Assessing the Quality of Randomization From Reports of Controlled Trials Published in Obstetrics and Gynecology Journals

Kenneth F. Schulz, MBA; Iain Chalmers, MBBS, MSc; David A. Grimes, MD; Douglas G. Altman

Objective.—To assess the methodologic quality of approaches used to allocate participants to comparison groups in randomized controlled trials from one medical specialty.

Design.—Survey of published, parallel group randomized controlled trials.

Data Sources.—All 206 reports with allocation described as randomized from the 1990 and 1991 volumes of four journals of obstetrics and gynecology.

Main Outcome Measures.—Direct and indirect measures of the adequacy of randomization and baseline comparisons.

Results.—Only 32% of the reports described an adequate method for generating a sequence of random numbers, and only 23% contained information showing that steps had been taken to conceal assignment until the point of treatment allocation. A mere 9% described both sequence generation and allocation concealment. In reports of trials that had apparently used unrestricted randomization, the differences in sample sizes between treatment and control groups were much smaller than would be expected due to chance. In reports of trials in which hypothesis tests had been used to compare baseline characteristics, only 2% of reported test results were statistically significant, lower than the expected rate of 5%.

Conclusions.—Proper randomization is required to generate unbiased comparison groups in controlled trials, yet the reports in these journals usually provided inadequate or unacceptable information on treatment allocation. Additional analyses suggest that nonrandom manipulation of comparison groups and selective reporting of baseline comparisons may have occurred.

(*JAMA.* 1994;272:125-128)

RANDOMIZATION eliminates selection biases in controlled trials. Unfortunately, investigators often address randomization improperly in the design and implementation phases of trials and neglect it in published reports.[1-3] Moreover, an analysis of prominent general journals revealed that among trials in which unrestricted randomization was used, the sample sizes in the two comparison groups were more similar than would be expected by chance.[3] Furthermore, results of only 4% of hypothesis tests comparing baseline characteristics were significant at the 5% level.

We conducted a systematic evaluation of reports of randomized controlled trials (RCTs) published in the two main US and the two main British journals of obstetrics and gynecology. The *American Journal of Obstetrics and Gynecology* (*AJOG*) and *Obstetrics and Gynecology* (*OG*) are published in the United States, and the *British Journal of Obstetrics and Gynaecology* (*BJOG*) and the *Journal of Obstetrics and Gynaecology* (*JOG*) are published in the United Kingdom.

Earlier research has suggested that the methodologic quality of RCTs in this specialty may be inadequate[4-7]; we anticipated that descriptions of adequate approaches to treatment assignment would be rarer in these journals than in general journals. We also hypothesized that (1) the reports published in the *BJOG* would be of better quality than those published in the other three journals because a concerted editorial effort had been made to improve the quality of reporting in the *BJOG*,[8-10] (2) the numbers of patients in the comparison groups of trials in which unrestricted randomization was used would be more similar than would be expected by chance, and (3) the percentage of reported statistically significant differences in baseline characteristics would be less than the expected 5%.

METHODS

We collected data from all reports (N=206) of trials published in the 1990 and 1991 volumes of the *AJOG*, the *BJOG*, the *JOG*, and *OG*. To identify eligible reports, we handsearched the journals and then cross-checked that search using the Oxford Database of Perinatal Trials[11] (issue 8) and MEDLINE. We included articles in which authors reported that individuals had been randomly allocated to parallel (uncrossed) groups. A report was included as long as it purported to refer to a randomized trial, even if the actual method described nonrandom allocation. We included only the first publications relating to particular trials.

We examined reports and collected data using methods similar to those used in the analysis of general journals.[3] For consistency of measurement across journals, one of us (K.F.S.) performed all of the assessments. To examine the reproducibility of items on the questionnaire, another of us (D.A.G.) assessed a sample (random number table) of 15 trials while blinded to the initial assessments. We found no notable differences on our main outcome measures. We entered data into an Epi-Info questionnaire.[12]

The reduction of bias in trials depends crucially on preventing foreknowledge of treatment assignment. Concealing assignments until the point of allocation prevents foreknowledge, but that process has sometimes been confusingly referred to as *randomization blinding*.[13]

From the London (England) School of Hygiene and Tropical Medicine (Mr Schulz); The United Kingdom Cochrane Centre, Oxford, England (Mr Schulz and Dr Chalmers); the Division of STD/HIV Prevention, National Center for Prevention Services, Centers for Disease Control and Prevention, Atlanta, Ga (Mr Schulz); the Department of Obstetrics, Gynecology, and Reproductive Sciences, University of California–San Francisco (Dr Grimes); and the Medical Statistics Laboratory, Imperial Cancer Research Fund, London (Mr Altman).

Presented in part at the Second International Congress on Peer Review in Biomedical Publication, Chicago, Ill, September 10, 1993.

Reprint requests to the Division of STD/HIV Prevention, Centers for Disease Control and Prevention, NCPS, Mailstop E-02, Atlanta, Ga 30333 (Mr Schulz).

This term, if used at all, has seldom been distinguished clearly from other forms of blinding (masking) and is unsatisfactory for at least three reasons. First, the rationale for generating comparison groups at random, including the steps taken to conceal the assignment schedule, is to eliminate selection bias. By contrast, other forms of blinding, used after the assignment of treatments, serve primarily to reduce ascertainment bias. Second, from a practical standpoint, concealing treatment assignment up to the point of allocation is always possible, regardless of the study topic, whereas blinding after allocation is not attainable in many instances, such as in trials conducted to compare surgical and medical treatments. Third, control of selection bias pertains to the trial as a whole, and thus to all outcomes being compared, whereas control of ascertainment bias may be accomplished successfully for some outcomes but not for others. Thus, concealment up to the point of allocation of treatment and blinding after that point address different sources of bias and differ in their practicability. In light of those considerations, we refer to the former as *allocation concealment* and reserve the term *blinding* for measures taken to conceal group identity after allocation.

We considered the following approaches to the generation of an allocation sequence as adequate: computer, random number table, shuffled cards or tossed coins, and minimization. We considered the following approaches to allocation concealment as adequate: central randomization (eg, by telephone to a trials office), a pharmacy, numbered or coded containers, and sequentially numbered, opaque, sealed envelopes. Nonrandom (often called systematic) approaches included alternate assignment and assignment by odd/even birth date or hospital number. Other terms are described elsewhere.[14,15]

Restriction forces sample sizes in comparison groups to be more similar than would occur by simple randomization.[14] Blocking is the most commonly used form. Our analyses of the differences in reported sample sizes of comparison groups has been limited to two-group, unrestricted trials. We categorized trials as "unrestricted" if the trial had not been reported as restricted or stratified (which are more likely to be restricted).

To assess whether authors reported appropriate measures of variability for means or medians when reporting baseline comparisons, we looked for the SD, range, or raw data. Unless otherwise indicated, we used χ^2 tests to compare nominally scaled variables. The Greenland and Robins approach was used to obtain confidence intervals for relative risks.[12]

Adequacy of Reported Methodologic Components of Randomization in 206 Randomized Controlled Trials From Four Journals

Randomization Component	% of Trials				
	Am J Obstet Gynecol (n=64)	*Br J Obstet Gynaecol* (n=48)	*J Obstet Gynaecol* (n=20)	*Obstet Gynecol* (n=74)	Total (N=206)
Use of adequate sequence generation	36	38	15	30	32
Use of adequate allocation concealment	19	44	5	19	23
Both generation and concealment adequate	9	15	5	7	9

RESULTS

We found 206 reports of trials in four journals. More than three quarters (78%) failed to provide information about the type of randomization. Despite purporting to be randomized trials, 11 reports (5%) described the use of a nonrandom method of assignment. Only 29 (14%) of the reports described the use of restriction (23 of the 29 descibed blocking). None reported the use of replacement randomization. Reports published in the *BJOG* stated the type of randomization more frequently than reports published in the other journals (48% vs 14%, $P<.001$, 1 *df*).

Only 32% of the reports specified an adequate method for generating random numbers, and the rates were similar among the four journals ($P=.27$, 3 *df*; Table). A computer random number generator was the most frequently specified method (18%), followed by a random number table (11%).

Almost half (48%) of the reports did not describe the mechanism used to allocate treatments. Authors specified use of envelopes most frequently (25%), but only one quarter of those (6% of all) stated that the envelopes had been sequentially numbered, opaque, and sealed. Fifteen reports (7%) specified the pharmacy, another 15 (7%) specified numbered bottles or containers, and five (2%) described central randomization. Ten (5%) stated that a list, table, or schedule had been used for allocation, and the other 11 used nonrandom, unconcealed assignment. Overall, only 23% stated an adequate approach to allocation concealment (Table). The proportion of reports describing adequate concealment varied markedly among the four journals ($P<.001$, 3 *df*). The *BJOG* had a rate 2.6 times higher than the other three journals combined (95% confidence interval, 1.6 to 4.1; $P<.001$). Only 9% described an adequate method for both sequence generation and allocation concealment (Table).

In 96 reports of apparently unrestricted trials, sample sizes of the treatment and control groups differed by less than would be expected due to chance alone. The Figure illustrates those differences in relation to total trial size. About five trials should fall outside the outer pair of straight lines—none did; about 48 should fall outside the inner pair of lines—only eight did ($P<.001$; χ^2 goodness of fit, 2 *df*). That 54% of the unrestricted trials had differences in group sizes of zero or one further indicates the similarity of group sizes. Surprisingly, only 36% of the blocked trials had differences in group sizes of zero or one.

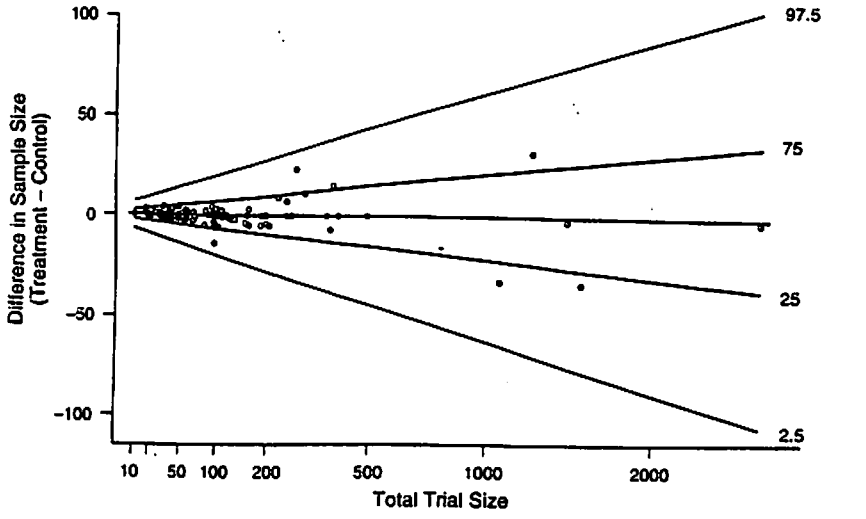
Authors presented comparisons of baseline characteristics in 84% of the reports. Comparisons presented as continuous variables were reported in 78% of the trials, and among those only 68% were accompanied by appropriate measures of variability. Reports in the *BJOG* were more likely than those in the other three specialty journals to present appropriate measures of variability, but the differences among the four journals were not statistically significant ($P=.22$, 3 *df*). In 41% of the 206 reports, either authors did not present baseline characteristics or did not report appropriate measures of variability.

Authors used hypothesis tests for baseline comparisons in 125 reports (61%) and presented results of 1076 tests. Of those results, only 2% were statistically significant at the 5% level, a departure from expectation ($P<.001$, z test).

In 50 (24%) of the reports, authors reported sample sizes to be based on prior statistical power calculations. The rates were 0% for the *JOG*, 18% for *OG*, 19% for the *AJOG*, and 52% for the *BJOG*. Trials published in the *BJOG* reported power calculations 3.3 times more frequently than those from the other three journals combined (95% confidence interval, 2.1 to 5.2; $P<.001$).

COMMENT

Randomized controlled trials provide the most valid basis for the comparison of interventions in health care. If im-



Relationship between the difference in numbers in treatment and control groups and total trial size for 98 unrestricted trials. Straight lines represent the expected distributions for 50% (inner) and 95% (outer) prediction intervals.[3]

properly conducted, however, trials purporting to be "randomized" can yield biased results. Indeed, bias has been detected in trials not reporting adequate allocation concealment.[13] Thus, for readers to have justifiable confidence in the internal validity of a trial, the report should demonstrate adequate randomization. Considering its central importance, we are surprised that authors have not been more meticulous in publishing clear reports of the randomization process.

Our estimate of 32% for adequate sequence generation may be generous, as it includes processes such as shuffled cards and tossed coins as "adequate." Because those methods open the production of assignment schedules to human perturbations and result in unreproducible results, we consider them less than optimal; others consider them unacceptable.[15] We recommend tables and computers not only because of reproducibility but also because of ease and speed.

Allocation concealment generally outweighs the importance of generating assignments per se,[16,17] yet only about half of the reports provided information adequate to assess that aspect of trial design and conduct. We judged that less than one quarter of the reports described adequate allocation concealment, but, even with many of those, further clarifying information should have been provided. Few reports stated who had prepared the randomization scheme; those who prepared the scheme should not have been involved in determining eligibility, administering treatment, or assessing outcome.

Reports of trials published in the *BJOG* provided more information than those in the other three journals. They more frequently included information about the type of randomization, reported an adequate approach to randomization concealment, and reported statistical power calculations. Also, the quality of reports in the *BJOG* matched or exceeded that found in the four general journals.[3] Even so, those of us (I.C. and D.G.A.) who had been involved in editorial efforts to improve the quality of reports in the *BJOG* were disappointed at how much room for improvement remains. The reports from the two US journals were comparable with each other but superior to those in the *JOG*. Editorial efforts similar to those made at the *BJOG* in the mid-1980s are now occurring at *OG*,[18] and those, too, may result in improved quality.

The relative sizes of comparison groups in the unrestricted trials should have reflected random variation. In other words, some discrepancy between the numbers in the comparison groups would be expected. We found the contrary, however, which supported an earlier finding.[3] The strong tendency for the comparison groups to be of equal or similar sizes may be explained by unreported use of (1) restriction, usually blocking; (2) replacement randomization; (3) a nonrandom method of assignment; or (4) nonrandom manipulation of assignments or data to balance sample sizes. Use of restriction would be the most palatable of these possible explanations, and it likely explains some instances. It probably does not explain most, however, because few trials reported restriction and because blocked trials yielded differences more disparate than those found in the unrestricted trials. We found no evidence of replacement randomization. We found evidence of nonrandom allocation; thus, its unidentified use in other trials may explain some of the similarities. This is hardly reassuring, however, given the risk of bias due to nonrandomness and difficulties with concealment.

The fourth potential explanation, nonrandom manipulation, has serious implications because it is the most likely to introduce selection bias. Our findings provide indirect evidence that it could have happened. Some investigators may have believed that they would increase the credibility of their trial if they presented comparison groups of equal size. Unfortunately for good science, but fortunately for those investigators, most readers probably shared their misconception. Paradoxically, the results of those possible manipulations have had exactly the opposite effect when analyzed in aggregate in our study. While our results indicate clearly that the set of trials that had supposedly used unrestricted randomization were not what they purported to be, the identification of any particular trial as suspect is impossible, as some trials would be expected to achieve similar numbers simply by chance.

While randomization assigns treatments without bias, it does not necessarily produce balanced groups with respect to prognostic factors. On strictly theoretical grounds, if randomization is properly implemented, establishment of comparability at baseline is unnecessary. Random assignment eliminates bias, even though, in a particular study, the groups compared may never be perfectly balanced for important prognostic variables. The process of randomization underlies significance testing, and that process is independent of prognostic factors, known or unknown.[19]

Baseline characteristics in RCTs should be addressed by authors, but the common, inappropriate use of hypothesis tests to compare characteristics concerns us.[20,21] That process assesses the probability that differences observed could have occurred by chance. In properly randomized trials, however, any observed differences have occurred by chance. As noted elsewhere,[21] "Such a procedure is clearly absurd." Hypothesis tests are superfluous, and their use in comparisons of baseline characteristics can mislead investigators and their readers. Rather, comparisons should be based on consideration of the prognostic strength of the variables measured and the magnitude of any chance imbalances that have occurred.[21]

Hypothesis tests in these reports resulted in many fewer statistically significant comparisons than expected. One plausible explanation for that discrepancy is that a few investigators may have decided not to report statistically significant comparisons, believing that by withholding that information they would increase the credibility of their reports. In fact, the opposite has occurred in this aggregated analysis. Investigators should report baseline comparisons on important prognostic variables, regardless of statistical significance. Not only are hypothesis tests superfluous and potentially misleading, but they can be harmful if they lead investigators to suppress any baseline imbalances.

None of our findings are particularly reassuring. As a whole, the trials from this medical specialty fared somewhat worse than the poor showing of the general journals.[3] Furthermore, these results probably represent what would be found in many other specialties as well. Although failure to report steps to reduce bias does not constitute direct evidence that those steps have not been taken, at least one study, in which clarification was sought from the authors of reports, has shown that inadequate reporting usually reflects inadequate methods.[22] Thus, while reporting must clearly be improved, deficiencies in the design and conduct of trials must also be addressed.

Omission of randomization details to date has probably been primarily an author-based phenomenon rather than a result of journal editors extracting important material from manuscripts. Moreover, refereeing and editorial work cannot improve what was actually done in a trial—only how well it was reported. Thus, the burden for improvement should fall primarily on investigators, although editors could stimulate that process.

Protestations from authors about lack of space do not constitute acceptable excuses for omission. Space will always be a limitation; the issue is the relative importance of the topics addressed. Authors frequently include information that has little bearing on scientific validity, while they omit critical elements of the randomization process. In a double-blind RCT, however, aspects other than randomization may be scientifically inconsequential to the analysis because they would have been applied equally to unbiased comparison groups. Certainly, we would not wish to promote a cavalier attitude toward other methodologic elements: surely some have to be adequately described for readers to interpret findings and extrapolate results. Yet, proper reporting of the randomization procedures should have the highest priority, and those trials that fail to provide such information should be interpreted cautiously.

At a minimum, reports of RCTs should include descriptions of (1) the type of randomization, (2) the method of sequence generation, (3) the method of allocation concealment, (4) the persons generating and executing the scheme, and (5) the comparative baseline characteristics, with proper interpretation. Furthermore, tolerance for groups of unequal sizes in unrestricted trials should be cultivated in addition to intolerance for hypothesis testing of baseline characteristics.

This work was supported by the Centers for Disease Control and Prevention. Dr Chalmers was supported by the National Health Service Research and Development Programme.

References

1. Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled trials: agenda for the editor. *Controlled Clin Trials.* 1980;1:37-58.
2. DerSimonian R, Charette LJ, McPeek B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med.* 1982;306:1332-1337.
3. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. *Lancet.* 1990;335:149-153.
4. Tyson JE, Furzan JA, Reisch JS, Mize SG. An evaluation of the quality of therapeutic studies in perinatal medicine. *J Pediatr.* 1983;102:10-13.
5. Thacker SB. The efficacy of intrapartum electronic fetal monitoring. *Am J Obstet Gynecol.* 1987;156:24-30.
6. Keirse MJNC. Amniotomy or oxytocin for induction of labor: re-analysis of a randomized controlled trial. *Acta Obstet Gynecol Scand.* 1988;67:731-735.
7. Grimes DA, Schulz KF. Randomized controlled trials of home uterine activity monitoring: a review and critique. *Obstet Gynecol.* 1992;79:137-142.
8. Bracken MB. Reporting observational studies. *Br J Obstet Gynaecol.* 1989;96:383-388.
9. Wald N, Cuckle H. Reporting the assessment of screening and diagnostic tests. *Br J Obstet Gynaecol.* 1989;96:389-396.
10. Grant A. Reporting controlled trials. *Br J Obstet Gynaecol.* 1989;96:397-400.
11. Chalmers I, Hetherington J, Newdick M, et al. The Oxford Database of Perinatal Trials: developing a register of published reports of controlled trials. *Controlled Clin Trials.* 1986;7:306-324.
12. Dean AG, Dean JA, Burton AH, Dicker RC. *Epi Info Version 5: A Word Processing, Database, and Statistics Program for Epidemiology on Microcomputers.* Atlanta, Ga: Centers for Disease Control; 1990.
13. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med.* 1983;309:1358-1361.
14. Pocock SJ. *Clinical Trials: A Practical Approach.* Chichester, England: John Wiley & Sons; 1983.
15. Meinert CL. *Clinical Trials: Design, Conduct, and Analysis.* New York, NY: Oxford University Press; 1986.
16. Hill AB. The clinical trial. *N Engl J Med.* 1952;247:113-119.
17. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline, I: control of bias and comparison with large co-operative trials. *Stat Med.* 1987;6:315-325.
18. Grimes DA. Randomized controlled trials: 'it ain't necessarily so.' *Obstet Gynecol.* 1991;78:703-704.
19. Fisher RA. *The Design of Experiments.* 8th ed. Edinburgh, Scotland: Oliver & Boyd Ltd; 1966.
20. Rothman KJ. Epidemiologic methods in clinical trials. *Cancer.* 1977;39 (suppl 4):1771-1775.
21. Altman DG. Comparability of randomised groups. *Statistician.* 1985;34:125-136.
22. Liberati A, Himel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. *J Clin Oncol.* 1986;4:942-951.

Printed and Published in the United States of America