Assessing Quality of Reports on Randomized Clinical Trials in Nursing Journals

Nicole Parent, RN, PhD, and James A. Hanley, PhD

Background. Several surveys have presented the quality of reports on randomized clinical trials (RCTs) published in general and specialty medical journals. The aim of these surveys was to raise scientific consciousness on methodological aspects pertaining to internal and external validity. These reviews have suggested that the methodological quality could be improved.

Objective. We conducted a survey of reports on RCTs published in nursing journals to assess their methodological quality. The features we considered included sample size, flow of participants, assessment of baseline comparability, randomization, blinding, and statistical analysis.

Methods. We collected data from all reports of RCTs published between January 1994 and December 1997 in *Applied Nursing Research*, *Heart & Lung* and *Nursing Research*. We hand-searched the journals and included all 54 articles in which authors reported that individuals have been randomly allocated to distinct groups. We collected data using a condensed form of the Consolidated Standards of Reporting Trials (CONSORT) statement for structured reporting of RCTs (Begg et al., 1996).

Results. Sample size calculations were included in only 22% of the reports. Only 48% of the reports provided information about the type of randomization, and a mere 22% described blinding strategies. Comparisons of baseline characteristics using hypothesis tests were abusively produced in more than 76% of the reports. Excessive use and unstructured reports of significance testing were common (59%), and all reports failed to provide magnitude of treatment differences with confidence intervals.

Conclusions. Better methodological quality in reports of RCTs will contribute to increase the standards of nursing research.

Address for correspondence: Nicole Parent, RN, PhD, Coordinator of the heart surgery sector, Montreal Heart Institute, 5000 Bélanger, Montreal, QC, H1T 1C8. E-mail: **nicole.parent@icm-mhi.org**

Key words: clinical trials, design, methods, meta-analysis

Background

A randomized clinical trial (RCT) provides the most valid basis for assessing the benefits of health care interventions. Results of a published RCT will be most influential on patient care if the reader can appreciate both the methodological quality of the trial and the quality of the report. The methodological quality of a trial itself depends on the accurate accomplishment of several fundamental steps. The quality of the report depends on whether the published RCT provides readers with adequate information on the design, implementation, analysis, and interpretation of the trial. Such a report will help readers judge both the internal and external validity of the trial. Unfortunately, surveys of reports on RCTs in leading medical journals have shown that investigators often neglect crucial aspects (Altman & Dore, 1990; Assmann, Pocock, Enos, & Kasten, 2000; Gotzsche, 1989; Liberati, Himel, & Chalmers, 1986; Pocock, Hughes, & Lee, 1987; Schulz, Chalmers, Grimes, & Altman, 1994). To our knowledge, no surveys have presented the quality of reports of RCTs published in nursing journals.

Several surveys have shown how RCTs are reported in general and specialty medical journals. For example, Assmann et al. (2000) conducted a systematic evaluation of 50 clinical trial reports published in the *British Medical Journal*, *Journal of the American Medical Association*, *The Lancet*, and the *New England Journal of Medicine* in 1997 to illustrate these problems. They found that the methods of randomization were often poorly described and about half the trials inappropriately used significance tests for baseline comparison. Schulz et al. (1994) reviewed 206 reports of trials published during 1990 and 1991 in two British and two U.S. obstetrics and gynecology journals. Only 32% of the reports specified how the randomization sequence was generated and only 22.8% reported on how the next available was concealed until the allocation of therapy. Schulz et al. (1994) stated that their findings not only highlight the importance of adequate methodological quality in RCTs, but also the importance of complete and reliable reporting. Without adequate reporting, assessing quality becomes impossible.

To improve the quality of reporting of clinical research, the Consolidated Standards of Reporting Trials (CONSORT) statement has been proposed for structured reporting of RCTs (Altman et al., 2001; Begg et al., 1996; Moher, Schulz, & Altman, 2001). The CONSORT statement is a checklist of 22 items and a flow diagram intended to assist authors, editors, and reviewers by ensuring that information pertinent to the trial is included in the study report. The checklist items pertain mainly to the methods, results, and discussion and identify key pieces of information necessary to evaluate the validity of the results. The flow diagram provides information about the progress of patients through a two-group parallel-design RCT, the type of trial most commonly reported. This present survey presented six major content areas that relate to crucial aspects of a trial's methodology and results. The purpose of this survey was to determine the extent to which reports of RCTs published in the nursing literature between 1994 and 1997, met the standards available in 1996 (Begg et al., 1996). The results of this survey have been presented, in part, at the annual meeting of the *Society for Clinical Trials* (Parent & Hanley, 2000).

Methods

We collected data from all reports of RCTs published in the four years between 1994 and 1997 in *Heart & Lung*, *Applied Nursing Research* and *Nursing Research*. We hand-searched the journals and included all consecutive articles in which authors reported that participants had been randomly allocated to distinct groups. One of us (N.P.) examined all reports individually and collected data using a condensed form of the CONSORT statement. This standardized evaluation covered the following six content areas: sample size calculation, flow of participants, randomization, blinding, assessment of baseline comparability, and statistical analysis (see Table 1). These content areas were chosen because they relate to important aspects of a trial's methodology (Begg et al., 1996; Moher, Dulberg, & Wells, 1994; Polit & Hungler, 1991).

Results

A total of 54 RCTs were published in the *Applied Nursing Research*, *Heart & Lung* and *Nursing Research* journals during the four years covered by our review. The median number of patients in the trials was 63; nine trials had fewer than 30 patients and 13 had more than 100 patients, including five trials with more than 200 patients. The large majority of trials involved nursing interventions in adult patients with cardiovascular diseases, in adolescents or in hospitalized infants. Examples of such interventions included exercise, early ambulation, relaxation, information, positioning, smoking cessation program, and other specialized care interventions.

Sample size and flow of participants

Information on whether the target sample size was based on prior statistical power calculations was included in only 12 (22%) of the reports (see Table 2). Nearly half (44%) gave no information about the flow of participants. In most reports, investigators only specified the numbers of participants initially randomized in treatment groups.

Randomization

Less than half of the reports (37%) described the method used for randomization: three trials reported the use of a computerized schema, five a table of random numbers, seven the use of coins, and five a random number draw. Half (50%) failed to provide information about the method used. Despite purporting to be randomized trials, seven reports (13%) described the use of a non-random method of assignment or did not use any method.

Blinding

In only 12 reports (22 %), a blinding strategy was specified by authors. Of those authors who did not specify a blinding strategy, we estimated that 28 (52%) could have provided blinding measures. Evidence for successful blinding of subjects, of persons carrying out the intervention, of outcome assessors and data analysts should be described. However, in many instances such as in RCTs conducted to compare educational programs or nursing interventions, blinding of subjects after allocation was not possible. Therefore, failure to employ any blinding strategy to reduce bias is understandable.

Table 1. Content areas used to assess quality of trial reports	
Sample Size	Sample size estimation gives the reader precise information on what potential intervention differences the RCT wishes to detect. The investigator should ensure that there is sufficient power to detect, as statistically significant, differences between groups considered to be of clinical interest (Friedman, Furberg, & DeMets, 1996). Alternatively, the investigator should ensure that there is a low probability of falsely concluding, from results that are not statistically significant, that a clinically meaningful result is absent. • Information on whether the target sample size was based on prior statistical power calculations should be included in the reports. These include the clinically important target difference between groups, the alpha (1-Type I error level), the statistical power (Type II error level) and the standard deviation for the measurements.
Flow of Participants	A flow diagram provides a trial profile summarizing participant progression in the study by intervention group. • A flow diagram should include the number of participants that were eligible, those who refused to participate, those who were randomized to treatment, those who completed the trial, and those who withdrew from the trial before follow-up was complete.
Randomization	Randomization is the assignment of patients to two or more treatment groups by chance alone. It implies that each participant has the same probability of receiving each of the possible treatments. Researchers use randomization to prevent bias in allocating patients to treatment and for its objectivity to remove bias (Piantadosi, 1997). Randomization guarantees that both observed and unobserved baseline differences between the treatment groups are attributable to chance. • Approaches to randomization considered as random included a coin toss, numbers drawn from hat, a table of random numbers, computer-generated random numbers, or random permuted blocks stratified by a factor. Non-random approaches to randomization included alternate assignment and assignment by odd/even number (for example, hospital procedure room number) were considered non-random.
Blinding	Blinding (masking) means that the patients on the study are unaware of which group they were allocated. Patients' knowledge of their treatment may influence them to report differentially symptoms, leading to biased results. Knowledge of the allocation schedule may also influence outcome assessors and data analysts, leading to biased results. • Blinding strategies should include efforts made to mask the allocation schedule to either patients, outcome assessors, or data analysts.
Assessment of Baseline Comparability	Assessment of baseline comparability in variables thought to affect the outcome allows the readers to evaluate whether the study groups were comparable before intervention was started. Baseline comparability should be addressed, but not with hypothesis tests. • Baseline demographic and clinical characteristics should be presented in a table with descriptive statistics (means and standard deviations should be reported for continuous variables; numbers and proportions should be reported for categorical variables).
Statistical Analysis	Continuous outcome measures are usually reported with means or difference in means (effect size), and binary outcomes measures as risk ratio, odds ratio, or risk difference. • Where p values are provided, exact values should be presented, rather than references to arbitrary levels. They provide a more precise statement as to the statistical significance of the trial result (e.g., $p = 0.008$ rather than $p < 0.01$). • The magnitude of treatment differences should be stated with confidence intervals.

Assessment of baseline comparability
Investigators presented comparisons of baseline characteristics using hypothesis tests in more than 76% of the reports. The differences reached statistical significance in four of the 41 studies (10%) at the 5% level for at least one variable.

Table 2. Quality of reports on RCTs in three nursing journals, 1994–1997

	Applied Nursing Research (n=12)	*Heart & Lung* (n=25)	*Nursing Research* (n=17)	Total (N=54)
Estimate of Sample Size				
Power Calculation Provided	3 (25%)	6 (24%)	3 (18%)	12 (22%)
Flow of participants				
Numbers eligible, randomized, refused, treated, completed, withdrew. Described	4 (33%)	13 (52%)	13 (76%)	30 (56%)
Randomization				
Computerized schema		2 (8%)	1 (6%)	3 (6%)
Table of random numbers	2 (17%)	1 (4%)	2 (12%)	5 (9%)
Coin toss or biased coin	1 (8%)	2 (8%)	4 (23%)	7 (13%)
Numbers drawn from hat	3 (25%)	1 (4%)	1 (6%)	5 (9%)
Alternate assignment	2 (17%)	2 (8%)	1 (6%)	5 (9%)
Odd or even numbers		1 (4%)		1 (2%)
Not specified	4 (33%)	15 (60%)	8 (47%)	27 (50%)
Not randomized		1 (4%)		1 (2%)
Blinding				
Blind to participants		1 (4%)		1 (2%)
Blind to outcome assessors	1 (8%)	3 (12%)		9 (16%)
Blind to participants and outcome assessors		2 (8%)		2 (4%)
Could have been to outcome assessors	9 (75%)	11 (44%)	8 (47%)	28 (52%)
Impossible to blind	2 (17%)	8 (32%)	4 (24%)	14 (26%)
Assessment of Baseline Comparability				
Used Statistical Tests				
No significant results	9 (75%)	17 (68%)	11 (65%)	37 (69%)
Significant results		3 (12%)	1 (6%)	4 (7%)
Did not use statistical tests	3 (25%)	5 (20%)	3 (17%)	11 (20%)
No baseline reported			2 (12%)	2 (4%)
Statistical Analysis				
Statistical significance reported				
Reported NS, Sig, or <.05	6 (50%)	7 (28%)	9 (53%)	22 (41%)
P values reported	6 (50%)	18 (72%)	8 (47%)	32 (59%)

Statistical analysis

Tests of statistical significance and corresponding *p* values were frequently used to present the estimated effect of the intervention on outcome measures. These tests assess how unusual the observed results would be assuming that the intervention under study has no effect (the null hypothesis). Statistical reporting of significance testing was inadequate in more than 40% of the reports surveyed. Investigators either failed to report exact *p* values or exact values of statistical tests, referring instead to nonsignificance (NS) or to arbitrary levels ($p<0.05$). No reports showed whether the observed differences were clinically important or used confidence intervals to quantify the treatment uncertainty or the precision in their estimates of effect.

Comments

Sample size

The intended size of a trial and the statistical justification of the intended size (e.g., power calculation and precision) should be specified in any report of RCT. In our survey, few of the reports (22%) specified the intended trial size, supported by a statement of statistical power. Similar results have been reported by others. In 45 trials published in the *British Medical Journal*, the *Lancet*, and the *New England Journal of Medicine* in 1985, statistical power was discussed in only 11% of the reports (Pocock et al., 1987). Two years later, information on statistical power was specified in 39% of the trials published in the same leading medical journals (n=80) and in the *Annals of Internal Medicine* (Altman & Dore, 1990). More recently, in 206 trials published in two British and two U.S. obstetrics and gynecology journals during 1990 and 1991, statistical power calculations were reported in only 24% of the reports (Schulz et al., 1994).

Pocock et al. (1987) argued that failure to report sample size calculations may suggest that the investigator had no preset trial size and reported the results at an arbitrary time, with the magnitude of treatment difference possibly affecting the decision to report. It may also suggest that the investigator had reported the trial before the intended trial size was achieved, because interim results showed a substantial treatment difference, or extended the trial beyond its intended size in order to achieve better statistical power. Inadequacies in trial size may cause the following misinterpretations of results. In a small trial that failed to reach statistical significance and where there was no preset trial size or the trial did not achieve the intended trial size, this may reflect a lack of statistical power and premature publication. The relationship between statistical power and negative studies (studies that failed to reach statistical significance) was illustrated by Moher et al. (1994) in a descriptive survey. They reviewed all (383) RCTs published in the *Journal of the American Medical Association*, the *Lancet*, and the *New England Journal of Medicine* in 1975, 1980, 1985, and 1990, and found that most trials with negative results did not have large enough sample sizes: only 16% and 36% had sufficient statistical power to detect a 25% or 50% relative treatment difference, respectively. They also found that among the 102 trials with negative findings, only 32% reported a sample size calculation. This percentage improved over time from 0% in 1975 to 43% in 1990. However, the percentage of trials with negative findings that lacked statistical power did not improve over time.

Flow of participants

All of the reports examined in the present review indicated the number of participants by intervention group. However, few provided complete information on the number of eligible participants, those who refused to participate, those who were randomized, treated, those who completed the trial, and those who withdrew from the trial before follow-up was complete. A flow diagram or a table should incorporate such information for each intervention group. Investigators should also report the numbers of participants who either received or complied with the treatment to which they were assigned, and those having received it initially, but who may have switched to another. The reasons for dropout could be incorporated into this diagram.

Because such problems in follow-up may lead to different ways of analyzing the results, investigators should report their approach to analysis (i.e., whether the main concern was efficacy or effectiveness). Efficacy refers to the potential effect of treatment under optimal circumstances. An analysis for efficacy would compare participants according to the treatment actually received and exclude those who complied poorly, those who switched over, and those who withdrew during the trial. Effectiveness analysis (intent to treat) refers to the actual effect of treatment in the "real world", comparing all participants according to their initial group assignment and, thus, including poor compliers, switch-overs, and withdrawals. Since a treatment that works under optimal circumstances may not necessarily work in clinical practice, treatment effectiveness is usually of priority in RCTs (Friedman et al., 1996). In our survey, none of the reports specified the approach to analysis.

Randomization

Information about the method used for randomization gives clear evidence that the trial was randomized. In our survey, such information was absent in 52% of the reports. Similar findings have

been reported in other surveys. For example, information about the method used for randomization was not provided in 51% and in 68% of the reports (Altman & Dore, 1990; Assmann et al., 2000; Schulz et al., 1994). For the present survey, although we considered coin toss as a "random" approach to randomization, this approach is less than optimal. Tables and computers are preferred because of reproducibility, and also their ease and speed.

Blinding

In our survey, 22% of reports specified the use of a blinding strategy, either of subjects and/or of outcome assessors. We estimated that among those who did not report blinding strategy, blinding could have been achieved in more than 52% of such trials. In 63 RCTs of primary treatment of early breast cancer reviewed in several medical journals, a similarly low percentage (8%) of the reports specified blinding of patients (Liberati et al., 1986).

Assessment of baseline comparability

Although tests of statistical significance to compare baseline variables are used in more than 75% in our survey, and in other surveys 50% to 61% (Altman & Dore, 1990; Assmann et al., 2000; Pocock et al., 1987; Schulz et al., 1994), these tests are not a useful way of assessing similarity. Assmann et al. (2000) argue that such significance testing is inappropriate. Tests of statistical significance to compare baseline variables assess indirectly whether randomization was adequate. If randomization is properly done, the null hypothesis that the two groups come from the same population is, by definition, true. Tests of statistical significance therefore calculate the probability of observing differences between groups, on a single characteristic, if chance were the only factor operating and, in any case, we should expect 5% of such comparisons to be significant at the 5% level. In our survey, of the authors who used hypothesis tests to compare baseline characteristics, 11% reached statistical significance at the 5% level, higher than the expected rate. In contrast, other surveys reported 2% and 6% of such tests reaching statistical significance (Assmann et al., 2000; Schulz et al., 1994).

Although one of the aims of randomization is to help produce balanced groups on baseline variables, undesirable differences between groups can nevertheless result, especially in small trials. The success of randomization to produce more balanced groups on baseline variables will be greater with increased sample size. Imbalances in important baseline characteristics can yield misleading results.

In the event that important baseline differences do exist, several factors will need to be considered to decide whether an adjustment for the target variables in the analysis will be needed, and how to make the adjustment. The prognostic strength of the variables that are imbalanced, the magnitude of the imbalances, and the degree to which adjustment for them or inclusion of them in the analysis model decreases bias of the estimated effect, should be the main elements for consideration (Altman, 1985). Even if the randomization is perfect, the inclusion of important prognostic factors as strata or in a multivariate analysis can substantially improve precision.

Statistical analysis

Tests of statistical significance were extensively used in our survey, statistical reporting of *p* values was frequently inadequate, and none of the reports surveyed made use of confidence intervals. In medical research, similar results have been shown by others, with 75% to 98% of the reports failing to provide magnitude of treatment differences with confidence intervals (Gotzsche, 1989; Pocock et al., 1987). Tests of statistical significance and corresponding *p* values are frequently overused and their results are often imperfectly understood (Altman, 1985; Altman & Dore, 1990; Assmann et al., 2000; Begg et al., 1996; Friedman et al., 1996; Gardner & Altman, 1986; Gotzsche, 1989; Liberati et al., 1986; Moher et al., 1994; Pocock et al., 1987; Schulz et al., 1994; Simon, 1986). In fact, *p* values reveal less information than what we ideally expect. For example, a very small *p* value (p=0.01) shows that the observed treatment difference would have little probability of occurring if chance alone was the cause. This may suggest that the target intervention is responsible for the observed treatment difference. The cutoff point of 0.05 allows the researchers to set a limit for judging this probability. Unfortunately, an "alpha" of 0.05 is sometimes falsely interpreted as a 5% chance of being wrong in one's decision. Worse still, it may be misinterpreted as the probability that the treatment does not work in 5% of patients. In fact, it is rather a 5% chance of concluding that there is a treatment difference when in fact there is none. In contrast to the limited value of *p* values, a confidence interval conveys information about both the magnitude of the treatment difference and the precision (random variability of this estimate) and, therefore, is preferable to *p* values (Rothman, 1986).

The confidence interval will reflect the random and unavoidable variability associated with any observed treatment difference. A confidence interval enables the reader to appreciate the range within which the true effect or treatment difference may plausibly lie. The upper limit of the confidence interval draws attention to the possibility that the treatment effect might be compatible with a true effect. This is especially useful in "negative" or inconclusive trials that failed to provide a significant treatment

difference, but where a clinically important difference may exist. Overall, these represent distinct advantages for using confidence intervals over just giving *p* values, usually dichotomized into "significant" or "non-significant". Confidence intervals are not a substitute method for significance testing. They convey information about *both* the strength of an association and the precision with which it is estimated. In contrast, tests of statistical significance do not distinguish between these two different concepts. They reflect the magnitude of the observed association and the sample size.

In recent years, several journals required the use of confidence intervals, and we have seen its use increased markedly. The International Committee of Medical Journal Editors, which designed the uniform requirements for manuscripts presentation, encourages investigators to avoid relying solely on statistical hypothesis testing, such as the use of *p* values and, instead, to quantify study findings and present them with indicators of measurements error or uncertainty, such as the confidence interval (International Committee of Medical Journal Editors, 1997). *P* values can be provided in addition to confidence intervals, but results should certainly not be reported solely with *p* values (Gardner & Altman, 1986).

Conclusion

We have reviewed methodological and statistical elements in the reporting of RCTs. Our survey showed that important information about methodology was commonly omitted in all three journals, and statistical reporting of test results was frequently inadequate and incomplete. Overall, the reports assessed in nursing research are comparable in terms of quality to those in medical research. The emphasis on the presentation of *p* values from hypothesis testing should be reduced and investigators should supplement or replace these tests with confidence intervals (Gardner & Altman, 1986; Rothman 1986). Overall, this survey stresses the importance of adequate methodologic quality in the trials and complete reporting of this methodology. Better quality in reports of RCTs will increase the standards of nursing research and impact on patient care. ♥

About the authors

Nicole Parent, RN, PhD, National Director of Research at the CCCN and coordinator of the surgical sector, Montreal Heart Institute.

James A. Hanley, PhD, Professor, Department of Epidemiology and Biostatistics, McGill University.

References

Altman, D.G. (1985). Comparability of randomized groups. *Statistician*, *34*, 125–136.

Altman, D.G., & Dore, C.J. (1990). Randomization and baseline comparisons in clinical trials. *Lancet*, *335*, 149–153.

Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001) The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine 134*, 663–694.

Assmann, S.F., Pocock, S.J., Enos, L.E., & Kasten, L.E. (2000). Subgroup analysis and other misuses of baseline data in clinical trials. *Lancet*, *355*, 1064–1069.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT Statement. *Journal of the American Medical Association, 276*, 637–639.

Friedman, L.M., Furberg, C.D., & DeMets, D.L. (1996). *Fundamentals of clinical trials* (3rd ed.). Mosby.

Gardner, M.J., & Altman, D.G. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal*, *292*, 746–750.

Gotzsche, P.C. (1989). Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal anti-inflammatory drugs in rheumatoid arthritis. *Controlled Clinical Trials*, *10*, 31–56.

International Committee of Medical Journal Editors. (1997). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine*, *126*, 36–47.

Liberati, A., Himel, H.N., & Chalmers, T.C. (1986). A quality assessment of randomized control trials of primary treatment of breast cancer. *Journal of Clinical Oncology*, *4*, 942–951.

Moher, D., Dulberg, C.S., & Wells, G.A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association*, *272*, 122–124.

Moher, D., Schulz, K.F., & Altman, D.G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Lancet*, *357*, 1191–1194.

Parent, N., & Hanley, J.A. (2000). Assessing quality of reports in randomized controlled trials in nursing journals. *Controlled Clinical Trials*, *21*(2S), 32S. Abstract presented at the 21st annual meeting of the Society for Clinical Trials. Toronto, Canada, April 16–20.

Piantadosi, S. (1997). *Clinical trials. A methodologic perspective*. New York: John Wiley & Sons, Inc.

Pocock, S.J., Hughes, M.D., & Lee, R.J. (1987). Statistical problems in the reporting of clinical trials. *New England Journal of Medicine*, *317*, 426–432.

Polit, D.F., & Hungler, B.P. (1991). Evaluating research reports. In *Nursing research: Principles and methods* (4th ed., pp. 583–596). Philadelphia: Lippincott.

Rothman, K.J. (1986). Significance questing. *Annals of Internal Medicine, 105*, 445–447.

Schulz, K.F., Chalmers, I., Grimes, D.A., & Altman, D.G. (1994). Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *Journal of the American Medical Association, 272*, 125–128.

Simon, R. (1986). Confidence intervals for reporting results of clinical trials. *Annals of Internal Medicine*, *105*, 429–435.