10/21807 (0.05%) 9/1938 (0.5%)
  - 1990-1994: 19/38712 (0.05%) 14/1306 (0.5%)
  - 1995-1999: 41/52060 (0.07%) 34/3594 (0.9%)
  - 2000-2004: 71/58537 (0.12%) 47/3859 (1.2%)

- $\chi^2_{\text{trend, df}=1} P < .0001$
Where exactly?

- low impact/specialty: 51
- BMJ: 2
- Archives Int Med: 2
- JAMA: 6
- Lancet: 27
- NEJM: 55
Characteristics

• Area of study
  – Cardiology 36
  – Cancer 30
  – HIV/AIDS 17
  – Critical care 10
  – Other 50

• Pharmacology 119/143

• 64 industry funded, 26 not reported, 53 not for profit
STOP-IT 1

- 76 of 143 did not report 1 or more of
  - planned sample size
  - interim analysis after which trial stopped
  - whether stopping rule informed decision

- median 66 events (IQ range 23-200)

- for 124 RCTs dichotomous outcomes
  - median RR 0.53 (IQ range 0.30-0.66)
  - fewer events larger treatment effects
    - OR 31, CI 12-82
Conclusions STOP-IT 1

- Epidemic of early stopped trials
- Large number in top journals
  - NEJM and Lancet big offenders
- Often methodologically flawed
- Majority events < 100
- Majority implausibly large effects
  - fewer the events, greater the effect
- Watch out for RCTs stopped early
  - high level of scepticism if large effects, few events
What about systematic reviews?

- stopped early trials overestimate
- Simulations with current stopping rules
  - overestimate minimal
  - true effect large, tRCTs small overestimate
  - true effect small, little weight in MA
tRCTs may still be a problem

- true effect modest
- appropriate stopping rules not applied
- publication bias
  - correcting RCTs never published
- tRCTs early in sequence of trials
  - correcting RCTs not yet done
  - correcting RCTs published late
    - Sterne and Simes, BMJ 1997, > 3 years difference
    - freezing/stifling effect
      - correcting RCTs never done at all
Remaining Questions

- Is there an overestimate & average size?

- What factors are associated with overestimate?

- Do meta-analyses including tRCTs suggest large or modest effects?

- Do meta-analyses with tRCTs suggest that conditions of overestimate exist (tRCTs high weight)?
Systematic Review of Meta-analyses including tRCTs

STOP-IT 2

Bassler D, Briel M, et al. JAMA 2010;303(12):1180-1187
Study design STOP-IT 2

- Obtain all tRCTs up to 2007
- Obtain meta-analyses in 2008
  - 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
Analysis STOP-IT 2

- **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: hierarchical 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
STOP-IT 2
Trial Flow

- 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)
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><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><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><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 \( p = 0.32 \)
Degrees of overestimates by number of events

![Graph showing the relationship between total tRCT event number and ratio of relative risk. The x-axis represents the total tRCT event number, ranging from 1 to 10,000. The y-axis represents the ratio of relative risk, ranging from 0 to 2.5. The graph displays a scatter plot of data points, with a ratio of RR = 0.71 indicated.](image)
Key Results STOP-IT 2

- Most effects small to moderate
- Large difference in effect (ratio of RR 0.71)
- Weight of tRCTs considerable
  - median weight 28%
    - interquartile range 12% to 40%
- 43 meta-analyses tRCTs < 100 events
  - 54% of > 20% weight
Cochrane Bias Methods Group

Discussion Meeting:

„Can early stopped trials result in misleading results of systematic reviews?“

Discussants:  Gordon Guyatt
              Steve Goodman
True Large Effect

Hypothetical

Empirical

Large effect

× difference from tRCTs to non-tRCTs

✓ Weight of tRCTs large
True Small Effect

Hypothetical

Empirical

Large effect difference from tRCTs to non-tRCTs

Weight of tRCTs large
True Small Effect with Freezing

Hypothetical

Empirical

Large effect difference from tRCT to non-tRCT

Weight of tRCTs large
Conclusions

• Most effects small or moderate, trials stopped early for benefit overestimate

• Small number events ➔ Overestimates large

• Clinicians, policymakers, methodologists cannot rely stopped early trials for effect estimates
Conclusion for Systematic Reviews

- If stopped early trials
  - alert to pooled result overestimate

- 3 Conditions coexist: Big Overestimate
  - tRCT small number of events (< 200)
  - big difference in RR tRCT and non-tRCTs
    - ratio of RRs < 0.70
  - tRCT has substantial weight in MA (> 20%)
What next?

• Potential correction of overestimates in tRCTs using Bayesian methods with conservative priors (Paul Glasziou, Rafael Perera)

• Examine RCTs that are launched after the publication of a tRCT asking the same research question - current practice of stopping RCTs sufficiently conservative?
Thank you very much for your attention!