JAMA®

Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects: Systematic Review and Meta-regression Analysis

Dirk Bassler; Matthias Briel; Victor M. Montori; et al.

Supplementary material	eSupplement http://jama.ama-assn.org/cgi/content/full/303/12/1180/DC1
Correction	Contact me if this article is corrected.
Citations	This article has been cited 1 time. Contact me when this article is cited.
Topic collections	Quality of Care; Evidence-Based Medicine; Statistics and Research Methods; Review; Prognosis/ Outcomes Contact me when new articles are published in these topic areas.
Related Articles published in the same issue	Randomized Controlled Trials Janet M. Torpy et al. *JAMA*. 2010;303(12):1216.

Subscribe http://jama.com/subscribe	Email Alerts http://jamaarchives.com/alerts
Permissions permissions@ama-assn.org http://pubs.ama-assn.org/misc/permissions.dtl	Reprints/E-prints reprints@ama-assn.org

Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects
Systematic Review and Meta-regression Analysis

Dirk Bassler, MD, MSc

Matthias Briel, MD, MSc

Victor M. Montori, MD, MSc

Melanie Lane, BA

Paul Glasziou, MBBS, PhD

Qi Zhou, PhD

Diane Heels-Ansdell, MSc

Stephen D. Walter, PhD

Gordon H. Guyatt, MD, MSc

and the STOPIT-2 Study Group

ALTHOUGH RANDOMIZED CON-trolled trials (RCTs) gener-ally provide credible evi-dence of treatment effects, multiple problems may emerge when in-vestigators terminate a trial earlier than planned,[1] especially when the decision to terminate the trial is based on the find-ing of an apparently beneficial treat-ment effect. Bias may arise because large random fluctuations of the estimated treatment effect can occur, particularly early in the progress of a trial.[2] When in-vestigators stop a trial based on an ap-parently beneficial treatment effect, their results may therefore provide mislead-ing estimates of the benefit.[3,4] Statistical modeling suggests that RCTs stopped early for benefit (truncated RCTs) will systematically overestimate treatment ef-fects,[5] and empirical data demonstrate that truncated RCTs often show implau-sibly large treatment effects.[6]

Empirical evidence addressing the mag-nitude of bias from stopping early, and fac-tors that may influence the magnitude of the bias, remain limited and the appro-priate interpretation of truncated RCTs

See also Patient Page.

Context Theory and simulation suggest that randomized controlled trials (RCTs) stopped early for benefit (truncated RCTs) systematically overestimate treatment effects for the outcome that precipitated early stopping.

Objective To compare the treatment effect from truncated RCTs with that from meta-analyses of RCTs addressing the same question but not stopped early (nontruncated RCTs) and to explore factors associated with overestimates of effect.

Data Sources Search of MEDLINE, EMBASE, Current Contents, and full-text journal content databases to identify truncated RCTs up to January 2007; search of MEDLINE, Coch-rane Database of Systematic Reviews, and Database of Abstracts of Reviews of Effects to identify systematic reviews from which individual RCTs were extracted up to January 2008.

Study Selection Selected studies were RCTs reported as having stopped early for benefit and matching nontruncated RCTs from systematic reviews. Independent re-viewers with medical content expertise, working blinded to trial results, judged the eligibility of the nontruncated RCTs based on their similarity to the truncated RCTs.

Data Extraction Reviewers with methodological expertise conducted data extrac-tion independently.

Results The analysis included 91 truncated RCTs asking 63 different questions and 424 matching nontruncated RCTs. The pooled ratio of relative risks in truncated RCTs vs matching nontruncated RCTs was 0.71 (95% confidence interval, 0.65-0.77). This difference was independent of the presence of a statistical stopping rule and the meth-odological quality of the studies as assessed by allocation concealment and blinding. Large differences in treatment effect size between truncated and nontruncated RCTs (ratio of relative risks <0.75) occurred with truncated RCTs having fewer than 500 events. In 39 of the 63 questions (62%), the pooled effects of the nontruncated RCTs failed to demonstrate significant benefit.

Conclusions Truncated RCTs were associated with greater effect sizes than RCTs not stopped early. This difference was independent of the presence of statistical stop-ping rules and was greatest in smaller studies.

JAMA. 2010;303(12):1180-1187   www.jama.com

a matter of controversy.[6-11] We therefore undertook a systematic review to deter-mine the treatment effect from truncated RCTs compared with meta-analyses of RCTs addressing the same research ques-tion that were not stopped early (nontrun-cated RCTs) and to explore factors asso-ciated with overestimates of effect.

METHODS

A prior report provides a detailed de-scription of the design and methods of

Author Affiliations: Department of Clinical Epidemi-ology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (Drs Bassler, Briel, Zhou, Walter, and Guyatt and Ms Heels-Ansdell); Depart-ment of Neonatology, University Children's Hospital Tuebingen, Tuebingen, Germany (Dr Bassler); Basel Institute for Clinical Epidemiology and Biostatistics, Uni-versity Hospital Basel, Basel, Switzerland (Dr Briel); Knowledge and Encounter Research Unit, Mayo Clinic, Rochester, Minnesota (Dr Montori and Ms Lane); and Centre for Evidence-Based Medicine, Department of Primary Health Care, University of Oxford, Oxford, UK (Dr Glasziou).
Members of the STOPIT-2 Study Group and their author affiliations are listed at the end of this article.
Corresponding Author: Victor Montori, MD, MSc, Knowledge and Encounter Research Unit, Mayo Clinic, Plummer 3-35, 200 First St SW, Rochester, MN 55905 (montori.victor@mayo.edu).

this study (Study of Trial Policy of Interim Truncation-2 [STOPIT-2]).[12] In summary, we conducted extensive literature searches to identify truncated RCTs and systematic reviews addressing the same question. We retrieved all RCTs included in the systematic reviews, extracted data and conducted new meta-analyses of the nontruncated RCTs addressing the outcome that led to the early termination of the truncated RCTs, and compared the relative risk (RR) generated by the truncated RCTs with the RR from all matching nontruncated RCTs.

Literature Search

We updated the database from our prior study following the same search strategy.[6] In January 2007 we searched MEDLINE, EMBASE, Current Contents, and full-text journal content databases from their inception for truncated RCTs. In addition, we identified truncated RCTs through hand searching, by personal contact with trial investigators, and by a citation search linked to 2 key articles.[6,13] For systematic reviews, we searched the Cochrane Database of Systematic Reviews, the Database of Abstracts of Reviews of Effects, and MEDLINE from their inception to January 2008.

Eligibility Criteria for Truncated RCTs and Matching Systematic Reviews

We included RCTs of any intervention reported as having stopped earlier than initially planned owing to interim results in favor of the intervention.

We excluded matching systematic reviews that did not have a methods section and did not describe a literature search that, at minimum, included MEDLINE.[12]

Identification, Retrieval, and Eligibility of Nontruncated RCTs

We retrieved the full text of all RCTs included in each systematic review. If a systematic review was published prior to the matching truncated RCT and thus did not include the truncated RCT, we updated this review.[12] Eligible nontruncated RCTs addressed the outcome that led to the early termination of the truncated RCT and stated clearly that allo-

cation was randomized. We assessed the eligibility of nontruncated RCTs based on the similarity of the question addressed by the matching truncated RCT (see Briel et al[12] for details).

Teams of 2 reviewers with relevant clinical expertise made independent eligibility and similarity decisions and resolved disagreement by discussion and, if necessary, by consulting a third party. Reviewers who judged eligibility were blinded to the results of the trials through electronic or manual masking.[12]

Data Extraction and Analysis

Working in pairs, reviewers with methodological expertise conducted data extraction independently.[12] From each RCT (truncated or nontruncated), we collected information about early termination, the journal of publication (we categorized *Annals of Internal Medicine*, *BMJ*, *JAMA*, *Lancet*, and *New England Journal of Medicine* as high-impact journals), the year of publication, methodological quality, data monitoring committees, stopping rules at the outset of the trial, interim analyses, and the measure of treatment effect for the outcome that terminated the truncated RCT. The only study characteristic tested for a statistically significant difference between truncated and nontruncated RCTs was publication in a high-impact journal.

We calculated an RR for each RCT in our study. For studies that provided results using continuous data, we estimated an approximate dichotomous equivalent.[12,14] For each question, we used meta-analysis for the pooled RR and a 95% confidence interval (CI) for all nontruncated RCTs. If more than 1 truncated RCT addressed the same question, we calculated a pooled RR and CI for those truncated RCTs. Pooled estimates of RRs were calculated using an inverse-variance weighted random-effects model.

We performed a z test for each meta-analysis to assess differences between the truncated and nontruncated RCTs with respect to their pooled RRs. As a summary measure we calculated a ratio of RRs, and its logarithm, for each question as follows:

Log[ratio of RRs]=log[RR of truncated RCT(s)/pooled RR of nontruncated

RCTs]=log[RR of truncated RCT(s)]−log[pooled RR of nontruncated RCTs].

We estimated the pooled log[ratio of RRs] using a random-effects inverse-variance meta-analysis and then, for purposes of presentation, back transformed to the overall ratio of RRs. To explore factors associated with the magnitude of the ratio of RRs, we performed a meta-regression analysis in which the dependent variable was the log[ratio of RRs] and independent variables were whether the truncated RCTs used a formal stopping rule and the number of outcome events in the truncated RCTs. When more than 1 truncated RCT addressed the same question, the stopping rule status was assigned to "has a rule" if at least 1 truncated RCT had a rule. Similarly, when there was more than 1 truncated RCT for the same question, we used the truncated RCT with the largest number of events as the source for our analyses of the influence of the number of events.

To allow consideration of methodological quality as a predictor and to test whether restriction to nontruncated RCTs that are more similar to the truncated RCTs would change the results, we constructed a second meta-regression described fully in a prior report.[12] In brief, this meta-regression used a hierarchical model with 2 levels: individual RCT (study) level and meta-analysis (question) level. The dependent variable in this analysis was the logarithm of the RR for each study. Predictor variables considered included a combined group variable (truncated RCT with a rule, truncated RCT without a rule, nontruncated RCT), number of events, concealment of allocation, use of blinding, and the interaction between the group variable and the other variables. We performed this meta-regression on different data sets based on various thresholds for the similarity of the nontruncated RCTs in each question to the matching truncated RCTs.

To test for an order effect (the hypothesis being that studies published earlier will have more responsive populations), for each review question we established where in the sequence of published studies (by publication date) the

truncated RCT stood and referred to this as the "rank" of the truncated RCT. We then calculated a "standardized rank" $[100 \cdot (\text{rank} - 1)/(\text{total number of studies} - 1)]$. If there was more than 1 truncated RCT in the review question, we used the median among the truncated RCTs as the standardized rank of truncated RCTs. We then divided the review questions into 2 groups (those with the standardized rank of the truncated RCT equal to or less than 50 [n=27] and those with the standardized rank of the truncated RCT greater than 50 [n=37]) and repeated the meta-analysis for each group.

As a secondary analysis, we compared the RR of the truncated RCT(s) with the pooled estimate for all trials including the truncated RCTs.

Analyses were performed using SAS version 9.2 (SAS Institute Inc, Cary, North Carolina); tests were 2-sided, and $P < .05$ was used as the threshold for statistical significance.

RESULTS

Literature Search

A total of 195 truncated RCTs formed the basis for the search for systematic reviews; we identified matching systematic reviews for 79 questions. We extracted 2488 nontruncated RCTs from 202 matching systematic reviews (of which 32 were updated). Of these 2488 studies, 22 (0.9%) proved to be truncated RCTs, which we added to the truncated RCT database. We excluded 2012 nontruncated RCTs based on insufficient similarity to the truncated RCTs or unclear randomization and 30

because the RR could not be calculated. The remaining 424 nontruncated RCTs and 91 matching truncated RCTs addressed 63 questions (**FIGURE 1**). An eSupplement reporting the references of the included studies is available at http://www.jama.com.

Study Characteristics

TABLE 1 describes the characteristics of the eligible studies. Compared with matching nontruncated RCTs, truncated RCTs were more likely to be published in high-impact journals (30% vs 68%, $P < .001$).

Quantification of Differences in Treatment Effect Size

Of 63 comparisons, the ratio of RRs was equal to or less than 1.0 in 55 (87%); the

Figure 1. Selection Process for Study Inclusion
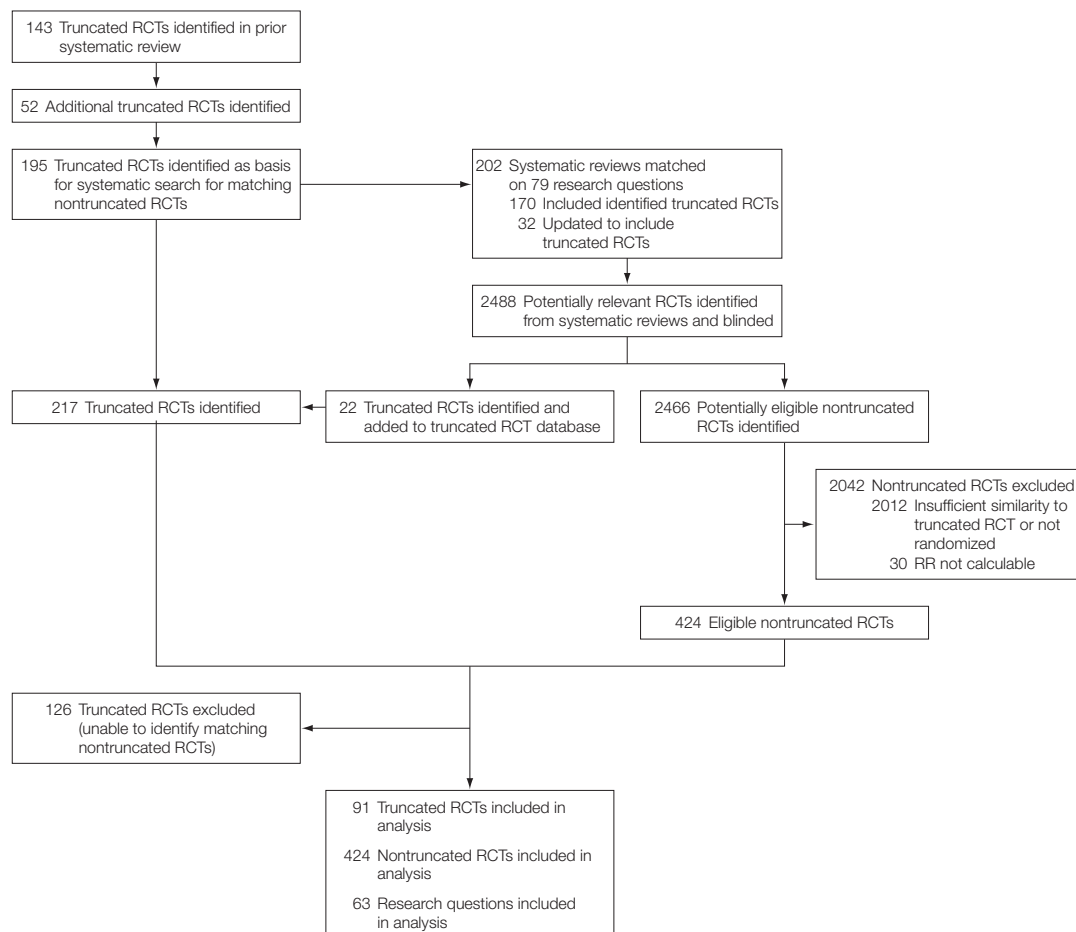
RCT indicates randomized controlled trial; RR, relative risk.

weighted average ratio of RRs was 0.71 (95% CI, 0.65-0.77; $P < .001$) (**FIGURE 2**). In 39 of 63 comparisons (62%), the pooled estimates for nontruncated RCTs were not statistically significant.

Comparison of the truncated RCTs with all RCTs (including the truncated RCTs) demonstrated a weighted average ratio of RRs of 0.85; in 16 of 63 comparisons (25%), the pooled estimate failed to demonstrate a significant effect.

Determinants of Differences in Treatment Effect Size

TABLE 2 summarizes the findings from the single-level meta-regression analysis to determine predictors of differences in the treatment effect size between truncated and nontruncated RCTs. In the univariable models, both the number of events ($P < .001$) and the presence of a statistical stopping rule ($P = .02$) were significant. When we included both variables in the model, only the number of events remained significant ($P < .001$). The results from the multilevel meta-regression confirmed significant interactions between the combined variable (truncated vs nontruncated RCT) and the number of events ($P < .001$). Large differences in treatment effect size between truncated and nontruncated RCTs (ratio of RRs < 0.75) occurred in truncated RCTs with fewer than 500 events (**FIGURE 3**).

The multilevel meta-regression analysis using the entire data set demonstrated that neither concealment of allocation ($P = .96$) nor blinding ($P = .32$) were significant predictors of the differences in treatment effect size.

Different Data Sets and Order of Publication

The findings were similar, irrespective of either the closeness of the match between nontruncated and truncated RCTs or the order of publication of the truncated RCTs relative to that of matching nontruncated RCTs. In the multilevel meta-regression analysis, adjusted ratios of RRs of truncated vs nontruncated RCTs were 0.64 when questions were very closely matched, 0.70 when they were moderately close, and 0.69

Table 1. Characteristics of Randomized Controlled Trials Stopped Early for Benefit and Those Not Stopped Early for Benefit Asking the Same Research Question

Characteristic	No. (%)	
	Stopped Early (n = 91)	Not Stopped Early (n = 424)
Year of publication		
Before 1975	2 (2)	9 (2)
1975-1984	1 (1)	29 (7)
1985-1994	27 (30)	141 (33)
1995-2004	50 (55)	230 (54)
2005-2008	11 (12)	15 (4)
Published in high-impact journal[a]	62 (68)	128 (30)
Area of study		
Cardiology	32 (35)	210 (49)
Hematology-oncology	10 (11)	24 (6)
HIV/AIDS	14 (15)	34 (8)
Critical care	8 (9)	38 (9)
Neurology	1 (1)	8 (2)
Other	26 (29)	110 (26)
Type of comparisons		
Active medication vs placebo	56 (62)	277 (65)
Active medication vs active medication	12 (13)	49 (12)
Nonpharmacological therapeutic interventions	15 (16)	46 (11)
Drug vs nonpharmacological therapeutic intervention	7 (8)	46 (11)
Nontherapeutic interventions	1 (1)	6 (1)
Quality of reporting of safeguards against bias		
Adequate randomization methods	61 (67)	195 (46)
Adequate allocation concealment	48 (53)	145 (34)
Blinding		
Patients	55 (60)	255 (60)
Care providers	40 (44)	225 (53)
Data collectors	25 (27)	146 (34)
Judicial assessors of outcomes	32 (35)	148 (35)
Data analysts	8 (9)	49 (12)
Reported planned sample size	71 (78)	220 (52)
Reported actual sample size	68 (75)	220 (52)
Funding		
For-profit organization	38 (42)	151 (35)
Not-for-profit organization/government funding	33 (36)	92 (22)
Other	1 (1)	29 (7)
Not reported	19 (21)	152 (36)
Authors' conflict of interest		
Reported	36 (40)	59 (14)
Conflict of interest	31 (34)	53 (13)
Employment at funding agency	21 (23)	42 (10)
Reported "no conflict of interest"	5 (6)	6 (1)
Not reported	55 (60)	365 (86)
DMC		
Existent	64 (70)	183 (43)
Not existent	5 (6)	28 (7)
Not mentioned	22 (24)	213 (50)
Composition of the DMC reported		
Yes	40 (62)	111 (61)
Names of members	33 (36)	40 (9)
Member affiliations	5 (5)	15 (3)
Member expertise	4 (4)	6 (1)
Sponsor representative a member		
Yes	3 (3)	17 (4)
No	21 (33)	38 (21)
Not mentioned	3 (5)	34 (18)

(continued)

Table 1. Characteristics of Randomized Controlled Trials Stopped Early for Benefit and Those Not Stopped Early for Benefit Asking the Same Research Question (continued)

Characteristic	No. (%)	
	Stopped Early (n = 91)	Not Stopped Early (n = 424)
Conflict of interest reported for DMC members		
Yes	1 (2)	3 (2)
No	0	3 (2)
Not mentioned	63 (98)	177 (96)
Blinding of DMC		
Yes	6 (9)	38 (21)
No	9 (14)	55 (30)
Not mentioned	49 (77)	90 (49)
Preplanned stopping rule		
Yes	69 (76)	55 (13)
Haybittle-Peto	8 (11)	2 (4)
Pocock	6 (9)	2 (4)
O'Brien-Fleming	22 (32)	5 (9)
Lan-DeMets	4 (6)	9 (16)
Other	20 (29)	37 (67)
Not specified	9 (13)	0
No	22 (24)	369 (87)
No. of planned interim analyses		
1	15 (16)	17 (4)
2	11 (12)	7 (2)
3	6 (7)	11 (3)
>3	12 (13)	10 (2)
Not reported	47 (52)	379 (89)
Definition of interim period		
By periods	10 (23)	14 (31)
By No. of recruited patients	17 (39)	13 (29)
By No. of events	2 (4)	2 (4)
By other means	10 (23)	7 (16)
Not reported	5 (11)	9 (20)

Abbreviations: DMC, data monitoring committee; HIV, human immunodeficiency virus.
[a] *Annals of Internal Medicine*, *BMJ*, *JAMA*, *Lancet*, *New England Journal of Medicine*.

when they were least close. The ratio of RRs of the group in which the truncated RCTs were published in early years (standardized rank ≤50) was 0.74 (95% CI, 0.66-0.83) and for the later years (standardized rank >50) was 0.68 (95% CI, 0.60-0.77). The *P* value for the difference between the 2 estimates was 0.33.

COMMENT
Summary of Findings

In this empirical study including 91 truncated RCTs and 424 matching nontruncated RCTs addressing 63 questions, we found that truncated RCTs provide biased estimates of effects on the outcome that precipitated early stopping. On average, the ratio of RRs in the truncated RCTs and matching nontruncated RCTs was 0.71. This implies that, for instance, if the RR from the nontruncated RCTs was 0.8 (a 20% relative risk reduction), the RR from the truncated RCTs would be on average approximately 0.57 (a 43% relative risk reduction, more than double the estimate of benefit). Nontruncated RCTs with no evidence of benefit—ie, with an RR of 1.0— would on average be associated with a 29% relative risk reduction in truncated RCTs addressing the same question.

In nearly two-thirds of comparisons, the pooled estimate for nontruncated RCTs failed to demonstrate a statistically significant effect. We found substantial heterogeneity in our analysis of the pooled ratio of RRs for truncated vs nontruncated RCTs, suggesting that differences between truncated and nontruncated RCT effects will differ across study questions. This heterogeneity could be partially explained by the total num-

ber of outcome events in the truncated RCTs, with larger differences between truncated and nontruncated RCTs in studies with a smaller number of events.

The methodological quality and the presence of a statistical stopping rule failed to predict the observed difference in the treatment effect.

Strengths and Limitations

We used rigorous search strategies and undertook an intensive independent evaluation of eligibility and similarity of several thousand RCTs blinded to the results. Our analysis had considerable statistical power to link the estimates of treatment effect from truncated and nontruncated RCTs addressing the same question and demonstrated consistent results across degrees of similarity of the question addressed by the truncated RCTs and the matching nontruncated RCTs.

Our literature search, while extensive, missed some truncated RCTs. Assessment of the 2488 RCTs included in the systematic reviews revealed 22 additional truncated RCTs not initially identified. Whether results would differ in other unidentified truncated RCTs remains speculative.

We relied on systematic reviews to identify nontruncated RCTs but did not assess the reviews' susceptibility to publication bias. However, we know that trials with positive findings have nearly 4 times the odds of being published compared with those with negative findings.[15] To the extent that publication bias is present, inclusion of unpublished studies would lead to a diminished pooled effect from the nontruncated RCTs. This would in turn likely lead to a larger gradient of effect between truncated and nontruncated RCTs. Thus, to the extent that publication bias exists, our results probably represent a conservative estimate of the exaggeration in treatment benefit associated with stopping early.

Relation to Recent Empirical Studies, Simulation, and Commentaries

Korn and colleagues recently reviewed the results of cancer trials stopped early and that either continued with further follow-up or released results early.[16] They

Figure 2. Pooled Ratio of Relative Risks (RRs) and 95% Confidence Intervals (CIs) for Truncated vs Nontruncated Randomized Controlled Trials (RCTs)

Matching Question No.	Truncated RCTs				Nontruncated RCTs						
	No.	Total Events	Sample Size	RR (95% CI)	No.	Total Events	Sample Size	RR (95% CI)	P Value	Ratio of RRs (95% CI)	Favors Truncated RCT / Favors Nontruncated RCT
1	1	15	38	0.36 (0.14-0.94)	24	918	3407	0.51 (0.40-0.64)	<.001	0.72 (0.27-1.92)	
2	1	1477	9297	0.79 (0.72-0.88)	4	1283	8928	0.76 (0.70-0.84)	<.001	1.03 (0.90-1.20)	
3	2	138	763	0.48 (0.26-0.90)	4	211	1388	0.51 (0.40-0.68)	<.001	0.93 (0.47-1.84)	
4	4	2122	30368	0.72 (0.66-0.80)	29	6264	66219	0.85 (0.80-0.90)	<.001	0.85 (0.75-0.96)	
5	2	182	1667	0.47 (0.22-1.04)	10	468	4253	0.52 (0.40-0.70)	<.001	0.91 (0.40-2.06)	
6	1	152	1627	0.62 (0.46-0.86)	7	1240	13617	0.78 (0.70-0.88)	<.001	0.79 (0.57-1.10)	
7	2	294	3112	0.63 (0.50-0.80)	22	3548	32832	0.76 (0.68-0.88)	<.001	0.83 (0.64-1.08)	
8	2	49	264	0.27 (0.14-0.54)	3	106	406	0.38 (0.24-0.60)	<.001	0.71 (0.32-1.62)	
9	2	41	913	0.26 (0.12-0.58)	9	307	3373	0.55 (0.42-0.74)	<.001	0.47 (0.20-1.10)	
10	1	183	305	0.76 (0.64-0.92)	1	125	220	0.64 (0.50-0.82)	<.001	1.19 (0.87-1.62)	
11	1	326	3825	0.73 (0.60-0.92)	22	1654	17571	0.84 (0.76-0.94)	.001	0.88 (0.69-1.12)	
12	3	395	3062	0.67 (0.56-0.82)	9	1225	4847	0.75 (0.64-0.90)	.001	0.90 (0.69-1.16)	
13	1	16	149	0.23 (0.08-0.80)	3	18	265	0.12 (0.04-0.44)	.001	1.99 (0.33-11.92)	
14	1	106	121	0.75 (0.64-0.88)	5	155	370	0.67 (0.54-0.86)	.002	1.11 (0.83-1.48)	
15	1	25	157	0.35 (0.16-0.82)	1	87	103	0.80 (0.70-0.92)	.002	0.43 (0.18-1.04)	
16	1	20	112	0.10 (0.02-0.42)	6	624	9199	0.81 (0.70-0.94)	.01	0.12 (0.03-0.52)	
17	2	141	929	0.73 (0.52-1.02)	3	850	11480	0.83 (0.74-0.96)	.01	0.88 (0.61-1.26)	
18	1	208	541	0.67 (0.54-0.84)	3	219	804	0.68 (0.52-0.90)	.01	1.00 (0.70-1.42)	
19	1	54	196	0.41 (0.24-0.70)	4	1017	2635	0.88 (0.80-0.98)	.01	0.46 (0.27-0.80)	
20	2	809	7966	0.69 (0.60-0.80)	3	2049	13876	0.82 (0.70-0.96)	.01	0.85 (0.69-1.06)	
21	2	19	46	0.21 (0.08-0.62)	3	79	388	0.57 (0.36-0.90)	.02	0.37 (0.12-1.20)	
22	3	131	389	0.57 (0.44-0.78)	38	4009	16909	0.88 (0.78-0.98)	.02	0.65 (0.48-0.90)	
23	4	320	2858	0.39 (0.24-0.66)	8	1018	4755	0.80 (0.64-1.00)	.04	0.48 (0.27-0.86)	
24	1	7	43	0.70 (0.48-1.06)	2	44	83	0.56 (0.32-1.00)	.04	1.26 (0.63-2.56)	
25	1	193	516	0.81 (0.66-1.02)	12	897	6242	0.79 (0.64-1.02)	.06	1.02 (0.73-1.42)	
26	2	226	1131	0.64 (0.50-0.82)	3	272	910	0.82 (0.66-1.02)	.07	0.78 (0.56-1.08)	
27	1	202	1229	0.72 (0.56-0.92)	6	749	3556	0.77 (0.58-1.06)	.10	0.92 (0.62-1.38)	
28	1	63	148	0.37 (0.24-0.58)	10	470	698	1.10 (0.98-1.22)	.10	0.34 (0.21-0.54)	
29	1	9	33	0.27 (0.08-1.12)	2	21	244	0.40 (0.14-1.24)	.11	0.68 (0.11-4.18)	
30	1	61	145	0.70 (0.48-1.04)	4	122	457	0.59 (0.30-1.16)	.12	1.19 (0.54-2.62)	
31	1	140	517	0.63 (0.48-0.84)	5	618	1092	0.82 (0.64-1.08)	.14	0.76 (0.51-1.14)	
32	1	264	13175	0.51 (0.40-0.66)	3	290	4414	0.86 (0.70-1.06)	.15	0.60 (0.43-0.84)	
33	2	723	6738	0.53 (0.46-0.62)	3	461	886	0.80 (0.60-1.10)	.15	0.67 (0.48-0.94)	
34	1	114	726	0.27 (0.18-0.42)	9	296	486	0.88 (0.74-1.06)	.16	0.31 (0.20-0.50)	
35	3	177	332	0.67 (0.56-0.84)	6	265	607	0.64 (0.34-1.22)	.18	1.05 (0.53-2.08)	
36	2	52	266	0.42 (0.22-0.78)	6	20	271	0.57 (0.24-1.38)	.21	0.73 (0.25-2.16)	
37	1	378	22071	0.58 (0.48-0.72)	3	674	14719	0.88 (0.72-1.08)	.21	0.66 (0.49-0.88)	
38	1	69	659	0.13 (0.06-0.28)	3	301	2556	0.33 (0.06-2.06)	.23	0.40 (0.06-2.86)	
39	1	30	75	0.63 (0.36-1.14)	3	131	232	1.34 (0.82-2.20)	.25	0.47 (0.22-1.02)	
40	1	43	95	0.53 (0.34-0.86)	1	18	37	0.67 (0.34-1.36)	.26	0.79 (0.34-1.84)	
41	1	51	508	0.38 (0.22-0.68)	6	259	2369	1.41 (0.76-2.62)	.27	0.27 (0.11-0.62)	
42	2	132	1718	0.52 (0.38-0.74)	16	578	3930	0.89 (0.72-1.12)	.31	0.59 (0.39-0.88)	
43	1	7	24	0.17 (0.02-1.18)	1	12	15	0.79 (0.50-1.26)	.32	0.21 (0.03-1.60)	
44	1	39	60	0.51 (0.32-0.80)	1	214	355	0.87 (0.66-1.18)	.35	0.58 (0.34-1.00)	
45	1	54	144	0.53 (0.34-0.84)	3	303	882	0.81 (0.54-1.26)	.35	0.65 (0.35-1.22)	
46	1	305	861	0.78 (0.66-0.94)	3	132	288	1.12 (0.88-1.46)	.36	0.69 (0.51-0.94)	
47	1	35	46	0.67 (0.48-0.96)	3	126	224	0.91 (0.76-1.12)	.37	0.73 (0.49-1.10)	
48	1	96	351	0.65 (0.46-0.94)	12	748	1811	1.07 (0.92-1.26)	.40	0.61 (0.41-0.90)	
49	2	18	162	0.06 (0.02-0.46)	4	137	1380	0.89 (0.66-1.22)	.46	0.07 (0.01-0.52)	
50	1	436	2161	0.39 (0.34-0.48)	4	363	1010	0.82 (0.50-1.40)	.46	0.48 (0.27-0.84)	
51	1	471	2135	0.70 (0.60-0.82)	4	618	3409	0.90 (0.68-1.22)	.47	0.78 (0.55-1.10)	
52	1	43	292	0.33 (0.18-0.64)	4	1036	2936	0.93 (0.74-1.16)	.50	0.36 (0.18-0.72)	
53	1	32	263	0.52 (0.26-1.04)	8	1180	5421	0.90 (0.68-1.22)	.51	0.58 (0.27-1.22)	
54	1	1558	19254	0.90 (0.82-1.00)	3	1847	21878	0.96 (0.86-1.08)	.51	0.93 (0.81-1.08)	
55	3	126	512	0.50 (0.30-0.88)	14	845	1325	0.95 (0.80-1.16)	.60	0.53 (0.29-0.96)	
56	1	5	20	0.42 (0.12-1.64)	2	51	81	0.91 (0.62-1.34)	.64	0.46 (0.11-1.90)	
57	1	9	20	0.13 (0.02-0.82)	1	13	59	0.83 (0.32-2.18)	.70	0.15 (0.02-1.26)	
58	1	26	701	0.18 (0.06-0.52)	1	17	1248	0.87 (0.34-2.24)	.77	0.21 (0.05-0.86)	
59	1	23	117	0.62 (0.42-0.94)	14	919	1766	0.98 (0.88-1.12)	.81	0.63 (0.41-0.98)	
60	2	49	99	0.61 (0.46-0.84)	2	88	160	0.97 (0.74-1.30)	.85	0.63 (0.41-0.96)	
61	1	47	100	0.47 (0.30-0.76)	10	700	1083	0.98 (0.82-1.20)	.86	0.48 (0.29-0.80)	
62	1	12	22	0.38 (0.16-0.96)	3	39	128	0.89 (0.24-3.40)	.87	0.43 (0.08-2.16)	
63	1	67	127	0.43 (0.30-0.66)	3	290	651	0.97 (0.56-1.72)	.91	0.45 (0.22-0.92)	

Random effects: $P<.001$ for heterogeneity; $I^2 = 57\%$

Test for overall effect: $z = 9.55$ ($P<.001$)

0.71 (0.65-0.77)

0.01 0.1 1.0 10

Ratio of Relative Risks (95% CI)

First column indicates number associated with the question addressed by each review that included 1 or more truncated and matching nontruncated RCTs. Results ordered by P values associated with results of nontruncated RCTs; size of the data markers indicates weight of review questions in meta-analysis.

Table 2. Meta-regression Model Investigating the Predictors of the Log[Ratio of Relative Risks][a]

	Univariable			Multivariable		
Independent Variable	β (95% CI)	P Value	$R^{2,b}$	β (95% CI)	P Value	$R^{2,b}$
Stopping rule	0.14 (0.02 to 0.27)	.02	0.08	0.07 (−0.05 to 0.19)	.24	
Additional No. of events in the truncated RCT (unit = 100 events)[c]	0.02 (0.01 to 0.03)	<.001	0.23	0.02 (0.01 to 0.02)	.001	0.24

Abbreviations: CI, confidence interval; RCT, randomized controlled trial.
[a] Model with the log[ratio of relative risks] as dependent variable. The meta-regression analysis was performed using data from 63 research questions.
[b] Reflects the proportion of the variability in the log[ratio of relative risks] explained by the statistical model.
[c] Given a mean ratio of 0.71, the addition of each 100, 500, 1000, or 1714 events to the truncated RCT would result in a new ratio of 0.72, 0.78, 0.87, or 1.00, respectively.

Figure 3. Weighted Bubble Plot Showing the Ratio of Relative Risks (RRs) vs the Total Number of Outcome Events in Truncated Randomized Controlled Trials (RCTs)
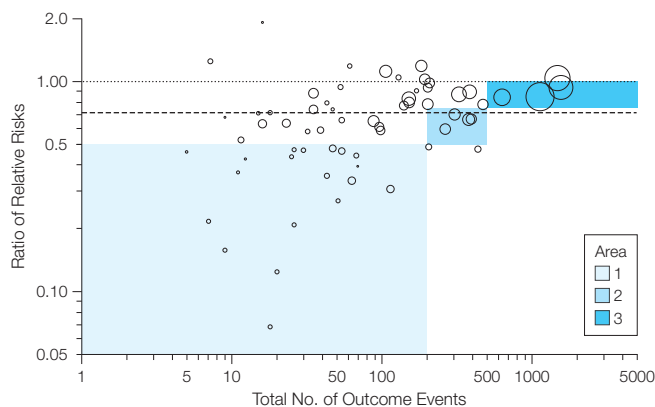
The size of each bubble is proportional to the magnitude of the inverse of the variance of the ratio of RR in the log scale. The dashed line indicates a ratio of RR of 0.71; the dotted line, a ratio of RR of 1.00. The shaded areas numbered 1 through 3 correspond to different degrees of overestimates of effect (ratios of RRs, 0.05-0.5; 0.5-0.75; 0.75-1.00): in area 1, very large overestimation (ratio of RR, 0.37; 95% confidence interval [CI], 0.31-0.44; P<.001) occurred in truncated trials with fewer than 200 events. In area 2, large overestimation (ratio of RR, 0.65; 95% CI, 0.56-0.77; P<.001) occurred in truncated trials stopped between 200 and 500 events. In area 3, truncated trials with more than 500 events led to moderate overestimation (ratio of RR, 0.88; 95% CI, 0.80-0.96; P=.003).

found that substantial differences between results at the time of early stopping and subsequent follow-up seldom occurred.[16] Freidlin and Korn published a related simulation study that supported these findings, suggesting that if the true effect is large, differences between stopped-early results and full follow-up results will differ little.[17]

These recent studies confirm that even when the true effect is large, studies stopped early still overestimate that effect. More important, the authors do not address circumstances in which the true underlying effect is small or absent. Clinicians seek the best estimate of an unknown true underlying effect with appropriate safeguards against bias. As Goodman[18] points out in a commentary on the simulation by Freidlin and Korn, "since we do not know what the true effect is, we can-

not know in any particular case whether the observed effect is biased or not; the fact that the trial is stopped early is not prima facie evidence that the estimate is wrong." We support this statement; unfortunately, neither do we know that the stopped-early result is close to the truth. Our findings suggest that often it is not.

Implications

Consensus exists that rigorous data monitoring practice requires a predefined statistical stopping rule.[19,20] Our findings, however, indicate that even a formal rule is insufficient to prevent bias consequent on stopping early and suggest the advisability of rules that require a large number of outcome events before early stopping is contemplated.

In this review we have focused only on RCTs stopped early for benefit. Al-

though ethical concerns make decisions regarding stopping trials early for safety more complex than those regarding stopping trials early for benefit, inferences regarding harm and those regarding benefit are equally susceptible to the bias associated with stopping early.

Our results have important implications for systematic reviews and ethics.[21,22] If reviewers do not note truncation and do not consider early stopping for benefit, meta-analyses will report overestimates of effects.[21] Investigators and funding bodies—in particular, drug and device manufacturers—have different but convergent interests to stop a study as soon as an important difference between experimental and control groups emerges, and journals have an interest in publishing the apparently exciting findings. Furthermore, data monitoring committees are well aware of their ethical obligation to ensure that patients are offered effective treatment as soon as it is clear that effective treatment is indeed available, providing a justification for stopping early.

However, data monitoring committees also have an ethical obligation to future patients who need to know more than whether data crossed a significance threshold; these patients need precise and accurate data on patient-important outcomes, of both risk and benefits, to make treatment choices.[22] Such patients will often number in the tens or hundreds of thousands and sometimes in the millions. To the extent that substantial overestimates of treatment effect are widely disseminated, patients and clinicians will be misled when trying to balance benefits, harms, inconvenience, and cost of a possible health care intervention. If the true treatment effect is negligible or

absent—as our results suggest it sometimes might be—acting on the results of a trial stopped early will be even more problematic. Thus, for trial investigators, our results suggest the desirability of stopping rules demanding large numbers of events. For clinicians, they suggest the necessity of assuming the likelihood of appreciable overestimates of effect in trials stopped early.

Authors/Members of the Study of Trial Policy of Interim Truncation-2 (STOPIT-2) Group: Dirk Bassler, MD, MSc (Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada, and Department of Neonatology, University Children's Hospital Tuebingen, Tuebingen, Germany); Matthias Briel, MD, MSc (Department of Clinical Epidemiology and Biostatistics, McMaster University, and Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Basel, Switzerland); Victor M. Montori, MD, MSc, Melanie Lane, BA, David N. Flynn, BS, Mohamed B. Elamin, MBBS, Mohammad Hassan Murad, MD, MPH, Nisrin O. Abu Elnour, MBBS, Julianna F. Lampropulos, MD, Amit Sood, MD, MSc, Rebecca J. Mullan, MSc, and Patricia J. Erwin, MLS (Knowledge and Encounter Research Unit, Mayo Clinic, Rochester, Minnesota); Paul Glasziou, MBBS, PhD, Clare R. Bankhead, DPhil, and Rafael Perera, DPhil, MSc (Centre for Evidence-Based Medicine, Department of Primary Health Care, University of Oxford, Oxford, United Kingdom); Qi Zhou, PhD, Diane Heels-Ansdell, MSc, Carolina Ruiz Culebro, MD, John J. You, MD, MSc, Sohail M. Mulla, Jagdeep Kaur, PhD, CRA, Kara A. Nerenberg, MD, MSc, Holger Schünemann, MD, PhD, Deborah J. Cook, MD, MSc, Kristina Lutz, Christine M. Ribic, MD, MSc, Noah Vale, MD, Stephen D. Walter, PhD, and Gordon H. Guyatt, MD, MSc (Department of Clinical Epidemiology and Biostatistics, McMaster University); German Malaga, MD, MSc (Universidad Peruana Cayetano Heredia, Lima, Peru); Elie A. Akl, MD, PhD (Department of Clinical Epidemiology and Biostatistics, McMaster University, and Departments of Medicine and Family Medicine, State University of New York at Buffalo); Ignacio Ferreira-Gonzalez, PhD (Cardiology Department, Vall d'Hebron Hospital, CIBER de Epidemiología y Salud Pública, Barcelona, Spain); Pablo Alonso-Coello, MD, PhD, and Gerard Urrutia, MD (Centro Cochrane Iberoamericano, Hospital Sant Pau, Barcelona, and CIBER de Epidemiologia y Salud Pública, Barcelona); Regina Kunz, MD, MSc, Heiner C. Bucher, MD, MPH, Alain J. Nordmann, MD, MSc, and Heike Raatz, MD, MSc (Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel); Suzana Alves da Silva, MD, MSc, and Fabio Tuche, MD (Teaching and Research Center of Pro-Cardiaco, Rio de Janeiro, Brazil); Brigitte Strahm, MD (Pediatric Hematology and Oncology Centre for Pediatrics and Adolescent Medicine, University Hospital Freiburg, Freiburg, Germany); Benjamin Djulbegovic, MD, PhD (Center for Evidence-based Medicine, USF Health Clinical Research, Tampa, Florida); Neill K. J. Adhikari, MD, MSc (Sunnybrook Health Sciences Centre and University of Toronto, Toronto, Ontario, Canada); Edward J. Mills, PhD (British Columbia Centre for Excellence in HIV/AIDS, University of British Columbia, Vancouver, Canada); Femida Gwadry-Sridhar, PhD (University of Western Ontario, London, Canada); Haresh Kirpalani, MD, MSc (Department of Clinical Epidemiology and Biostatistics, McMaster University, and Children's Hospital Philadelphia, Philadelphia, Pennsylvania); Heloisa P. Soares, MD (Mount Sinai Medical Center, Miami Beach, Florida); Paul J. Karanicolas, MD, PhD (Memorial Sloan-Kettering Cancer Center, New York, New York); Karen E. A. Burns, MD, MSC (St. Michael's Hospital, Keenan Research Centre and Li Ka Shing Knowledge Institute, University of Toronto); Per Olav Vandvik, MD, PhD (Department of Medicine, Gjøvik, Innlandet Hospital Trust, Norway); Fernando Coto-Yglesias, MD (Hospital Nacional de Geriatría y Gerontología San José, Costa Rica); Pedro Paulo M. Chrispim, MSc (National School of Public Health, Rio de Janeiro); and Tim Ramsay, PhD (Ottawa Hospital Research Institute, University of Ottawa, Ottawa, Ontario, Canada).

Online-Only Material: An online list of references for the RCTs analyzed in this article is available at http://www.jama.com

REFERENCES

1. Psaty BM, Rennie D. Stopping medical research to save money. *JAMA*. 2003;289(16):2128-2131.
2. Wheatley K, Clayton D. Be skeptical about unexpected large apparent treatment effects. *Control Clin Trials*. 2003;24(1):66-70.
3. Pocock S, White I. Trials stopped early: too good to be true? *Lancet*. 1999;353(9157):943-944.
4. Schulz KF, Grimes DA. Multiplicity in randomised trials: subgroup and interim analyses. *Lancet*. 2005;365(9471):1657-1661.
5. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. *Control Clin Trials*. 1989;10(4)(suppl):209S-221S.
6. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA*. 2005;294(17):2203-2209.
7. Bassler D, Montori VM, Briel M, et al. Early stopping of randomized clinical trials for overt efficacy is problematic. *J Clin Epidemiol*. 2008;61(3):241-246.
8. Sydes MR, Parmar MK. Interim monitoring of efficacy data is important and appropriate. *J Clin Epidemiol*. 2008;61(3):203-204.
9. Trotta F, Apolone G, Garattini S, Tafuri G. Stopping a trial early in oncology: for patients or for industry? *Ann Oncol*. 2008;19(7):1347-1353.
10. Sargent D. Early stopping for benefit in National Cancer Institute–sponsored randomized Phase III trials. *J Clin Oncol*. 2009;27(10):1543-1544.
11. Goodman SN. Stopping at nothing? some dilemmas of data monitoring in clinical trials. *Ann Intern Med*. 2007;146(12):882-887.
12. Briel M, Lane M, Montori VM, et al. Stopping randomized trials early for benefit: a protocol of the Study Of Trial Policy Of Interim Truncation-2 (STOPIT-2). *Trials*. 2009;10:49.
13. Pocock SJ. When (not) to stop a clinical trial for benefit. *JAMA*. 2005;294(17):2228-2230.
14. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life. *Med Care*. 2003;41(5):582-592.
15. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev*. 2009;(1):MR000006.
16. Korn EL, Freidlin B, Mooney M. Stopping or reporting early for positive results in randomized clinical trials. *J Clin Oncol*. 2009;27(10):1712-1721.
17. Freidlin B, Korn EL. Stopping clinical trials early for benefit: impact on estimation. *Clin Trials*. 2009;6(2):119-125.
18. Goodman SN. Stopping trials for efficacy. *Clin Trials*. 2009;6(2):133-135.
19. DAMOCLES Study Group; NHS Health Technology Assessment Programme. A proposed charter for clinical trial data monitoring committees. *Lancet*. 2005;365(9460):711-722.
20. Pocock SJ. Current controversies in data monitoring for clinical trials. *Clin Trials*. 2006;3(6):513-521.
21. Bassler D, Ferreira-Gonzalez I, Briel M, et al. Systematic reviewers neglect bias that results from trials stopped early for benefit. *J Clin Epidemiol*. 2007;60(9):869-873.
22. Mueller PS, Montori VM, Bassler D, Koenig BA, Guyatt GH. Ethical issues in stopping randomized trials early because of apparent benefit. *Ann Intern Med*. 2007;146(12):878-881.