

REVIEW ARTICLE

# Systematic reviewers neglect bias that results from trials stopped early for benefit

Dirk Bassler<sup>a,b</sup>, Ignacio Ferreira-Gonzalez<sup>a,c,d</sup>, Matthias Briel<sup>a,e</sup>, Deborah J. Cook<sup>a,f</sup>, P.J. Devereaux<sup>a,f</sup>, Diane Heels-Ansdell<sup>a</sup>, Haresh Kirpalani<sup>a,g</sup>, Maureen O. Meade<sup>a,f</sup>, Victor M. Montori<sup>h</sup>, Anna Rozenberg<sup>i</sup>, Holger J. Schünemann<sup>a,j,k</sup>, Gordon H. Guyatt<sup>a,f,\*</sup>

<sup>a</sup>The Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

<sup>b</sup>The Department of Neonatology, University Children's Hospital, Tübingen, Germany

<sup>c</sup>The Department of Cardiology, Epidemiology Unit, Vall d'Hebron Hospital, Barcelona, Spain

<sup>d</sup>The Department of Internal Medicine, Universitat Autònoma de Barcelona, Barcelona, Spain

<sup>e</sup>Basel Institute for Clinical Epidemiology, University Hospital Basel, Basel, Switzerland

<sup>f</sup>The Department of Medicine, McMaster University, Hamilton, Ontario, Canada

<sup>g</sup>The Department of Pediatrics, McMaster University, Hamilton, Ontario, Canada

<sup>h</sup>Knowledge and Encounter Unit, Mayo Clinic College of Medicine, Rochester, MN, USA

<sup>i</sup>The Department of Anesthesia, McMaster University, Hamilton, Ontario, Canada

<sup>j</sup>The Department of Medicine, University at Buffalo, NY, USA

<sup>k</sup>The Clinical Research Development and INFORMATION Translation Unit, Italian National Cancer Institute Regina Elena, Rome, Italy

Accepted 3 December 2006

## Abstract

**Objective:** To examine how authors of systematic reviews that include randomized clinical trials (RCTs) that are stopped early for benefit (truncated RCTs—tRCTs) address the potential for overestimation of treatment effects and to determine the weight of the tRCTs on pooled results.

**Study Design and Setting:** We searched the Cochrane Library and MEDLINE and evaluated systematic reviews that include at least one tRCT. We documented approaches that authors used to address potential overestimates of treatment effect introduced by including tRCTs. We assessed the impact of tRCTs in meta-analyses on the outcomes that led to their early termination.

**Results:** Of 96 systematic reviews that included at least one tRCT, 44 (46%) included > 1 tRCT, 68 (71%) did not mention truncation at all, and 2 (2%) documented early stopping for benefit as a criterion for methodological quality. Of 47 meta-analyses in which authors reported, or we could calculate the contribution of the tRCTs to the pooled result, the tRCTs contributed more than 40% of the weight in 16/47 (34%).

**Conclusion:** Most systematic reviews and meta-analyses including tRCTs fail to consider the possible overestimates of effect that may result from early stopping for benefit. We recommend safeguards that address this possibility. © 2007 Elsevier Inc. All rights reserved.

**Keywords:** Clinical trial methodology; Systematic reviews and meta-analysis; Bias; Stopping rules; Truncation of studies; Heterogeneity in meta-analysis

## 1. Background

As randomized clinical trials (RCTs) accrue patients and measure outcomes, interim results sometimes suggest large treatment effects that appear unlikely to be due to chance. Consequently, the investigators may conclude that one

treatment is superior to the other and stop the trial before reaching the target sample size.

Unfortunately, though generating low *P*-values, these apparent effects may be due to chance. Random fluctuations will occur early in RCTs, and premature termination on the basis of these results risks attribution of these random fluctuations to true underlying treatment effects [1–3]. Indeed, a systematic review suggested that treatment effects observed in such truncated RCTs (tRCTs) are often implausibly large, particularly when the number of events is small [4].

The prevalence of tRCTs is increasing [4]. Because of their large treatment effects, tRCTs are often published in high profile medical journals and receive considerable

\* Corresponding author. Department of Clinical Epidemiology and Biostatistics, Health Sciences Centre Room 2C12, McMaster University, 1200 Main Street West, Hamilton, ON L8N 3Z5, Canada. Tel.: 905-525-9140 ext. 22900; fax: 905-524-3841.

E-mail address: [guyatt@mcmaster.ca](mailto:guyatt@mcmaster.ca) (G.H. Guyatt).

public attention [4]. It is therefore unlikely that systematic reviews will fail to identify tRCTs; indeed tRCTs may trigger systematic reviews. If reviewers fail to note truncation, and do not consider early stopping for benefit as a source of potential overestimation of treatment effects, meta-analyses may report spuriously large treatment effects [5]. We therefore conducted a systematic review of meta-analyses that included tRCTs to determine how investigators addressed the potential overestimation associated with truncation, and to determine the impact of these studies on pooled results of meta-analyses.

## 2. Methods

### 2.1. Eligibility criteria

We endeavored to identify all systematic reviews that included a tRCT. We included reviews that described a population, experimental intervention, and control intervention similar to a previously identified tRCT. Additionally, reviews included a methods section and a literature search that included Medline.

### 2.2. Literature search

We began with a repository of 143 tRCTs that we had previously identified by an electronic literature search in MEDLINE, EMBASE, Current Contents, and full-text journal content databases up to November 2004 [4]. We updated this database through a hand search in Pediatrics from 1990 to 2004 and through personal contacts with colleagues who alerted us to tRCTs that they identified after the publication of the systematic review in 2005 [4].

To capture relevant systematic reviews, we performed a computerized search of the Cochrane Database of Systematic Reviews, the Database of Abstracts of Reviews of Effects (both contained in the Cochrane Library 2005, Issue 3), and MEDLINE in September 2005. We used text words and MESH terms based on the study population and the intervention of each of the identified tRCTs, using a validated filter for systematic reviews [6] without language restrictions. We contacted experts in the relevant fields to identify further systematic reviews.

### 2.3. Study selection

Two investigators (D.B. and M.B.) screened the abstracts and obtained the full-text reports of potentially eligible systematic reviews. After obtaining full reports of all candidate reviews, the same investigators assessed eligibility. When necessary, we contacted review authors for further information.

### 2.4. Data extraction

Four investigators (D.B., I.F.G., M.B., and A.R.) extracted relevant data independently and in duplicate, using

a standardized form. Reviewers recorded the year of publication, content area, type of comparisons, total number of included RCTs and the number of tRCTs per review, features related to the methodological quality of RCTs, and whether the Cochrane Collaboration had conducted the reviews. Reviewers identified the systematic reviews that included the tRCTs in any meta-analyses that addressed the outcome that led to the early termination, and the subset of these meta-analyses that the authors had designated as the primary analyses. If the authors of the systematic reviews did not identify a primary analysis, we considered the meta-analysis of the outcome that was first reported in the results section in the abstract as the primary meta-analysis. We included meta-analyses that focused on a component endpoint of a composite outcome that led to truncation and those in which the outcome was a subcategory (e.g., cause specific mortality) of the outcome that led to truncation (e.g., overall mortality).

Reviewers noted the model (e.g., random effects or fixed effects) used in the meta-analyses for the outcome that led to the early termination, the weight attributed to the truncated trials in these meta-analyses, and the approaches of authors to deal with the stopped-early trials.

## 3. Statistical analysis

We described categorical variables with frequencies and percentages and continuous nonnormal data with medians and interquartile ranges. We performed an overall analysis and also analyzed the data separately by the field of study to explore a possible cluster effect.

For reports that failed to calculate the weight of tRCTs in the relevant meta-analyses, we performed our own meta-analysis using the same model specified in the systematic review. If authors did not mention any specific model, we replicated the meta-analyses using either a fixed or random effects model and compared our results with those reported in the systematic review, selecting the model that best matched the results reported. We used the algorithm in Rev Man version 4.2 to calculate the weights (Mantel–Haenszel method under the assumption of fixed effects and DerSimonian and Laird method under the assumption of random effects). We used SPSS version 11.0 (SPSS Inc, Chicago, IL) and Rev Man version 4.2 for all data analyses.

## 4. Results

### 4.1. Truncated RCTs

Hand searching and personal contacts yielded 16 tRCTs in addition to the 143 identified in the initial electronic search [4] (an online list of these trials is available). We

searched for systematic reviews corresponding to the 159 tRCTs.

#### 4.2. Selected systematic reviews

Our literature search identified 146 potentially eligible systematic reviews (Fig. 1). Of these reviews, 50 did not include a tRCT (in 46 because the tRCT was published after the systematic review terminated its search and in four because the review apparently failed to identify the tRCT). Table 1 describes the characteristics of the 96 eligible systematic reviews that include at least one tRCT (an online list of these systematic reviews is available). Forty-five of the 96 reviews investigated cardiology interventions (18 of which investigated the effects of statins); review characteristics did not differ by area of study (data available from authors). Most systematic reviews compared pharmacological interventions with placebo. The Cochrane Collaboration published 32% of the reviews. The median number of RCTs included per systematic review was 11.5; 46% (44/96) included more than one tRCT.

#### 4.3. Assessment of methodological quality and approaches to deal with truncation

Of the 96 systematic reviews, 36 (37%) did not address the methodological quality of the included studies, 49 (51%) assessed quality using individual components, seven reported quality assessment by score, and four reported both individual components and an overall score. Neither the Jadad scale [7] used by the majority of reviewers who calculated a quality score nor any of the other scoring systems used address truncation. Two reviews included truncation as an individual criterion for the assessment of methodological quality.

Sixty-eight of the 96 systematic reviews (71%) failed to mention that an included trial was stopped early for benefit (Table 1); this was true for 53 of the 72 (74%) systematic reviews that conducted a meta-analysis including the outcome that led to early termination of a truncated trial (Table 2). Of the total of 28 reviews that did identify the tRCTs, seven acknowledged that including a truncated RCT might introduce bias. Two of these seven reviews included subgroup analyses excluding the tRCT.

#### 4.4. Meta-analyses including studies stopped early

Of the 96 identified systematic reviews, 72 included at least one meta-analysis for the outcome that led to early termination of a truncated study, of which 57 (79%) were either reported or considered as primary analysis (Table 2). In 24 (33%) of 72 reviews, investigators analyzed the data using only a fixed effects model (Table 2). The median number of RCTs included in the 57 primary meta-analyses was eight. Of the 57, 27 reported the weight of each RCT contributing to the pooled results and we calculated the weight in an additional 20 primary meta-analyses. We failed to calculate the weight in 10 primary meta-analyses because the authors did not report the results of the individual studies in a manner allowing our replicating their analysis ( $n = 6$ ), they were individual-patient meta-analysis ( $n = 3$ ), or they used a Bayesian approach ( $n = 1$ ). Of these 47 primary meta-analyses, the weight of the stopped-early studies was more than 40% in 16 (34%) meta-analyses (Table 2). The weight of the stopped-early studies was available in 142 secondary meta-analyses and was more than 40% in 70 (49%). Of all the meta-analyses (primary and secondary) conducted using the outcome that led to truncation, the weight was available in 189. The weight of the tRCTs was more than 40% in 88 of these 189 meta-analyses (47%).

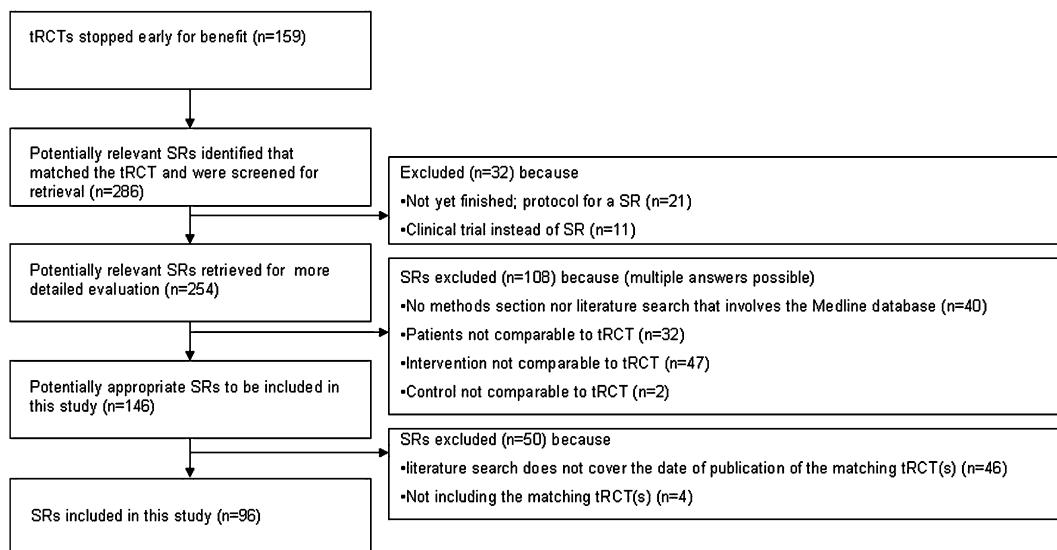


Fig. 1. Screening process for SRs. Abbreviation: SR, Systematic review.

Table 1  
Characteristics of SRs that include at least 1 randomized clinical trial stopped early for benefit ( $N = 96$ )

Characteristic	No./Total (%)	Median (IQR)
Year of publication		
1992–1994	1/96 (1)	
1995–1997	5/96 (5.2)	
1998–2000	21/96 (21.9)	
2001–2003	41/96 (42.7)	
2004–01/2006	28/96 (29.2)	
Area of study		
Cardiology (nonstatins)	27/96 (28.1)	
Cardiology (statins)	18/96 (18.8)	
HIV/AIDS	9/96 (9.4)	
Hematology–oncology	7/96 (7.3)	
Neurology	7/96 (7.3)	
Critical care	6/96 (6.3)	
Pediatrics	5/96 (5.2)	
Other <sup>a</sup>	17/96 (17.2)	
Type of comparisons		
Drug vs. placebo	25/96 (26)	
Drug vs. standard therapy <sup>b</sup>	15/96 (15.6)	
Drug vs. placebo and drug vs. standard therapy <sup>b</sup>	45/96 (46.8)	
Nonpharmacological intervention vs. standard therapy <sup>b</sup>	11/96 (11.3)	
Included studies per SR		11.5 (5.25–25)
SRs from the Cochrane Collaboration	31/96 (32.3)	
SRs not mentioning truncation of tRCT	68/96 (70.8)	
Methodological quality assessment of studies		
None	36/96 (37.5)	
Individual components	49/96 (51)	
Score	7/96 (7.3)	
Individual components and score	4/96 (4.2)	
Individual components <sup>c</sup>		
Adequate randomization	44/96 (45.8)	
Allocation concealment	43/96 (44.8)	
Blinding	48/96 (50)	
Follow-up	42/96 (43.7)	
Truncation of study	2/96 (2.1)	
SRs including > 1 tRCT	44/96 (45.8)	
Including two tRCTs	32/96 (33.2)	
Including three tRCTs	9/96 (9.3)	
Including four tRCTs	3/96 (3.1)	
SRs with primary MA for the outcome that led to early termination including the tRCT(s)	57/96 (59.4)	
SRs with MAs for the outcome that led to early termination including the tRCT(s)	72/96 (75)	
Number of MAs in SRs for the outcome that led to early termination including the tRCT(s)	299	

Abbreviations: HIV, Human Immunodeficiency Virus; IQR, Interquartile range; MA, Meta-analysis, SR, Systematic review.

<sup>a</sup> Includes obstetrics and gynecology, endocrinology, gastroenterology, pulmonology, rheumatology, nephrology, ophthalmology, and ear nose and throat disorders.

<sup>b</sup> Standard therapy may include other drugs.

<sup>c</sup> Several components may be reported per review.

## 5. Comment

### 5.1. Findings

Most of the 96 systematic reviews identified failed to mention that some of the included RCTs were stopped early for benefit, and only two reported truncation as a potential source of bias. Of those systematic reviews that included a tRCT in meta-analyses focusing on the outcome that led to truncation, the tRCTs carried more than 40% of the weight in 34% of the primary meta-analyses and in 49% of the secondary meta-analyses.

### 5.2. Strengths and limitations

To our knowledge, this is the first study highlighting deficiencies in the reporting of systematic reviews and meta-analyses, including RCTs stopped early for benefit, and demonstrating that tRCTs have a substantial impact on pooled results. Although the initial sample of tRCTs was compiled through a systematic electronic literature search [4] that has now been complemented through a hand search and personal contacts, tRCTs are difficult to identify. Truncation is often not mentioned in the abstract and may not even be explicitly labeled in the methods. Indeed, a review of data monitoring committee use and practice in RCTs suggests that authors may not identify that their trials are stopped early for benefit [8]. Thus, there may be relevant tRCTs that we failed to identify.

### 5.3. Implications

A previous simulation study has explored the extent to which application of statistical stopping rules in clinical trials can explain heterogeneity of treatment effects in meta-analyses of related trials [5]. Despite a growing awareness that stopping trials early because of apparent benefit may inflate treatment effects [1–4], authors of systematic reviews have paid little attention to the potential bias introduced through inclusion of tRCTs in meta-analyses. Most of the systematic reviews we identified did not even mention that they included one or more tRCTs. Including tRCTs in systematic reviews and meta-analyses risks overestimating treatment effects; failure to identify the possible overestimate creates an even greater risk of spurious findings ultimately influencing clinical practice.

We recommend, that authors of systematic reviews and meta-analyses report truncation of studies and explore truncation as an explanation for heterogeneity in their results. The findings of our study have immediate implications for authors of systematic reviews, the Cochrane Collaboration among them, as well as for the recommendations on how to report the results of meta-analyses, summarized in the QUOROM [9] guidelines.

Table 2

Characteristics of SRs including the tRCTs in MAs for the outcome that led to the early termination ( $N = 72$ )

Characteristic	No./total (%)		
<b>Model for MAs</b>			
Fixed effects model	24/72 (33.3)		
Random effects model	18/72 (25)		
Fixed and random effects model	13/72 (18)		
Other model	9/72 (12.5)		
Model not specified	8/72 (11.1)		
<b>Approach to deal with tRCT(s) in MAs</b>			
Truncation not mentioned at all	53/72 (73.6)		
Truncation mentioned without specification that this may create bias	12/72 (16.6)		
Truncation mentioned and chance of bias acknowledged, no subgroup analysis	5/72 (6.9)		
Truncation mentioned and subgroup analysis performed with and without tRCT(s)	2/72 (2.8)		
	Primary MAs	Secondary MAs	All MAs
MAs in SRs ( $N = 72$ ) for the outcome that led to early termination including the tRCT(s)	57	242	299
MAs that report the number of included RCTs: No./total (%)	57/57 (100)	237/242 (98)	294/299 (98.6)
RCTs included in MA: median (IQR)	8 (4–12)	5 (3–7)	5 (3–8)
MAs that report the weight of individual RCTs: No./total (%)	27/57 (47.3)	142/242 (58.5)	169/299 (56)
Weight of tRCT(s) in MAs <sup>a</sup>	$N = 47$	$N = 142$	$N/189$
< 20%	19/47 (40.4)	26/142 (18.4)	45/189 (23.8)
20%–40%	12/47 (25.5)	46/142 (32.4)	58/189 (30.7)
> 40	16/47 (34.1)	70/142 (49)	88/189 (46.5)

Abbreviations: IQR, Interquartile range; MA, Meta-analysis; No., Number; RCT, Randomized clinical trial; SR, Systematic review.

<sup>a</sup> We performed calculations for the weight in 20 primary meta-analyses.

## Acknowledgments

Dr. Ferreira is supported by Carlos III Spanish Institute of Health Research Fellowship Award (FIS). Dr. Briel is supported by the Swiss National Science Foundation. Dr. Montori is a Mayo Foundation Scholar. Dr. P.J. Devereaux is supported by a Canadian Institutes of Health Research, New Investigator Award. Dr. Cook is a Research Chair of the Canadian Institutes for Health Research.

## References

- [1] Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet* 2005;365:1657–61.
- [2] Pocock S, White I. Trials stopped early: too good to be true? *Lancet* 1999;353:943–4.
- [3] Pocock SJ. When (not) to stop a clinical trial for benefit. *JAMA* 2005;294:2228–30.
- [4] Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA* 2005;294:2203–9.
- [5] Hughes MD, Freedman LS, Pocock SJ. The impact of stopping rules on heterogeneity of results in overviews of clinical trials. *Biometrics* 1992;48:41–53.
- [6] Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. *BMJ* 2005;330(7482):68.
- [7] Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;17:1–12.
- [8] Sydes MR, Altman DG, Babiker AB, Parmar MK, Spiegelhalter DJ. Reported use of data monitoring committees in the main published reports of randomized controlled trials: a cross-sectional study. *Clin Trials* 2004;1:48–59.
- [9] Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. *Lancet* 1999;354:1896–900.