Randomized Trials Stopped Early for Benefit
A Systematic Review

Victor M. Montori, MD, MSc

P. J. Devereaux, MD

Neill K. J. Adhikari, MD

Karen E. A. Burns, MD

Christoph H. Eggert, MD

Matthias Briel, MD

Christina Lacchetti, MHSc

Teresa W. Leung, BHSc

Elizabeth Darling, RM, BHSc

Dianne M. Bryant, PhD

Heiner C. Bucher, MD, MPH

Holger J. Schünemann, MD, PhD

Maureen O. Meade, MD, MSc

Deborah J. Cook, MD, MSc

Patricia J. Erwin, MLS

Amit Sood, MD

Richa Sood, MD

Benjamin Lo, MD

Carly A. Thompson, BHSc

Qi Zhou, PhD

Edward Mills, PhD

Gordon H. Guyatt, MD, MSc

Context Randomized clinical trials (RCTs) that stop earlier than planned because of apparent benefit often receive great attention and affect clinical practice. Their prevalence, the magnitude and plausibility of their treatment effects, and the extent to which they report information about how investigators decided to stop early are, however, unknown.

Objective To evaluate the epidemiology and reporting quality of RCTs involving interventions stopped early for benefit.

Data Sources Systematic review up to November 2004 of MEDLINE, EMBASE, Current Contents, and full-text journal content databases to identify RCTs stopped early for benefit.

Study Selection Randomized clinical trials of any intervention reported as having stopped early because of results favoring the intervention. There were no exclusion criteria.

Data Extraction Twelve reviewers working independently and in duplicate abstracted data on content area and type of intervention tested, reporting of funding, type of end point driving study termination, treatment effect, length of follow-up, estimated sample size and total sample studied, role of a data and safety monitoring board in stopping the study, number of interim analyses planned and conducted, and existence and type of monitoring methods, statistical boundaries, and adjustment procedures for interim analyses and early stopping.

Data Synthesis Of 143 RCTs stopped early for benefit, the majority (92) were published in 5 high-impact medical journals. Typically, these were industry-funded drug trials in cardiology, cancer, and human immunodeficiency virus/AIDS. The proportion of all RCTs published in high-impact journals that were stopped early for benefit increased from 0.5% in 1990-1994 to 1.2% in 2000-2004 ($P<.001$ for trend). On average, RCTs recruited 63% (SD, 25%) of the planned sample and stopped after a median of 13 (interquartile range [IQR], 3-25) months of follow-up, 1 interim analysis, and when a median of 66 (IQR, 23-195) patients had experienced the end point driving study termination (event). The median risk ratio among truncated RCTs was 0.53 (IQR, 0.28-0.66). One hundred thirty-five (94%) of the 143 RCTs did not report at least 1 of the following: the planned sample size (n=28), the interim analysis after which the trial was stopped (n=45), whether a stopping rule informed the decision (n=48), or an adjusted analysis accounting for interim monitoring and truncation (n=129). Trials with fewer events yielded greater treatment effects (odds ratio, 28; 95% confidence interval, 11-73).

Conclusions RCTs stopped early for benefit are becoming more common, often fail to adequately report relevant information about the decision to stop early, and show implausibly large treatment effects, particularly when the number of events is small. These findings suggest clinicians should view the results of such trials with skepticism.

JAMA. 2005;294:2203-2209 www.jama.com

WHEN RANDOMIZED clinical trials (RCTs) identify larger than expected treatment effects, investigators may conclude, before completing the trial as planned, that one treatment is superior to the other. Such trials often receive considerable attention. For example, consider the impact on current practice of published RCTs that were stopped early documenting the effect of β-blockers in patients undergoing vascular surgery[1] and of tight control of glucose levels using insulin in patients in the intensive care unit.[2]

Little is known, however, about the prevalence of truncated trials in the medical literature.

Author Affiliations are listed at the end of this article.
Corresponding Author: Gordon H. Guyatt, MD, MSc, Department of Clinical Epidemiology and Biostatistics, Health Sciences Centre Room 2C12, McMaster University, Hamilton, Ontario, Canada L8N 325 (guyatt@mcmaster.ca).

Clinicians face challenges when interpreting the results of truncated RCTs. Taking the point estimate of the treatment effect at face value will be misleading if the decision to stop the trial resulted from catching the apparent benefit of treatment at a "random high." When this occurs, data from future trials will yield a more conservative estimate of treatment effect, the so-called regression to the truth effect.[3]

Thus, clinicians must attend not only to the usual methodological safeguards against bias[4] but also to characteristics that affect the decision to stop a trial early. Such characteristics include the plausibility of the treatment effect, the planned sample size, the number of interim analyses after which the investigators stopped the RCT, and the statistical methods used to monitor the trial and to adjust estimates, *P* values, and confidence intervals for interim analyses. While the Consolidated Standards of Reporting Trials (CONSORT) statement recommends reporting of sample size estimations, interim analyses, and stopping rules,[5] little is known about the extent to which investigators reporting the results of RCTs stopped early for benefit adhere to these recommendations.

While RCTs stopped early for reasons other than benefit might share some characteristics with RCTs stopped early for benefit, their implications are very different. Trials stopped early because of harm or futility tend to result in decreased use or prompt discontinuation of useless or potentially harmful interventions. In contrast, trials stopped early for benefit may result in the quick identification, approval, and dissemination of promising new treatments. Because of these profound differences, we focused our work on the epidemiology and reporting quality of RCTs involving interventions stopped early for benefit.

METHODS
Eligibility Criteria
We included RCTs of any intervention reported as having stopped earlier than planned because of results in favor of the experimental intervention. There were no exclusion criteria.

Literature Search
An experienced reference librarian (P.J.E.) and one of the investigators (V.M.M.) developed sensitive search strategies to identify RCTs stopped early for benefit. First, we identified search terms from the trials that our research group knew were stopped early for benefit. Next, we created database-specific strategies incorporating terms for randomized trials, for early termination (eg, *halted*, *closed*, *closure*, *terminated*, *stopped*, *early*, *prematurely*), and for procedures associated with this decision (eg, *data and safety monitoring board* [DSMB], *interim analysis*, and *stopping rules or boundaries* and their specific names [eg, *O'Brien-Fleming*]). In November 2004, we searched MEDLINE, EMBASE, and Current Contents from their inception. Also, we searched databases that included the full text of journals (OVID, ScienceDirect, Ingenta, and HighWire Press, including articles in full text from approximately 1700 journals since 1993), and we searched the Web sites of selected general medicine journals that offered access to full text at that time (*Lancet*, *New England Journal of Medicine*, *JAMA*, *Annals of Internal Medicine*, and *BMJ*).

Study Selection
Two investigators (V.M.M. and C.H.E.) working in duplicate scanned all abstracts and obtained the full-text reports of records that indicated or suggested that an RCT was stopped early (ie, a review article on stopping rules, a clinical review that discussed an RCT stopped early for any reason, or the report of an RCT stopped early). After obtaining all reports of the candidate trials (eg, reports of methods and baseline characteristics, reports of preliminary results, discussions of the stopping rules and decisions), the same reviewers independently assessed eligibility from full-text articles from the formal search, from investigators' recall, and from searches of the full-text databases.

To estimate the denominator for calculating the prevalence of RCTs stopped early for benefit, we identified RCTs published for each year from 1975 to 2004 by a search of MEDLINE using the publication type *randomized controlled trial* (the National Library of Medicine and the US Cochrane Center have collaborated to accurately index RCTs published from 1966 to 2002 in the MEDLINE database using this term [available at http://www.cochrane.us/central.htm]). Trials stopped early for benefit yield positive and, typically, large treatment effects. Investigators faced with such results may surmise that the probability of publishing their results in high-impact journals is great. If this is so, then journals with lower impact factors are likely at lower risk of publishing such RCTs. In keeping with this hypothesis, we twice calculated the proportion of RCTs stopped early for benefit, first using the MEDLINE-indexed journals as the denominator, and then restricting the sample to the 10 general medicine journals publishing RCTs and the 5 journals in cardiology and hemato-oncology (ie, the clinical fields most commonly generating truncated RCTs) with the highest impact factor.

Data Extraction
Forming review pairs, 12 reviewers (N.K.J.A., K.E.A.B., C.H.E., M.B., E.D., D.M.B., H.C.B., H.J.S., M.O.M., D.J.C., A.S., and R.S.) trained in health research methodology conducted data extraction independently and in duplicate, using a standardized form and a data collection manual. Reviewers collected information about the content area and the type of interventions tested, the type of end point driving study termination, the treatment effect, the length of follow-up, sample size estimation and total sample studied, the existence and role of the DSMB in stopping the study, the number of interim analyses planned and conducted, and the existence and type of monitoring methods, statistical boundaries, and adjustment procedures for interim analyses and early stopping. We also noted the general methodological quality features of each trial (eg, reporting of randomization and

blinding) and the funding agents and their role in the study. Finally, we noted in which sections of the manuscript the authors reported that they stopped the study early for benefit.

We entered the data in an electronic database such that duplicate entries existed for each study; when the 2 entries did not match, 2 additional investigators also trained in health research methods (V.M.M., C.L.) reviewed all the reports of the included trials and together adjudicated the entry.

Data Analysis

The φ statistic provided a measure of interobserver agreement independent of chance regarding RCT eligibility.[6] We used the $χ^2$ test for trend with 1 df to evaluate whether there was a trend over time in the proportion of published RCTs stopped early. Using logistic regression, we estimated the extent of association between quality of reporting of CONSORT-recommended items (planned sample size, interim analysis, and stopping rules) as the dependent variable and year (in 5-year intervals) and journal of publication (5 medical journals with highest 2004 impact factor that publish RCTs vs others) as independent variables. Logistic regression also provided methods for estimating the extent of association between the total number of events (ie, end points driving termination) in the trial and the calculated effect size using several thresholds (both variables dichotomized at their median values and at the largest quartile for events and for effect size). We expressed these associations using odds ratios (ORs) and associated 95% confidence intervals (CIs). Analyses were performed using SAS version 11.0 (SAS Institute Inc, Cary, NC); $P<.05$ was used to determine statistical significance.

RESULTS

Results of Literature Search

Our literature search generated 487 abstracts, from which we identified 90 eligible trials. From databases supplying full-text publications we identified an additional 42 trials (29 from *New En-*

Table 1. Characteristics of Randomized Clinical Trials (RCTs) Stopped Early for Benefit (N = 143)

Characteristic	No./Total (%)	
	RCTs Stopped Early/ RCTs Indexed in MEDLINE	RCTs Stopped Early/ RCTs in Top-Impact Journals
Year of publication		
1975-1979	1/6574 (0.01)	0/620 (0)
1980-1984	1/12 653 (0.008)	1/1175 (0.1)
1985-1989	10/21 807 (0.05)	9/1938 (0.5)
1990-1994	19/38 712 (0.05)	15/3106 (0.5)
1995-1999	41/52 060 (0.08)	35/3594 (1.0)
2000-2004	71/58 537 (0.1)	47/3859 (1.2)
Area of study		
Cardiology	36 (25)	
Cancer (hematology-oncology)	30 (21)	
HIV/AIDS	17 (12)	
Critical care	10 (7)	
Other areas	50 (35)	
Type of comparisons		
Active medication vs placebo	76 (53)	
Active medication vs active medication	31 (22)	
Nonpharmacological therapeutic interventions (eg, invasive procedures, rehabilitation)	23 (16)	
Drug vs nonpharmacological therapeutic intervention	12 (8)	
Nontherapeutic interventions (eg, education)	1 (1)	
Type of end point driving decision to stop trial		
Dichotomous		
Single	95 (66)	
Composite	32 (22)	
Continuous	16 (11)	
Quality of reporting of safeguards against bias		
Adequate randomization method	84 (59)	
Adequate allocation concealment	76 (53)	
Blinding		
Participants	77 (54)	
Clinicians	61 (43)	
Data collectors	39 (27)	
Data analysts	7 (5)	
Judicial assessors of outcomes	58 (41)	
Reported planned sample size	115 (80)	
Sections reporting RCT stopped early		
Title	2 (1)	
Abstract	95 (66)	
Introduction	25 (18)	
Methods		
Not statistical section	16 (11)	
Statistical section	38 (27)	
Results		
First paragraph	57 (40)	
Elsewhere in Results	38 (27)	
Discussion	57 (40)	
Funding		
For-profit agency (eg, pharmaceutical industry)	64 (45)	
Only source reported	36 (56)	
Along with not-for-profit/government funding	28 (44)	
Not-for-profit organization/government agency only	53 (37)	
Not reported	26 (18)	
Report of competing interests		
No report of competing interests	100 (70)	
Reported employment with funding agency	24 (17)	
Reported potential conflict other than employment	16 (11)	
Reported no competing interests	3 (2)	

Abbreviation: HIV, human immunodeficiency virus.

gland Journal of Medicine, 8 from *Lancet,* 3 from *JAMA*, 1 from *BMJ*, and 1 from *Annals of Internal Medicine*); our research group identified 13 additional RCTs for a total of 145 potentially eligible RCTs, of which 143 proved eligible on consensus review. Chance-independent agreement about eligibility between reviewers was ex-cellent ($\phi = 0.98$). **TABLE 1** describes the characteristics of the 143 eligible RCTs (an online list of published studies is available at http://www.jama.com).

Epidemiology of RCTS Stopped Early for Benefit

The relative prevalence of RCTs reporting having stopped early for benefit, cal-culated using either RCTs published in all MEDLINE-indexed journals or the subset of RCTs published in high-impact journals as denominators, is in-creasing (Table 1) ($P < .001$ for linear trend): 55 were published in *New England Journal of Medicine*, 27 in *Lancet*, 6 in *JAMA*, 2 in *Annals of Internal Medicine*, 2 in *BMJ*, and 51 in other publications.

Most RCTs stopped early studied phar-macological interventions and were con-ducted in clinical areas rich in RCTs and in which patients could derive relatively rapid benefit from starting a new effec-tive therapeutic intervention (eg, acute coronary syndromes, lung cancer). In 32 RCTs, 15 of which were in cardiology, a composite end point drove early termi-nation. Trials had planned sample sizes of up to 18 000 patients (median, 400; in-terquartile range [IQR], 148-1100). A for-profit agency funded 64 of the 117 trials with declaring a funding source.

Stopping Characteristics of Trials Stopped Early for Benefit

TABLE 2 describes the stopping char-acteristics of these RCTs. In the 105 RCTs that reported a planned sample size and prematurely stopped recruit-ing, investigators on average enrolled 63% (SD, 25%) of the target sample. Trials were stopped after following up patients for a median duration of 13 (IQR, 3-25) months. Trials with di-chotomous end points as the events driving termination were stopped af-ter accruing a median of 66 (IQR, 23-195) events.

Data and Safety Monitoring Board

Of the 143 RCTs, 99 (69%) reported having a DSMB (or a person or group assigned to that role) and 68 reported its composition; of these, 2 included a sponsor representative. Fourteen of the 99 DSMBs were blinded to group allocation when assessing interim report data. Eighty-five reports (ie, 86% of those with a DSMB) indicated that the DSMB recommended stop-ping the trial; in 1 trial, the DSMB showed the interim data to the investi-gators without making a recommenda-

Table 2. Stopping Characteristics of Randomized Clinical Trials Stopped Early for Benefit (N = 143)

Characteristic	No. (%)
Type of stop	
Stopped recruitment	104 (73)
Continued follow-up	51 (49)
Stopped follow-up	53 (51)
Stopped follow-up after completing recruitment	30 (21)
Could not be determined	9 (6)
Interim analyses: definition of interval	
After enrolling a set number of participants	51 (36)
After a calendar period after date of trial start	32 (22)
After a set number of end points accrue	6 (4)
After a set follow-up (eg, patient-years of observation)	8 (6)
Ad hoc	6 (4)
Not reported	40 (28)
Monitoring methods/stopping boundaries	
No method or α spending function used*	28 (20)
Method specified	
O'Brien-Fleming boundary	38 (27)
With Lan-DeMets α spending function	15 (19)
Haybittle-Peto boundary	16 (11)
Pocock boundary	10 (7)
Triangular boundaries	3 (2)
Prespecified *P* value (α spending function not reported)	15 (10)
Other boundaries/α spending functions	13 (9)
Monitoring methods not specified	20 (14)
Role of monitoring method/stopping boundary in trial termination	
Results exceeded stopping boundary	90 (63)
Unrelated to stopping boundary/no stopping boundaries or rules in place	46 (32)
Trial continued despite results exceeding stopping boundary	3 (6)
Unclear	7 (5)
Adjustments for early stop/interim analyses	
None reported	129 (90)
Reported	14 (10)
On point estimate, confidence interval, and *P* value	5 (36)
On confidence interval or *P* value only	9 (64)
Adjusted estimates reported in abstract	11 (8)
Who made decision to stop	
Executive committee	109 (76)
Following recommendation from data and safety monitoring board	77 (71)
Data and safety monitoring board	8 (6)
Not-for-profit sponsor	2 (1)
Not reported	24 (17)

*To ensure that multiple interim analyses will not inflate the type I error (ie, risk of concluding there is a difference when there is not), the α spending function preserves a study-wide α level (often set at .05) by allocating varying α levels to conclude that a difference exists at each interim analysis. Typically, the earlier the analysis, the smaller the α level (eg, ≤.001 for the first analysis).

tion; in another, the DSMB recommended trial continuation; in the other 12, the DSMB action was not specified.

Interim Analyses and Stopping Rules

Of the 143 RCTs, 91 (64%) reported the number of planned interim analyses. Of these, 2 reported not planning any interim analyses (and both stopped after 1 interim analysis), 33 planned 1 interim analysis, 25 planned 2, and 31 planned more than 2 (up to 39). Of the 103 RCTs that reported the interim analysis after which they stopped the RCT, 55 stopped after the first analysis, 29 after the second, and 19 after 3 or more analyses (up to 25).

In approximately one third (48/143) of RCTs stopped early for benefit, a statistical approach to monitoring the trial was either not used or not specified in the report. Most RCTs (n=109) were stopped by the trial's executive committee (or the investigator in that role); 90 were stopped after the results exceeded a stopping boundary.

Only 8 RCTs reported all 4 key methodological elements: planned sample size, the interim analysis after which the RCT was stopped, the stopping rules used to inform that decision, and adjusted estimates for interim analysis and early stopping. Sixty-seven RCTs reported the first 3 elements. Compared with other journals, the 5 medical journals with the highest impact factors were more likely to publish RCTs reporting these 3 features (49/91 vs 18/52; OR, 2.2; 95% CI, 1.1-4.5). After taking into account journal of publication, the proportion of truncated RCTs reporting these 3 features has improved over time (logistic regression for each 5-year period: OR, 1.5; 95% CI, 1.02-2.1; P=.04).

Treatment Effect

Among the 99 trials with dichotomous unfavorable end points, 67 reported summary effect estimates. Of these, 41 reported a median risk ratio (RR) of 0.56 (IQR, 0.35-0.67), 22 reported a median hazard ratio of 0.48 (IQR, 0.32-0.59), and 4 reported a median OR of 0.30 (IQR, 0.15-0.45). We estimated RR for all RCTs with dichotomous outcomes reporting event rates (n=126): the median RR was 0.53 (IQR, 0.28-0.66).

Trials that accrued fewer events (ie, end points driving termination) estimated larger treatment effects; this finding is consistent at the median (66 events) and 75th percentile of events (195 events) and at the median (RR, 0.53) and top quartile (RR, 0.28) of treatment effect (median events–median effect: OR, 28; 95% CI, 11-73; median events–top quartile effect: OR, 28; 95% CI, 6-123; top quartile events–median effect: OR, 26; 95% CI, 6-116). There were no RCTs with more than 195 events in which the point estimate of the RR was less than 0.50.

COMMENT

Findings

We found an increasing prevalence of RCTs reported to have stopped early for benefit, with clustering of publication in the top general medical journals. Many RCTs evaluated cardiovascular or cancer interventions and were funded by for-profit agencies. These RCTs stopped after recruiting approximately 64% of the planned sample and after a median of 13 months of follow-up and 1 interim analysis, documenting a median of 66 patients experiencing the end points (events) driving termination and estimating a median RR of 0.53. We found a strong and inverse association between the number of events and the estimated treatment effect.

We noted limited reporting of critical features specific to the decision to stop the trial: only 67 (47%) of the 143 trials reported the planned sample size, the interim analysis after which the decision was made to stop the RCT, and the stopping rule used to inform this decision; only 8 trials reported these 3 elements and in addition reported adjusted estimates for interim analysis and early stopping. Reporting of these 3 characteristics was better in RCTs published in the top 5 medical journals; nevertheless, half of these RCTs did not report 1 or more of them.

Limitations and Strengths

The sample of 143 RCTs stopped early for benefit is the result of a systematic search. Since RCTs stopped early are not indexed accordingly and authors do not systematically report early stopping in abstracts, searching MEDLINE and similar databases will miss eligible RCTs. We therefore searched in databases of full-text journals. Because we identified trials in these databases mostly by reporting of characteristics associated with stopping early, we may have overestimated the quality of reporting of these characteristics in the included RCTs. Nonetheless, we found the quality of reporting of 3 CONSORT-recommended characteristics (planned sample size, interim analyses, and stopping rules) to be lacking.

Strengths of our study include the comprehensive and careful data collection, including independent judgment and abstraction of data at all stages with adequate reliability; and the use of targeted, relevant analyses.

Implications

Truncated RCTs reported as having stopped early for benefit are becoming more prevalent. Investigators, patients and their advocates, institutional review boards, and funding agencies may have different but convergent interests to stop a study as soon as an important difference between experimental and control groups emerges. Increased impact and publicity may motivate investigators, journals, and funding agencies. A commitment to promptly offer participants in the less favorable group the better treatment choice may motivate investigators, patients and their advocates, and the DSMB. The opportunity to save research dollars by truncating a trial may motivate the funding agencies.

An increase in the prevalence of RCTs with planned sample sizes and an increase in the prevalence of RCTs stopped early that report doing so may also contribute to the increase in prevalence. Less likely, the increase is due to poorly planned studies, inadequate

comparators, or to the increasing availability of highly effective treatments.

As in 22% of the trials, use of a composite end point to monitor the trial may lead to decisions to stop driven by the least patient-important outcome that makes up the composite end point (eg, angina in a composite outcome of death, myocardial infarction, and angina).[7] Consequently, few events important to patients may accrue. Even when investigators do not use composite end points, few events will accrue in the end points not driving the decision to stop early for benefit. These end points may include patient-important beneficial events (eg, overall survival rather than progression-free survival[8]) or adverse events. Lack of adequate safety data may in turn affect the perceived and actual risk-benefit ratios (overestimating the benefit, underestimating the risk) of implementing the intervention in clinical practice.[9] This underscores the difficult decision faced by trialists and members of the DSMB who need to balance their ethical obligation to those in the trial (such as offering effective treatment to patients in the placebo group as soon as efficacy becomes evident) with their obligation to future patients,[10] who will need precise and accurate data on patient-important outcomes to make treatment choices.[11]

Clinicians making inferences on the basis of results in truncated trials face important challenges. The larger-than-expected treatment effect may in fact be a chance finding (ie, may result from analyzing the data at a "random high"). Take, for example, the RCT evaluating the efficacy of bisoprolol in patients with a positive dobutamine echocardiography result and undergoing elective vascular surgery.[1] At the time the trial was stopped, investigators had enrolled 112 patients (the authors had planned to recruit 266 patients, expecting an RR of 0.50), and the results had exceeded the O'Brien-Fleming boundary for benefit. The RR for the primary end point (cardiac death or nonfatal myocardial infarction) was 0.09 (95% CI, 0.02-0.37). This very large treat-ment effect is likely too good to be true. It is inconsistent with the researchers' expectations; with the magnitude of effect (ie, RRs of 0.65 to 0.85) of β-blockers in tens of thousands of patients with acute myocardial infarction or chronic management of congestive heart failure; with results in day-to-day clinical practice[12]; and with results of other trials, including a recently conducted RCT in 496 patients undergoing vascular surgery that showed no significant effect of β-blockers on cardiac death or nonfatal myocardial infarction.[13]

Investigators have reported instances in which study results exceeded stopping boundaries, but because the results yielded implausibly large treatment effects, the DSMB and the investigator agreed to continue the trial and ultimately found no significant benefit.[14,15] Furthermore, statistical simulation studies have shown that RCTs can overestimate the magnitude of the treatment effect depending on the timing (ie, the fraction of the total planned sample size or expected number of events) of the decision to stop.[10]

Adjustment of the estimate and of the precision of the estimate (or the P value) based on the number of interim "looks" at the data (using methods known as α spending functions) can provide insight into the true treatment effects underlying trials that stop early. For example, an RCT of intensive insulin therapy vs control in patients admitted to the intensive care unit found a large treatment effect on the risk of dying in the intensive care unit (RR, 0.58; 95% CI, 0.38-0.78). After adjusting for 4 interim analyses using the LanDeMets α spending function,[16] the apparent effect decreased substantially, and the upper confidence limit approached no difference (0.68; 95% CI, 0.45-0.98). The authors reported the adjusted estimates.[2]

An RCT of prolonged vs brief thromboprophylaxis with warfarin after total hip arthroplasty illustrates the potential impact of an adjusted analysis. The study found a large treatment effect on thromboembolic complications (1/184 vs 9/176; RR, 0.11; 95% CI, 0.01-0.83; P = .03) and stopped earlier than planned following 1 interim analysis.[17] Adjusting for this interim analysis and for stopping at one third of the planned sample[18] shows that the apparent effect decreases (RR, 0.29), the 95% CI is wider and consistent with harm (0.04-2.3), and the P value is no longer significant (P = .24). As with all but 14 of the RCTs in our cohort, these authors did not report adjusted results.

The risk of overestimating treatment effects is not uniform across all truncated RCTs. We found a very strong association between number of events and the magnitude of treatment effects—this association remained very strong across a number of cutpoints for calculation, up to almost 200 events. These results suggest that the risk of overestimating treatment effects decreases markedly when the number of events is very large. The threshold above which one can be confident that the results do not represent substantial overestimates of effect remains uncertain.

A number of observations suggest that investigators, journal editors, and clinical experts are not mindful of the problematic inferences that may arise from truncated RCTs. Top journals continue to publish results of trials stopped early but do not require authors to note the early stopping in the abstract and to report details that would allow readers to carefully evaluate the decision to stop early. Professional organizations continue to issue recommendations on the basis of trials stopped early for benefit, including those reporting very few end points that seem most likely to overestimate effects. Such recommendations include the use of perioperative β-blockers in patients undergoing vascular surgery.[19,20] Until investigators, editors, and guideline panels become more cautious in their interpretation of trials stopped early, the risk of prematurely translating unreliable study findings into clinical practice will continue.

Conclusion

Randomized clinical trials stopped early for benefit are becoming increasingly common, particularly in top medical journals. Adequate descriptions of the methods used to inform the decision to truncate the trial are often lacking. Trials stopped early for benefit, particularly those with few events, often report treatment effects that are larger than typical of interventions that have been definitively studied. These considerations suggest that clinicians should view results of RCTs stopped early for benefit with skepticism.

Author Affiliations: Department of Clinical Epidemiology and Biostatistics (Drs Montori, Devereaux, Adhikari, Burns, Bryant, Schünemann, Meade, Cook, Zhou, Mills, and Guyatt and Ms Lacchetti and Darling), Bachelor of Health Sciences Programme (Mss Leung and Thompson), and Department of Medicine (Drs Devereaux, Mead, Cook, Lo, and Guyatt), McMaster University, Hamilton, Ontario; Department of Medicine, Mayo Clinic College of Medicine, Rochester, Minn (Drs Montori, Eggert, A. Sood, and R. Sood and Ms Erwin); Interdepartmental Division of Critical Care, University of Toronto, Toronto, Ontario (Drs Adhikari and Burns); Basel Institute for Clinical Epidemiology, University Hospital, Basel, Switzerland (Drs Briel and Bucher); and Departments of Medicine and Social and Preventive Medicine, University at Buffalo, Buffalo, NY and Division of Clinical Research Development and INFORMAtion Translation, Italian National Cancer Institute Regina Elena, Rome, Italy (Dr Schünemann).

REFERENCES

1. Poldermans D, Boersma E, Bax JJ, et al; Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. *N Engl J Med*. 1999;341:1789-1794.
2. van den Berghe G, Wouters P, Weekers F, et al. Intensive insulin therapy in the critically ill patients. *N Engl J Med*. 2001;345:1359-1367.
3. Pocock S, White I. Trials stopped early: too good to be true? *Lancet*. 1999;353:943-944.
4. Guyatt G, Rennie D. *Users' Guides to the Medical Literature: A Manual of Evidence-Based Clinical Practice*. Chicago, Ill: AMA Press; 2002.
5. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*. 2001;285:1987-1991.
6. Cook R, Farewell V. Conditional inference for subject-specific and marginal agreement: two families of agreement measures. *Can J Stat*. 1995;23:333-344.
7. Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I, et al. Validity of composite end points in clinical trials. *BMJ*. 2005;330:594-596.

8. Cannistra SA. The ethics of early stopping rules: who is protecting whom? *J Clin Oncol*. 2004;22:1542-1545.
9. Juurlink DN, Mamdani MM, Lee DS, et al. Rates of hyperkalemia after publication of the Randomized Aldactone Evaluation Study. *N Engl J Med*. 2004;351:543-551.
10. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. *Control Clin Trials*. 1989;10(4 suppl):209S-221S.
11. Guyatt G, Montori V, Devereaux PJ, Schunemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. *ACP J Club*. 2004;140:A11-A12.
12. Devereaux PJ, Yusuf S, Yang H, Choi PT, Guyatt GH. Are the recommendations to use perioperative beta-blocker therapy in patients undergoing noncardiac surgery based on reliable evidence? *CMAJ*. 2004;171:245-247.
13. Yang H, Raymer K, Butler R, Parlow J, Roberts R, Tech M. Metoprolol after vascular surgery (MAVS). *Can J Anaesth*. 2004;51:A7.
14. Wheatley K, Clayton D. Be skeptical about unexpected large apparent treatment effects: the case of an MRC AML12 randomization. *Control Clin Trials*. 2003;24:66-70.

15. Slutsky AS, Lavery JV. Data safety and monitoring boards. *N Engl J Med*. 2004;350:1143-1147.
16. Lan K, DeMets D. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70:659-663.
17. Prandoni P, Bruchi O, Sabbion P, et al. Prolonged thromboprophylaxis with oral anticoagulants after total hip arthroplasty: a prospective controlled randomized study. *Arch Intern Med*. 2002;162:1966-1971.
18. Lan KK, Wittes J. The B-value: a tool for monitoring data. *Biometrics*. 1988;44:579-585.
19. Palda VA, Detsky AS. Perioperative assessment and management of risk from coronary artery disease. *Ann Intern Med*. 1997;127:313-328.
20. Eagle KA, Berger PB, Calkins H, et al; American College of Cardiology/American Heart Association. ACC/AHA guideline update for perioperative cardiovascular evaluation for noncardiac surgery—executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Update the 1996 Guidelines on Perioperative Cardiovascular Evaluation for Noncardiac Surgery). *J Am Coll Cardiol*. 2002;39:542-553.