The Online Journal of

CURRENT CLINICAL TRIALS

Volume 4, Number 197 950826

The Methodologic Quality of Randomization

Schulz KF, Chalmers I, Altman DG, Grimes DAS, Doré CJ, The methodologic quality of randomization as assessed from reports of trials in specialist and general medical journals [article]. Online J Curr Clin Trials [serial online] 1995 August 28; 4(Doc No 197);[81 paragraphs]. 2 figures; 16 tables.

THE METHODOLOGIC QUALITY OF RANDOMIZATION AS ASSESSED FROM REPORTS OF TRIALS IN SPECIALIST AND GENERAL MEDICAL JOURNALS

Kenneth F. Schulz, Iain Chalmers, Douglas G. Altman, David A.Grimes Caroline J. Doré

Division of STD/HIV Prevention, National Center for Prevention Services, Centers for Disease Control and Prevention, Atlanta, GA 30333 USA

Document Classification: Research Report

Keywords: randomization, randomized trials, journals, specialized, journals, general, obstetrics and gynecology

Date of Publication: 19950826

ABSTRACT

(1) **Objective:** To assess the quality of randomization from reports of trials in a sample of specialist journals, and to compare those results with a similar assessment from a sample of general medical journals.

(2) **Design:** Evaluation of all 206 reports of parallel-group randomized trials published in the 1990 and 1991 volumes of four journals of obstetrics and gynecology and of 81 reports of trials published during 1987 in four general medical journals.

(3) **Results:** Of the reports published in the specialist and in the general medical journals, only 32% and 48%, respectively, reported having used an adequate method to generate random numbers; only 23% and 26%, respectively, contained information showing that steps had been taken to conceal assignment until the point of treatment allocation; and merely 9% and 15%, respectively, described adequate methods of both sequence generation and allocation concealment. In those 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 by chance, and that feature was more marked in the specialist journals. In reports of trials in which hypothesis tests had been used to compare baseline characteristics, only 2% of tests reported in specialist journals and 4% of tests reported in general journals were statistically significant, lower than the expected rate of 5%.

(4) **Conclusions:** Generating unbiased comparison groups requires proper randomization, yet the reports in these specialist and general journals usually provided inadequate or unacceptable information. Additional analyses suggest that nonrandom manipulation of comparison groups and selective reporting of baseline com-

parisons may have occurred.

INTRODUCTION

(5) Randomization avoids selection biases in controlled trials of prevention and treatment. Forty years ago, Austin Bradford Hill[1] wrote that "having used a random allocation, the sternest critic is unable to say when we eventually dash into print that quite probably the groups were differentially biased through our predilections or through our stupidity." Unfortunately, the process of randomization is often improperly addressed in the design and implementation phases of controlled trials, and it is often neglected in published reports. For example, in 132 reports of trials on cancer topics, only a third of the authors reported how the randomization had been carried out and many of the methods specified were, in fact, nonrandom[2].

(6) Even in some of the most highly regarded medical journals, the quality of reporting leaves considerable room for improvement. Less than a fifth of the reports of clinical trials published during 1979–80 in the *New England Journal of Medicine* (*NEJM*), the *Lancet*, the *Journal of the American Medical Association*, and the *British Medical Journal* (*BMJ*) described the method of randomization[3]. An analysis by Altman and Doré of reports of trials published during 1987–8 in the *BMJ*, the *Lancet*, the *NEJM*, and the *Annals of Internal Medicine* (*Annals*) revealed that only 34% of the articles specified both the method used to generate random numbers and the mechanism used to allocate treatments; and even when methods were specified, they were often not methodologically sound[4]. The analysis also revealed that 49% of trials had reported baseline data unsatisfactorily, and that 58% had inappropriately used hypothesis tests to compare baseline variables[4].

(7) Altman and Doré's survey of widely read and highly regarded general medical journals prompted a suggestion that the standard of reporting in specialist medical journals was likely to be even worse[5]. To investigate that possibility, we undertook a similar study of reports of randomized controlled trials (RCTs) published in a sample of specialist journals in one field—obstetrics and gynecology. As far as we are aware, no systematic analysis of randomization and allocation in a sample of journals in this specialty had been undertaken prior to this work, although assessments of reports of trials in this field had indicated that the methodological quality may indeed be worse than that of reports published in general medical journals[6-9]. We have already reported the major findings of our research[10]. This paper presents the complete findings and contrasts them with those from general journals[4]. It also provides a more extensive discussion of randomization concepts.

(8) We conducted a systematic evaluation of reports of RCTs published in the two main U.S. 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[10].

(9) We analyzed the reported approaches to treatment assignment and to comparison

of baseline characteristics. First, as indicated above, we suspected that reports in the specialist journals would be of lower quality than reports published in the general medical journals. Second, we thought that the reports published in the *BJOG* would be of better quality than those published in other journals of obstetrics and gynecology, since a concerted editorial effort had been made to improve the quality of trials reporting in that journal[10,11]. Third, we postulated that, as had been demonstrated using reports of trials published in the general medical journals, the numbers of patients in the comparison groups of trials that had apparently used simple randomization would too often be similar[4,10]. Fourth, we suspected that the percentage of statistically significant differences in characteristics measured at baseline would be less than the expected 5%, as had been suggested by the report from the general medical journals[4,10].

METHODS

(10) We collected data from 206 reports of trials published in the 1990 and 1991 volumes of the *AJOG*, the *BJOG*, the *JOG*, and *OG*, using a handsearch to try to ensure that we identified all the eligible reports[10]. In addition, we searched both the Oxford Database of Perinatal Trials[12] (issue 8) and MEDLINE as a cross-check. Our study was restricted to reports of parallel (uncrossed) group trials, comparing two or more treatments, in which allocation was stated to have been randomized. Initial selection was based on the abstract and a cursory inspection of the main text. A report was included as long as it purported to refer to a randomized trial, even if the actual method of allocation described was nonrandom. We included only the first publications relating to particular trials[10].

(11) We examined reports and collected data with methods similar to those used in Altman and Doré's[4] analysis of general medical journals. We tested the data collection instrument in a pilot study of the 1989 volumes of the same journals. For consistency of measurements across journals, one of us (KF Schulz) performed all of the assessments[10]. To examine the reproducibility of items on the questionnaire, in an approximate way, another of us (DA Grimes) assessed a sample (random number table) of 15 trials while unaware of the initial assessments. We found no notable differences on our main outcome measures[10]. The data were entered interactively into an EPI-INFO questionnaire with on-line editing and checking capability[13].

(12) Restriction forces the 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 numbers of participants reported to have been assigned to comparison groups have been limited to two-group trials that were apparently "unrestricted." We categorized trials as "unrestricted" if they met all the following criteria: (1) the trial had not been reported to have been restricted; (2) the type of randomization for the trial had either not been stated, or had been stated to have been "simple" or "unrestricted"; and (3) the trial had not been reported to have been "stratified" (since stratified trials are more likely to be restricted)[10]. We only included trials in this analysis report in which the authors provided the comparison group sizes at randomization.

(13) In assessing the results of hypothesis tests of baseline characteristics, the level of significance was assumed to have been 0.05 if it had not been stated explicitly. We did not test data for which the authors had not presented test results. If some baseline characteristics within a particular report had variability presented and others did not, our assessment was based on the method used when variability had been presented.

(14) If means or medians are reported for continuous baseline characteristics, appropriate information about variability should also be reported, e.g., the standard deviation, range, or raw data[10]. We counted a report as providing appropriate information about variability if the authors presented it on at least one continuous variable. We used this definition consistently for both specialist and general journals. The definition for the use of appropriate information about variability in the original report of the general journals[4] required that the appropriate information be reported for all the continuous variables.

(15) Unless otherwise indicated, chi-squared tests were used for comparing nominally scaled variables[10]. The Greenland and Robins approach was used to obtain confidence intervals for relative risks[13]. Because Bartlett's test for homogeneity of variance was typically statistically significant at $p < 0.05$, we used Kruskal–Wallis one-way analysis of variance tests to compare continuous variables among journals. The Kruskal–Wallis test is a nonparametric test that does not assume normal distributions, yet it retains most of the power of a parametric test.

(16) To facilitate direct comparison of the reports published in the specialist journals with those published in the general journals, we have reproduced data presented in Altman and Doré's[4] article, together with some information not presented in the original publication. In particular, the one report of a trial that had used deterministic allocation and that was excluded from that paper[4] was included in parts of this analysis for conformity with our analysis of randomization in the specialty journals.

Terminology

(17) The *only* unique strength of randomization as an element in the design of treatment comparisons is that, if successfully accomplished, it prevents selection bias in allocation of treatments. Its success in this respect depends on fulfilling two interrelated, prior conditions. First, a schedule, based on some chance (random) process, must be generated for assigning people to comparison groups in the trial. Second, steps must be taken to secure strict implementation of that schedule of random assignments. Generating a schedule of random assignments presents fewer problems than ensuring strict adherence to it. The key to achieving strict adherence to a schedule of random allocation is to prevent foreknowledge of treatment assignments among those involved in recruiting participants to the trial.

(18) **Simple (unrestricted) randomization** can be achieved using one of the long-established methods of "drawing lots," such as repeated coin-tossing, throwing dice, or dealing previously shuffled cards. More commonly, it is achieved by referring to a printed list of random numbers or to a list of random assignments generated by a computer. In trials using large samples, simple randomization usu-

ally generates unbiased comparison groups of relatively similar sizes. In trials using small samples, simple randomization, although generating unbiased comparison groups, will sometimes result in groups that differ quite substantially in size. With simple randomization, 5% of trials result in an imbalance in size greater than $1.96\sqrt{N}$ (where N is the total study size)[4].

(19) **Replacement randomization** involves preparing a randomization list as in simple randomization and then checking it for a prespecified inequality in treatment numbers[14]. Based on criteria delineated *a priori*, if the inequality is sufficiently large, then a whole new simple randomization list is generated to replace the first one. While this procedure may seem inappropriate to some, as long as it happens before the trial starts, it is appropriate and has some good properties[14].

(20) **Balanced (restricted) randomization** is used to ensure not only that comparison groups will be unbiased, but also that they will be of approximately the same size. The most frequently used method of achieving balanced randomization is by "blocking." Blocking ensures that the numbers of participants to be assigned to each of the comparison groups will be balanced within blocks of, say, every 10 consecutively entered participants. The block size may remain fixed throughout the trial or it may be randomly varied, to reduce the likelihood of foreknowledge of treatment assignment among those recruiting participants. If the blocking remains fixed, that increases the chances of the concealment failing.

(21) Other approaches to restricted randomization include the "biased-coin" and "replacement-randomization" methods[14]. Restricted randomization is also used to generate separate randomization schedules for subsets of participants defined by potentially important prognostic factors (i.e. stratified randomization), such as disease severity and study centers. Another good approach that incorporates both the general concepts of stratification and restricted randomization is minimization, which can be used to make small groups closely similar with respect to several characteristics[14].

(22) Another restricted randomization method is the "restricted-shuffled" approach. It involves determining the desired sample size, apportioning the number of specially prepared cards for each treatment according to the allocation ratio, inserting the cards into opaque envelopes, sealing, and shuffling to produce a form of random assignment. The envelopes should then be sequentially numbered. This approach is less than optimal because shuffling determines the assignment sequence, one large block makes the later assignments obvious, and envelopes (themselves less than optimal) conceal the assignment. Nonetheless, the restricted shuffled approach can produce satisfactory allocations in many situations.

(23) **Deterministic (systematic) methods of assignment** may at first glance appear to be reasonable, but fail under closer scrutiny for both theoretical and practical reasons. These methods, such as assignment based on date of birth, case record number, date of presentation, or an odd or even number in the order of presentation in a consecutive series of participants, are just not random. Sometimes they are referred to as "quasi-random," but even that may give a falsely optimistic impres-

sion. For example, in some populations, the day of the week on which birth occurs is not a matter of chance[15]. While an element of chance is certainly involved in some of these approaches, to confirm that the assignments were at least close to being random would entail undertaking a separate, more time-consuming study than the primary substantive study planned.

(24) The most important weakness with deterministic methods is that concealing the basis for the assignment schedule is usually impossible, which allows foreknowledge of treatment assignment among those recruiting participants to the trial. For those reasons the *British Medical Joumal* has decided not to publish trials that have used such allocation schemes when randomization was feasible[16]. If authors report using nonrandom methods of allocation, readers should look askance at the results. Using an appropriate, random approach is easier in the short and long run, and is reproducible.

(25) **Preventing foreknowledge of treatment assignment** is a crucially important element of trial design. When assessing a potential participant's eligibility for a trial, those responsible for recruiting participants should remain unaware of the next assignment in the sequence until after the decision about eligibility has been made. Then, after the assignment has been made, they should not be able to alter the assignment or the decision about eligibility. The ideal is for the process to be impervious to any influence by the individual making the allocation. This condition is most likely to be achieved if an assignment schedule generated using true randomization is administered by someone who is not responsible for recruiting participants, for example, someone based in a trial office, or pharmacy. If organizing "central randomization" in that way is not possible, then other precautions are required to try to prevent manipulation of the schedule of random assignment by those recruiting participants to the trial. These include, for example, using numbered or coded bottles, ampoules, or other containers, and using serially numbered, sealed, opaque envelopes (all three attributes required)[2,4,14]. Simply using an open list ("table" or "schedule") of random numbers is as open to manipulation as is dependence on one of the deterministic, nonrandom methods of assignment.

(26) The process of concealing treatment assignment until after a decision has been made to enter a participant into a trial has sometimes been referred to as "randomization blinding"[17]. While those authors perceptively and appropriately coined a term for the process of preventing foreknowledge, that term, if used at all, has seldom been distinguished clearly from other forms of blinding (masking). We believe that the terminology should be clarified 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 comparing surgical with medical treatments. Third, control of selection bias pertains to the trial as a whole, and thus to all outcomes being com-

pared, whereas control of ascertainment bias may be accomplished successfully for some outcomes, but not for others. Thus, concealing 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[10].

(27) **Comparison of baseline characteristics** of the treatment groups is an important first step in trial reporting. Although randomization assigns treatments without selection bias, it does not necessarily produce groups that are similar in important prognostic factors[10]. **Chance imbalances** can and do occur. The probabilistic argument is that, on average, randomized groups will have the same characteristics. In practice, however, a particular trial is likely to have one or more characteristics unequally split between groups. Large studies generate serious imbalances less frequently, but smaller studies using simple randomization are susceptible to substantial covariate imbalances[18].

(28) Such imbalances in baseline characteristics cause concern, however, only when they involve characteristics of prognostic importance. If the imbalance is substantial, they can be confounding variables, albeit by chance, but confounding nonetheless. Testing for statistically significant differences (hypothesis testing) is not a valid basis on which to assess comparability in respect to baseline characteristics. Comparability must be assessed in terms of the prognostic strength of the variables and the magnitude of any imbalance[18-20].

RESULTS

(29) **Source of reports:** Of the 206 reports of trials published in the specialist journals, 64 were found in the *AJOG*, 48 in the *BJOG*, 20 in the *JOG*, and 74 in *OG*[10]. Of the 81 eligible reports of trials published in general medical journals, 80 were those analyzed by Altman and Doré, who had selected the first 20 reports to be published after 1 January 1987 in each of the *Annals of Internal Medicine* (*Annals*), the *British Medical Journal* (*BMJ*), the *Lancet*, and the *New England Journal of Medicine* (*NEJM*)[4]. The remaining report, published in the *NEJM*, was a trial which, although purporting to be randomized, had actually used a deterministic method of allocation. Other trials in the both the specialist and general medical journals used deterministic methods, but they were not included because they did not purport to be randomized.

(30) **Type of randomization:** Over three-quarters (78%) of the reports of trials published in the specialist journals failed to provide information about the type of randomization (Table 1A). Moreover, 11 reports (5%), about a quarter of those providing any information at all, clearly state that a deterministic (nonrandom) method of assignment had been used, despite their claims to be reporting randomized trials[10].

(31) Only 29 (14%) of the reports in the specialist journals described the use of restriction

[10], and of the 23 reports mentioning the use of blocking, only 15 (65%) stated the size of blocks. In the remaining reports of trials that had used restriction, four

had used a restricted shuffled approach, one the biased coin method, and one minimization. No reports stated the use of replacement randomization[10]. Only four trial reports stated that simple (unrestricted) randomization had been used, but a majority of the trials in which the approach had not been stated explicitly had probably actually used simple randomization.

(32) Among the four specialist journals, reports published in the *BJOG* stated the type of randomization more frequently than reports published in the other journals ($p <$ 0.001, 3 df)[10]. The differences among the other three journals in this respect were not statistically significant (p = 0.36, 2 df). Reports published in the *BJOG* also more frequently reported using a restricted approach to randomization ($p <$ 0.001, 3 df).

(33) Overall, the quality of reporting in the general journals[4] was marginally better than that in the specialist journals: 69% of the reports provided no information about the type of randomization (Table 1B) as compared to the 78% from the specialist journals. Reports from the general journals were more likely to report a restricted approach and less frequently reported nonrandom methods of assignment than those published in the specialist journals. None of the general journals, however, reported "type of randomization" and "restriction" as frequently as these details were provided in reports published in the *BJOG*.

(34) **Stratification:** Only 9% of the reports of trials in the specialist journals reported the use of stratification (Table 2A), and fewer than half of those reported the use of blocking or minimization. By contrast, 39% of the reports in the general medical journals reported the use of stratification (Table 2B), but, as in the specialist journals, only about half the reports mentioned the use of blocking or minimization[4].

(35) **Methods for generating random numbers:** Only 32% of reports published in the specialist journals specified an adequate method for generating random numbers, with the rates being similar for the four journals (p = 0.27, 3 df)(Table 3A)[10]. A computer random number generator was the most frequently specified method (18%), followed by a random number table (11%)[10]. Other random processes used in 3% of the trials included shuffled cards and tossed coins.

(36) A higher proportion (48%) of the reports published in the general medical journals reported an acceptable approach to generating random numbers (Table 3B)[4]. As with the specialist journals, a computer random number generator (23%) and a random number table (20%) were the most frequently specified methods[4].

(37) **Treatment allocation methods:** Almost half (48%) of the reports of trials in the specialist journals, and a somewhat lower proportion (44%) of those published in the general medical journals, did not describe the mechanism used to allocate treatments (Table 4A and Table 4C)[4,10]. A quarter of the reports in the specialist journals described the use of envelopes, but only a quarter of those reports stated that the envelopes had been sequentially numbered, opaque, and sealed[10]. Fifteen trials specified that the allocation had been prepared by the pharmacy, another 15 that numbered bottles or containers had been used, and 5 that a form of central randomization had been organized[10]. Five percent of the reports stated that a list,

table, or schedule had been used for allocation; in a further five percent, some form of deterministic assignment procedure had been used[10].

(38) Overall, only 23% of the reports published in the specialist journals (Table 4A) and 26% in the general journals (Table 4C) reported an adequate approach to allocation concealment[4,10]. The proportion of trials in which adequate allocation concealment appeared to have been achieved varied markedly ($p < 0.001$, 3 df) among the four specialist journals (Table 4A). The *BJOG* had a rate that was 2.6 times higher than the other three combined (95% CI 1.6–4.1, $p = 0.001$), a rate which more than matched the highest rate among the general medical journals (Table 4C)[10].

(39) **Overall quality of randomization and allocation:** Fifty-one reports of trials (25%) published in the specialist journals included information both on the method used to generate random numbers and on the mechanism used to allocate treatment, but only 19 (9%) described both an adequate method of generating random numbers and an adequate method of allocation concealment[10]. The proportions for each of the four specialist journals were 15% for the *BJOG*, 9% for the *AJOG*, 7% for *OG*, and 5% for the *JOG*, but the differences among those proportions were not statistically significant ($p = 0.46$, 3 df).

(40) Twenty-seven reports of trials (34%) published in the general medical journals included information both on the method used to generate random numbers and on the mechanism used to allocate treatment, but only 12 (15%) described both an adequate method of generating random numbers and an adequate method of allocation concealment. The proportions for each of the four general medical journals were 25% for the *Annals*, 0% for the *BMJ*, 10% for the *Lancet*, and 24% for the *NEJM*[4].

(41) **Relative size of treatment groups at the time of randomization in apparently unrestricted trials:** In the 96 reports of apparently unrestricted trials published in the specialist journals, the differences in sample sizes between the treatment and control groups were much smaller than would be expected by chance alone[10]. In Figure 1, 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 8 did. The differences in group sizes were much smaller than would be expected by the play of chance ($p < 0.001$, Chi-squared goodness-of-fit, 2 df). The discrepancy between the observed and expected differences in 43 reports published in the general medical journals was similar, but less marked (Figure 2). A further indicator of the similarity of group sizes in the specialty journals is that 54% of the unrestricted trials had differences in group sizes of zero or one (45% in the general journals). Surprisingly, the blocked trials in the specialty journals yielded differences that were less similar overall, with 36% of the trials having differences in group sizes of zero or one[10].

(42) **Comparisons of Baseline Characteristics:** Comparisons of baseline characteristics were presented in 84% of the reports published in the specialist journals (Table 5A), and in 92% of those published in the general medical journals (Table 5C)[4,10]. Among the specialist journals, comparisons of baseline characteristics were most often presented in the *BJOG*, least often in the *JOG*, with reports in the *AJOG* and

OG having intermediate and similar rates. However, those differences were not statistically significant ($p = 0.17$; 3 df).

(43) The median numbers of comparisons of baseline characteristics in those reports in which comparisons were presented was 6 in the specialist journals and 9 in the general medical journals. Among the specialist journals, reports in the *AJOG* and *OG* tended to present a larger number of comparisons ($p = 0.008$, 3 df) (Table 5A).

(44) Comparisons of baseline characteristics presented as continuous variables were reported in 78% of the reports published in the specialist journals (Table 5A), and in 86% of the reports published in the general medical journals (Table 5C)[4,10]. The frequency of reporting of appropriate measures of variability (of those presenting at least one continuous variable) was similar in the specialist journals (68%) and general medical journals (67%). Reports in the *BJOG* were more likely than those in the other specialist journals to present appropriate measures of variability, but the differences among the four specialist journals were not statistically significant ($p = 0.22$; 3 df).[10] Overall, either authors did not present baseline characteristics or did not report an appropriate measure of variability in at least one instance in 41% of the reports published in specialist medical journals and 36% of reports published in general medical journals[4,10].

(45) **Use of hypothesis tests to compare baseline characteristics:** Hypothesis tests were used to compare baseline characteristics in 61% of the reports of trials published in the specialist journals (Table 6A) and 58% of the reports published in general medical journals (Table 6B)[4,10]. Hypothesis tests were presented more often in the American specialist journals than in the British specialist journals ($p < 0.001$, 3 df), and more often in the *Annals* than in the other general medical journals[4].

(46) Overall, 1,076 hypothesis tests were presented in 125 reports in the specialist journals. Only 2% of these were statistically significant at the 5% level, which is itself a statistically significant departure from expectation ($p < 0.001$, *z*-test).[10] In the general medical journals, 600 tests were presented in 46 reports of trials. Only 4% of these were statistically significant at the 5% level[4].

(47) **Power calculations:** In 50 (24%) of the reports of trials in the specialist journals, the sample sizes were reported 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*. Reports published in the *BJOG* thus reported power calculations over three times more frequently than those from the other three journals combined (RR = 3.3, 95% CI 2.1–5.2, $p < 0.001$)[10].

(48) Sample size was reported to have been based on prior statistical power calculations in 31 (39%) of the reports published in the general medical journals. The rates were 30% for the *Annals*, 40% for the *BMJ*, 35% for the *Lancet*, and 50% for the *NEJM*[4].

DISCUSSION

(49) Randomization is the only reliable way to create comparable comparison groups with respect to unknown, unmeasured, or imperfectly measured prognostic factors.

For that reason, RCTs are widely accepted as providing the most valid basis for comparing interventions in health care. Indeed, of the various measures to control bias within a trial, proper randomization is arguably the only one that can be confidently assumed to apply to the trial as a whole. All of the other steps which may be taken in an attempt to control biases ("blinding," and analysis by "intention-to-treat," for example) may have been achieved successfully for some of the outcomes assessed in a trial, but not for others. Furthermore, and perhaps even more importantly, the virtual total success of randomization can be guaranteed for all trials. By contrast, other measures used to control bias cannot be implemented for some trials and, if they can be, frequently only partial success can be attained. Indeed, the success of double-blinding and analysis by intention-to-treat hinge upon successful randomization: the concept underlying intention-to-treat analyses is simply the preservation of the randomized allocation.

(50) Considering how centrally important randomization is to any assessment of the validity of a treatment comparison, we are surprised that authors and editors have not been more meticulous in publishing clear reports of the process used to assign participants to comparison groups[10]. As Mosteller and his colleagues[2] put it: "When the randomization leaks, the trial's guarantee of lack of bias runs down the drain."

Were the descriptions of the process of generating and applying treatment assignments in specialist journals worse than those in general journals?

(51) While descriptions of the process of treatment assignment were of generally poor quality in both specialist and general journals, the specialist journals were, on average, less satisfactory. Overall, only 9% of the reports of trials published in the specialist journals, compared with only 15% of those in the general medical journals, clearly stated that adequate methods of both random number generation and allocation concealment had been used. However, the time frames for the assessments were different. The data from the general journals were collected from reports published 3 to 4 years earlier that those from the specialist journals. Some of the general journals, *The Lancet* for example, have instituted new statistical review procedures that would have likely produced more favorable results if we had collected new data from the general journals concurrently with data collected from the specialist journals. Thus, the disparity we observed between the specialist and general journals probably is an underestimate of the differences between them.

(52) Less than a quarter of the reports of trials published in the specialist journals provided information on the type of treatment assignment, but a nonrandom method accounted for a quarter of those, which amounts to 5% of all the trials. That rate is at the lower end of the range (5–10%) found in earlier surveys of reports of "randomized" trials[2,12,21], but substantially higher than the rate of 1% found in our study of general medical journals.

(53) Blocking was reported in only 11% of trials reported in the specialist journals as compared to 28% in the general journals, but the true rates are likely to be higher.

The information on the size of the blocks used was missing in over one-third of those reports in both types of journals. Stratification was mentioned explicitly much less frequently in the specialist journals (9%) than in the general medical journals (39%). However, in both types of journal, about half of the reports in which stratification had been mentioned made no reference to blocking. Although blocking had probably been used in a higher proportion than that, its use should be stated explicitly because stratification is not effective unless blocking, or other form of restriction, has been used as well.

(54) The method of generating random numbers was less well reported in the specialist than in the general medical journals. In only about one-third of the reports in the specialist journals did we conclude that an acceptable approach had been used, as compared to almost half of the reports in the general journals. Moreover, that is a generous estimate, since it includes processes such as shuffled cards and tossed coins as adequate[10]. Because those methods are subject to human perturbations in the production of allocation schedules and are not reproducible, they are certainly less than optimal[2], if not unacceptable[22]. We recommend random number tables and computer random number generators not only for being more reliable and reproducible, but also for being easier and faster[10].

(55) In the process of allocating treatment such that foreknowledge of the allocation is prevented, allocation concealment is generally more important than generation of the randomized assignments per se[1,10,23], yet only 52% of the reports in the specialist journals, and 56% of those in the general medical journals, provided information adequate to assess that aspect of trial design and conduct. We judged many of those stated approaches to have been inadequate, however, and even with those we judged to have been adequate, many reports should have provided further important information (see below). In sum, only 23% of the studies reported in the specialist journals and 26% of the studies reported in the general medical journals appear to have used and reported an adequate approach to allocation concealment.

Were the descriptions of treatment assignment published in the BJOG of better quality than those published in other obstetric and gynecology journals?

(56) Reports of trials published in the *BJOG* were more informative than those in the other three specialist journals. Although the frequency of reports providing evidence that an acceptable approach to generating random numbers had been used was similar in the four journals, reports of trials published in the *BJOG* more frequently included information about the type of randomization, and they were nearly 3 times more likely to report an adequate approach to allocation concealment than the other three journals combined. Furthermore, reports published in the *BJOG* were 3 times more likely than those published in the other specialist journals combined to have reported the use of statistical power calculations. Among the four journals, the overall methodological quality of reports published in the *BJOG* was highest, with those published in the two journals from the U.S. being similar and

superior to reports published in the *JOG*[10]. Editorial efforts similar to those made at the *BJOG* in the mid-1980s are now occurring at *OG*[24], and those too may result in improved quality of reports[10].

(57) Overall, the methodological quality of reports in the *BJOG* was commensurate with that found in reports published in the four general medical journals. Indeed, in important respects (such as allocation concealment and prior statistical power calculations) the quality of reports published in the *BJOG* matched or exceeded that found in the best of the four general medical journals. Even so, those of us (IC and DGA) who had been involved in the editorial efforts to improve the quality of reports of trials published in the *BJOG* were disappointed to find just how much room for improvement still remained[10]. Some rather basic errors of commission and omission continued to be made.

Were the numbers of patients in the comparison groups too often similar in trials that had apparently used simple randomization?

(58) Restricting randomization to balance the numbers in comparison groups in a trial is useful not only to retain statistical power, but also to control for any time trends that may exist in treatment efficacy and outcome measurement during the course of the trial. It is essential if benefits from stratification are to be attained. Nevertheless, restriction can be thought of as primarily cosmetic in large trials. Simple, unrestricted randomization will usually suffice if trials are sufficiently large to ensure a reasonable balance of numbers in the groups. Some discrepancy between the treatment group numbers will normally result, but that will not usually have an important effect on the power of the study[4].

(59) Differences in the comparison group sizes at the time of randomization in those trials using simple randomization should reflect random variation. In other words, some discrepancy between the numbers in the comparison groups is to be expected. Our analysis of reports of trials in general medical journals showed that the reported sizes of the comparison groups tended to be much too similar[4]. That finding was confirmed in our analysis of reports published in the specialist journals[10]. Not only were the similarities we found in the specialist journals unlikely to be due to the play of chance, they were even more marked than those revealed in our study of the general journals.

(60) The strong tendency for the comparison groups to be of equal or similar sizes in these two studies may be explained by: (1) failure to report the use of blocking; (2) failure to report the use of replacement randomization; (3) failure to report the use of a restricted shuffled envelope method; (4) failure to report the use of a nonrandom method of assignment, such as alternation or odd-even date; or (5) "rectification" of an imbalance in sample sizes by nonrandom manipulation of assignments or data[10].

(61) Use of blocking would be the most palatable of those possible explanations, but it is unlikely to explain many cases since so few trials reported blocking, and, in particu-

lar, since the blocked trials yielded more disparate differences than the unrestricted trials[10]. Replacement randomization would also be an acceptable explanation, but no evidence for its use was found[10]. A less acceptable alternative would be that the restricted shuffle approach had been used. Only 5% of the trials that specified a type of randomization used that method, however, so it seems an unlikely explanation for the similarity in the reported size of the comparison groups. A more likely explanation is the use of deterministic, nonrandom allocation[10]. Of the reports in the specialist journals in which a method of generation was stated, 14% used that approach (11 of 77; Table 3A). Thus, the unidentified use of nonrandom, deterministic allocation may explain at least some of the similarities in the numbers of participants assigned to the comparison groups. Unfortunately, that explanation implies that an even higher proportion of trials had used an unacceptable allocation approach, and, in any case, it would only account for some of the effect seen.

(62) The last possibility, nonrandom manipulation of treatment groups, has serious implications because it is the most likely of the possible explanations to introduce selection bias into the comparisons made. Nonrandom manipulations may occur at any time during the trial, from the point of enrollment and allocation to the analysis of the data. However, regardless of whether an investigator alters assignments or differentially drops participants after randomization, those manipulations introduce selection bias.

(63) Although we have not found direct evidence of nonrandom manipulations, the strong tendency for the comparison groups to be of equal or similar sizes provides indirect evidence from both types of journal. Possibly some investigators believe that they will increase the credibility of their trial reports if they present comparison groups of equal size. Most readers may share their misconception. Paradoxically, in aggregate, the results of those manipulations have had exactly the opposite effect. While our results clearly indicate that the set of trials supposedly using simple randomization are not what they purport to be, we cannot identify any particular trials as suspect, as some trials would be expected to achieve almost equal numbers simply by chance[10].

Was the percentage of statistically significant differences in characteristics measured at baseline at the expected level of 5%?

(64) On strictly theoretical grounds, if randomization is properly implemented, establishment of comparability at baseline is not necessary[10]. Random assignment permits the use of probability theory to depict the extent to which any difference in outcome between treatment groups is likely to be due to chance. Although, in a particular study, the groups compared may never be perfectly balanced for important prognostic variables, randomization makes it possible to ascribe a probability distribution to the difference in outcomes between the comparison groups, and a probability can then be assigned to the observed difference between them. The process of randomization underlies significance testing, and that process is independent of prognostic factors, known or unknown[10,25]. As Fisher[25] stated, randomization

"relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which his [sic] data may be disturbed."

(65) In practice, however, comparison of baseline characteristics in the trial groups is useful for at least two reasons. First, evidence that reasonable similarity in baseline characteristics has been achieved will tend to support a claim that randomization has been implemented correctly. Second, the point estimates of effects may be improved by statistical adjustment to take account of chance baseline imbalances in important prognostic variables, and, if modeled properly, it may also increase precision[18]. Moreover, the preferable procedure is to adjust for variables because they are known to be prognostic rather than because imbalance was observed.

(66) Although comparisons of baseline characteristics were presented in a majority of the reports published in the specialist and general medical journals (84% and 92%, respectively), many of the reported comparisons were deficient. In reports in which at least one continuous variable (such as a mean or median) had been presented, 32% were either unaccompanied by measures of variability, or accompanied by a measure that was inappropriate (most frequently the standard error). A similar proportion (33%) of the reports published in the general medical journals were deficient in that respect.

(67) An even more worrying deficiency, which was present in 61% of the reports in the specialist journals and 58% of the reports in the general journals, was the inappropriate use of hypothesis tests to compare the distribution of baseline characteristics in the comparison groups. Using hypothesis tests to compare baseline characteristics in RCTs assesses the probability that the differences observed have occurred by chance, when, in properly randomized trials, it is known already that any differences observed *have* occurred by chance[10]. As noted by Altman[20] elsewhere, "Such a procedure is clearly absurd." Hypothesis tests are superfluous and their use in comparisons of baseline characteristics can mislead investigators and their readers[10]. For example, substantial differences between comparison groups could be judged as unimportant merely because the *p*-value happened to be greater than 0.05. Therefore, comparisons should be based on consideration of the prognostic strength of the variables measured and the magnitude of any chance imbalances that have occurred[20].

(68) Although use of hypothesis tests inappropriately addresses the assessment of baseline imbalances in prognostic characteristics, such tests might, in principle, be used by investigators who are concerned that randomization may not have been executed effectively in their studies[4,20]. Occasionally gross imbalances, quite incompatible with random variation, are revealed in that way by other investigators[8]. Finding several statistically significant differences between the comparison groups may suggest that randomization has not been achieved; but use of tests in that way will often pose substantial problems of interpretation.

(69) Concern that randomization may not have been executed correctly seems an unlikely explanation for the use of hypothesis tests by the authors of the 61% of reports in the specialist journals in which test results were presented: only 2% of

the tests reported were statistically significant at the 5% level, a discrepancy from expectation which is highly unlikely to reflect chance. This observed frequency of "statistically significant" test results in the specialist journals is more extreme in its departure from the expected 5% than the value of 4% we found in the general medical journals.

(70) A plausible explanation for those discrepancies is that, when comparing baseline characteristics using hypothesis tests, investigators may decide not to report a statistically significant result, believing that by withholding that information they will increase the credibility of their reports. In fact, the opposite has occurred. Having too few statistically significant results in aggregate has hurt the credibility of these trials. Investigators must report baseline comparisons on important prognostic variables whether they are statistically significantly different or not. Clearly, not only are hypothesis tests superfluous and potentially misleading, they can be positively harmful if they lead investigators to drop important variables from baseline comparisons[10].

CONCLUSIONS

(71) Our findings are not reassuring. 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 methodology[26]. Thus, while reporting clearly must be improved, the deficiencies in the design and conduct of trials must also be urgently addressed[10].

(72) Although, as predicted, descriptions of the process of generating and applying treatment assignments in reports of trials published in specialist journals were of somewhat poorer quality than those published in general journals, the standard of reporting in both samples leaves a great deal to be desired. Although the quality of reports of trials published in the *BJOG* was indeed better than that of those published in the other three obstetric and gynecology journals (and of comparable quality to those in the best general medical journals), the *BJOG* has considerable room for improvement.

(73) We have confirmed that the numbers of participants in the comparison groups of trials which have apparently used simple randomization were too often too similar, and that the observed percentage of statistically significant differences in characteristics measured at baseline is much less than the expected value of 5%. Those are disturbing findings in that they suggest nonrandom manipulation of comparison groups and selective reporting of baseline comparisons.

(74) Because the quality of randomization is of such fundamental importance in controlling selection biases in treatment comparisons, the reporting of randomization procedures deserves to be given higher priority, by methodologists, investigators, authors and journal editors. At a minimum, reports of randomized controlled trials should include descriptions of (i) the type of randomization; (ii) the method of sequence generation; (iii) the method of allocation concealment; (iv) the persons generating and executing the scheme; and (v) the comparative baseline charac-

teristics, 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[10].

(75) While improving the standard of reporting is surely a shared responsibility, omission of randomization details to date has probably been primarily due to authors and not to 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, arguably the burden for improvement should fall primarily upon investigators and authors, although editors could stimulate that process[10].

(76) Protestations from authors about lack of space are not an acceptable excuse for omission. Space will always be a limitation (albeit, much less so in electronically published reports); the issue is the relative importance of the topics addressed. Information with little bearing on scientific validity has been included in many reports while critical elements of the randomization process have been omitted. Yet in a well-executed, blinded, randomized controlled trial, aspects other than randomization are almost scientifically inconsequential to the treatment comparisons since they would have been applied equally to unbiased comparison groups. Certainly, we would not wish to promote a cavalier attitude toward the other methodological elements of trials: they must be adequately addressed, and surely some have to be adequately described for readers to interpret the findings and extrapolate the results. Yet, proper reporting of the randomization procedures should be of the highest priority, and those trials that fail to provide such information should be interpreted cautiously[10].

OTHER

(77) **Date of Submission:** 1995 February 16

(78) **Date of Acceptance:** 1995 June 6

(79) **Acknowledgments:** This work was supported by the Centers for Disease Control and Prevention. Iain Chalmers is supported by the National Health Service Research and Development Programme. As stated in the Introduction, the *Journal of the American Medical Association* has published many of the major findings of our research and similar portions of selected text [10]. This paper presents the complete findings and contrasts them with results from the *Lancet*[4] pertaining to general journals. We reproduced data from the *Lancet* paper by Altman and Doré[4] along with more complete findings not in the original publication.

(80) **Address for Requests for Reprints:** Kenneth F.Schulz, PhD, Division of STD/HIV Prevention, NCPS, MS E-02, VCenters for Disease Control and Prevention, Atlanta, GA 30333 USA; fax 404-639-8609.

(81) **Current Authors Addresses: Kenneth F. Schulz:** Division of STD/HIV Prevention, National Center for Prevention Services, Centers for Disease Control and Prevention, Atlanta, GA USA and The UK Cochrane Centre, Oxford , UK and London School of Hygiene and Tropical Medicine, London, UK; **Iain Chalmers:** The UK

Cochrane Centre, Oxford, UK; **Douglas G. Altman:** Medical Statistics Laboratory, Imperial Cancer Research Fund, London, UK; **David A. Grimes:** Department of Obstetrics, Gynecology, and Reproductive Sciences, University of California, San Francisco, CA USA; **Caroline J. Doré:** Medical Statistics Unit, Royal Postgraduate Medical School, London, UK.

REFERENCES

1 . HILL AB. The clinical trial. N Engl J Med 1952;247:113–119.

2. 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.

3. DERSIMONIAN R, CHARETTE LJ, MCPEEK B, MOSTELLER F. Reporting on methods in clinical trials. N Engl J Med 1982;306:1332–1337.

4. ALTMAN DG, DORÉ CJ. Randomization and baseline comparisons in clinical trials. Lancet 1990;335:149–153.

5. PIGNON JP, POYNARD T. Statistics in clinical trials (letter). Lancet 1990;335: 614.

6. 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.

7. THACKER SB. The efficacy of intrapartum electronic fetal monitoring. Am J Obstet Gynecol 1987;156:24–30.

8. KEIRSE MJNC. Amniotomy or oxytocin for induction of labor: Re-analysis of a randomized controlled trial. Acta Obstet Gynecol Scand 1988;67:731–735.

9. GRIMES DA, SCHULZ KF. Randomized controlled trials of home uterine activity monitoring: A review and critique. Obstet Gynecol 1992;79:137–142.

10. SCHULZ KF, CHALMERS I, GRIMES DA, ALTMAN DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA 1994;272:125–128.

11. GRANT A. Reporting controlled trials. Br J Obstet Gynaecol 1989;96:397–400.

12. CHALMERS I, HETHERINGTON J, NEWDICK M, MUTCH L, GRANT A, ENKIN 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.

13. DEAN AG, DEAN JA, BURTON AH, DICKER RC. Epi Info, Version 5: A Word Processing, Database, and Statistics System for Epidemiology on Microcomputers. Atlanta, Ga: Centers for Disease Control and Prevention; 1990.

14. POCOCK SJ. Clinical Trials: A Practical Approach. Chichester, England: John Wiley and Sons; 1983.

15. MACFARLANE AJ. Variations in numbers of births and perinatal mortality by day of week in England and Wales. BMJ 1978;2:1670–1673.

16. ALTMAN DG. Randomization: essential for reducing bias. BMJ 1991;302:1481–1482.

17. 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.

18. LAVORI PW, LOUIS TA, BAILAR JC, POLANSKY M. Designs for experiments—parallel comparisons of treatment. N Engl J Med 1983;309:1291–1299.

19. ROTHMAN KJ. Epidemiologic methods in clinical trials. Cancer 1977;39 (suppl 4): 1771–1775.

20. ALTMAN DG. Comparability of randomized groups. Statistician 1985;34:125–136.

21. EVANS M, POLLOCK AV. Trials on trial: a review of trials of antibiotic prophylaxis. Arch Surg 1984;119:109–113.

22. MEINERT CL. Clinical Trials: Design, Conduct, and Analysis. New York, NY: Oxford University Press; 1986.

23. 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.

24. GRIMES DA. Randomized controlled trials: "it ain't necessarily so". Obstet Gynecol 1991;78:703–704.

25. FISHER RA. The Design of Experiments. 8th ed. Edinburgh, Scotland, UK: Oliver & Boyd Ltd, 1966.

26. LIBERATI A, HIMEL HN, CHALMERS TC. A quality assessment of randomized controlled trials of primary treatment of breast cancer. J Clin Oncol 1986;4:942–951.

Table 1a. The type of randomization stated in the four specialist journals, 1990 and 1991.

Type of randomization stated	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
Simple (unrestricted)	0% (0)	6% (3)	5% (1)	0% (0)	2% (4)
Balanced (restricted)	6% (4)	35% (17)	5% (1)	9% (7)	14% (29)
Deterministic (nonrandom)	3% (2)	6% (3)	5% (1)	7% (5)	5% (11)
Other	0% (0)	0% (0)	5% (1)	0% (0)	0% (1)
Not Stated	91% (58)	52% (25)	80% (16)	84% (62)	78% (161)
Total	100% (64)	100% (48)	100% (20)	100% (74)	100% (206)

Table 1b. **The type of randomization stated in the four general medical journals, 1987+.**

Type of randomization stated	Ann Intern Med	Br Med J	Lancet	N Engl J Med	Total
Simple (unrestricted)	0% (0)	0% (0)	5% (1)	0% (0)	1% (1)
Balanced (restricted)	30% (6)	30% (6)	20% (4)	33% (7)	28% (23)
Deterministic (nonrandom)	0% (0)	0% (0)	0% (0)	5% (1)	1% (1)
Other	0% (0)	0% (0)	0% (0)	0% (0)	0% (0)
Not Stated	70% (14)	70% (14)	75% (15)	62% (13)	69% (56)
Total	100% (20)	100% (20)	100% (20)	100% (21)	100% (81)

Table 2a. **Stratified trials in the specialist journals.**

Stratification status	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
Stratified*	5% (3)	19% (9)	0% (0)	8% (6)	9% (18)
Stratified & blocked*	2% (1)	10% (5)	0% (0)	3% (2)	4% (8)

* Includes the trial that used minimization.

Table 2b. **Stratified trials in the general journals.**

Stratification status	Ann Intern Med	Br Med J	Lancet	N Engl J Med	Total
Stratified*	60% (12)	25% (5)	20% (4)	48% (10)	39% (31)
Stratified & blocked*	25% (5)	15% (3)	15% (3)	24% (5)	20% (16)

* Includes the trial that used minimization.

Table 3a. **Methods for generating random numbers in the specialist journals.**

Method for Generating Random Numbers	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
Computer*	20% (13)	21% (10)	5% (1)	16% (12)	18% (36)
Random number table*	13% (8)	8% (4)	10% (2)	11% (8)	11% (22)
Other possible random processes*	3% (2)	8% (4)	0% (0)	3% (2)	4% (8)
Deterministic (nonrandom)	3% (2)	6% (3)	5% (1)	7% (5)	5% (11)
Not stated	61% (39)	56% (27)	80% (16)	64% (47)	63% (129)
Total	100% (64)	100% (48)	100% (20)	100% (74)	100% (206)
Any adequate random process	36% (23)	38% (18)	15% (3)	30% (22)	32% (66)

* Adequate random process.

Table 3b . **Methods for generating random numbers in the general journals.**

Method for generating randon numbers	Ann Intern Med	Br Med J	Lancet	N Engl J Med	Total
Computer*	25% (5)	20% (4)	10% (2)	38% (8)	23% (19)
Random number table*	20% (4)	30% (6)	10% (2)	19% (4)	20% (16)
Other possible random processes*	5% (1)	0% (0)	15% (3)	0% (0)	5% (4)
Deterministic (nonrandom)	0% (0)	0% (0)	0% (0)	5% (1)	1% (1)
Not stated	50% (10)	50% (10)	65% (13)	38% (8)	51% (41)
Total	100% (20)	100% (20)	100% (20)	100% (21)	100% (81)
Any adequate random process	50% (10)	50% (10)	35% (7)	57% (12)	48% (39)

* Adequate random process.

Table 4a. Allocation concealment methods in the specialist journals.

Allocation method	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
Numbered or	8%	10%	0%	7%	7%
coded containers*	(5)	(5)	(0)	(5)	(15)
Pharmacy	6%	10%	0%	8%	7%
concealed*	(4)	(5)	(0)	(6)	(15)
Centrally concealed*	2%	6%	0%	1%	2%
(e.g. telephone)	(1)	(3)	(0)	(1)	(5)
Sequentially numbered					
opague, sealed	3%	17%	5%	3%	6%
envelopes*	(2)	(8)	(1)	(2)	(13)
envelopes—	20%	25%	10%	16%	19%
other	(13)	(12)	(2)	(12)	(39)

* Adequate allocation concealment method.

Table 4b. **Continuation of Table 4a.**

Allocation method	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
List/table/ schedule	9% (6)	4% (2)	0% (0)	3% (2)	5% (10)
Deterministic (nonrandom)	3% (2)	6% (3)	5% (1)	7% (5)	5% (11)
Not stated or described	48% (31)	21% (10)	80% (16)	55% (41)	48% (98)
Total	100% (64)	100% (48)	100% (20)	100% (74)	100% (206)
Use of an adequate allocation concealment method	19% (12)	44% (21)	5% (1)	19% (14)	23% (48)

Table 4c. **Allocation concealment methods in the general journals.**

Allocation method	Ann Intern Med	Br Med J	Lancet	N Engl J Med	Total
Numbered or coded containers*	5% (1)	5% (1)	10% (2)	19% (4)	10% (8)
Pharmacy concealed*	20% (4)	10% (2)	0% (0)	5% (1)	9% (7)
Centrally concealed* (e.g. telephone)	10% (2)	0% (0)	0% (0)	10% (2)	5% (4)
Sequentially numbered, opaque, sealed envelopes*	0% (0)	0% (0)	0% (0)	10% (2)	2% (2)
Envelopes—other	20% (4)	20% (4)	10% (2)	19% (4)	17% (14)

* Adequate allocation concealment method.

Table 4d. Continuation of Table 4c.

Allocation method	Ann Intern Med	Br Med J	Lancet	N Engl J Med	Total
List/table/ /schedule	10% (2)	15% (3)	15% (3)	5% (1)	11% (9)
Deterministic (nonrandom)	0% (0)	0% (0)	0% (0)	5% (1)	1% (1)
Not stated or described	35% (7)	50% (10)	65% (13)	29% (6)	44% (36)
Total	100% (20)	100% (20)	100% (20)	100% (21)	100% (81)
Use of an adequate allocation concealment method	35% (7)	15% (3)	10% (2)	43% (9)	26% (21)

Table 5a. **Baseline comparisons in the specialist journals.**

Baseline	Am J	Br J	J Obstet	Obstet	Total
comparisons	Obstet	Obstet	Gynaecol	Gynecol	
	Gynecol	Gynaecol			
	(*n*=64)	(*n*=48)	(*n*=20)	(*n*=74)	(*n*=206)
All variables reported:					
≥1 Presented for each treatment group	81% (52)	94% (45)	75% (15)	82% (61)	84% (173)
Median number presented (range)	7 (1–34)	5 (1–32)	4 (2–7)	7 (1–58)	6 (1–58)

Table 5b. Continuation of Table 5a.

Baseline comparisons	Am J Obstet Gynecol (*n*=64)	Br J Obstet Gynaecol (*n*=48)	J Obstet Gynaecol (*n*=20)	Obstet Gynecol (*n*=74)	Total (*n*=206)
Continuous only:					
≥1 Continuous variable reeported for each group	80% (51)	85% (41)	70% (14)	74% (55)	78% (161)
Median number presented (range)	4 (1–12)	4 (1–18)	3 (1–7)	4 (1–14)	4 (1–18)
Appropriate variability*	65% (33)	81% (33)	57% (8)	64% (35)	68% (109)
Inappropriate variability†	24% (12)	17% (7)	36% (5)	24% (13)	23% (37)
No measure of variability reported	12% (6)	2% (1)	7% (1)	13% (7)	9% (15)
Overall poor reporting of baseline comparisons‡	47% (30)	23% (11)	55% (11)	45% (33)	41% (85)

* Standard deviation, range, centiles, or raw data reported for at least one.

† Standard error or confidence interval reported for at least one without reporting at least one baseline comparison with appropriate variability.

‡ Authors did not present baseline comparisons or did not report appropriate variability.

Table 5c. **Baseline characteristics in the general journals.**

Baseline comparisons	Ann Intern Med (*n*=20)	Br Med J (*n*=20)	Lancet (*n*=20)	N Engl J Med* (*n*=20)	Total* (*n*=80)
All variables reported:					
≥1 Presented for each treatment group	95% (19)	100% (20)	80% (16)	95% (19)	92% (74)
Median number presented (range)	10 (4–39)	8.5 (2–22)	8.5 (1–39)	9 (2–39)	9 (1-39)

* Does not include the one deterministic, nonrandom trial.

Table 5d. **Continuation of Table 5c.**

Baseline comparisons	Ann Intern Med (*n*=20)	Br Med J (*n*=20)	Lancet (*n*=20)	N Engl J Med[*] (*n*=20)	Total[*] (*n*=80)
Continuous only:					
≥1 Continuous variable reported for each group	95% (19)	90% (18)	75% (15)	85% (17)	86% (69)
Median number presented (range)	5 (1–15)	4 (1–19)	3 (1–13)	3 (1–22)	4 (1–22)
Appropriate variability†	74% (14)	67% (12)	73% (11)	53% (9)	67% (46)
Inappropriate variability‡	26% (5)	28% (5)	13% (2)	18% (3)	22% (15)
No Measure of variability reported	0% (0)	6% (1)	13% (2)	29% (5)	12% (8)
Overall poor reporting of baseline comparisons§	30% (6)	30% (6)	40% (8)	45% (9)	36% (29)

[*] Does not include the one deterministic, nonrandom trial.

† Standard deviation, range, centiles, or raw data reported for at least one.

‡Standard error or confidence interval reported for at least one without reporting at least one baseline comparison with appropriate variability.

§Authors did not present baseline comparisons or did not report appropriate variability.

Table 6a. **Use of hypothesis tests (tests of statistical significance) to compare baseline characteristics in the specialist journals.**

Comparing baseline Variables	Am J Obstet Gynecol (*n*=64)	Br J Obstet Gynaecol (*n*=48)	J Obstet Gynaecol (*n*=20)	Obstet Gynecol (*n*=74)	Total (*n*=206)
Trials using hypothesis tests	72% (46)	35% (17)	20% (4)	78% (58)	61% (125)
In those trials using hypothesis tests:					
Specified test methods	87% (40)	82% (14)	75% (3)	85% (49)	85% (106)
Mean number tested	10.3	6.2	4.0	8.3	8.6
Total number tested	472	106	16	482	1076
Percent (Number) statistically significant at p<0.05	2% (11)	2% (2)	0% (0)	2% (9)	2% (22)

Table 6b. **Use of hypothesis tests (tests of statistical significance) to compare baseline characteristics in the general journals.**

Comparing baseline variables	Ann Intern Med (*n*=20)	Br Med J (*n*=20)	Lancet (*n*=20)	N Engl J Med[*] (*n*=20)	Total[*] (*n*=80)
Trials using hypothesis tests:	70% (14)	50% (10)	55% (11)	55% (11)	58% (46)
In those trials using hypothesis tests:					
Specified test methods	64% (9)	70% (7)	82% (9)	82% (9)	74% (34)
Mean number tested	10.7	8.5	12.7	20.5	13.0
Total number tested	150	85	140	225	600
Percent (number) statistically significant at $p<0.05$	5% (8)	3.5% (3)	2% (3)	4% (10)	4% (24)

[*] Does not include the one deterministic, nonrandom trial.

Figure 1 . The relationship between the difference in sample sizes in the treatment and control groups and total study size for 96 unblocked trials in the specialist journals. The straight lines represent the expected distribution due to the play of chance. Total study size is shown on a square root scale to make the confidence interval lines straight. The 95% prediction interval is approximately:±1.96√(total study size). [Figure originally in *The Journal of the American Medical Association,* July 13, 1994, Volume 272, Page 127: reproduced with permission.]

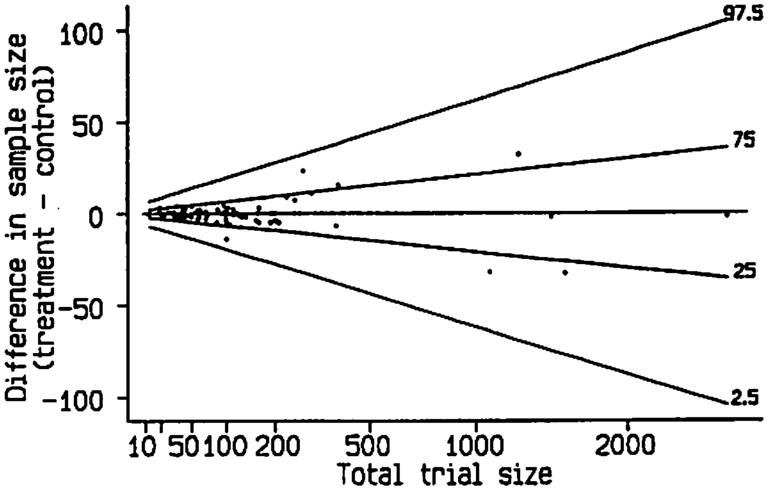
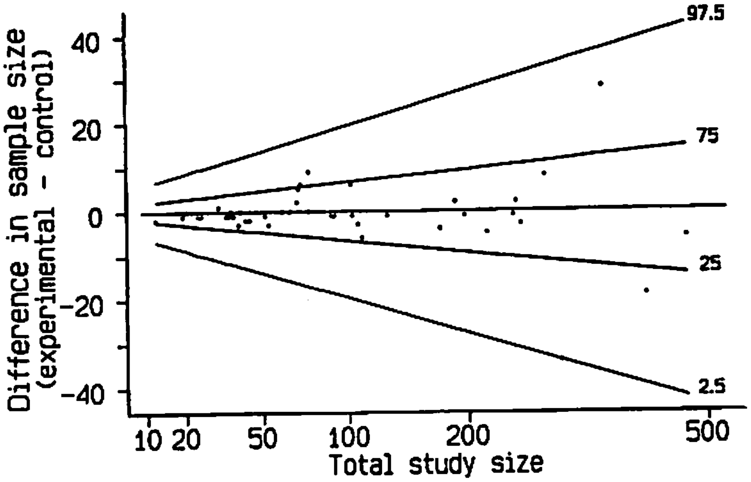


Figure 2 . The relationship between the difference in sample sizes in the treatment and control groups and total study size for 43 unblocked trials in the general journals. The straight lines represent the expected distribution due to the play of chance. Total study size is shown on a square root scale to make the confidence interval lines straight. The 95% prediction interval is approximately: ±1.96√(total study size). [Figure originally in *The Lancet*, January 20, 1990, Volume 335, Page 151: reproduced with permission.]


Difference in sample size
(experimental - control)
Total study size