Methodological Quality and Bias in Randomised Controlled Trials

A thesis submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy
of the University of London

Kenneth Fredrick Schulz

Department of Epidemiology and Population Sciences
London School of Hygiene and Tropical Medicine
University of London
Keppel Street, London WC1E 7HT

Abstract

To evaluate the methodological quality of randomised trials in recently published articles and to examine the associations between methodological quality and bias, three related investigations were undertaken. First, to ensure the development of useful measures for the adequacy of randomisation, approaches to allocation were assessed as reported in 206 parallel group trials published in recent volumes of journals of obstetrics and gynaecology. Next, a study was conducted of associations between methodological quality and treatment effects. The material analyzed came from 250 trials in 33 meta-analyses on pregnancy and childbirth topics. Finally, the reported approaches to blinding and handling of exclusions were assessed from a random sample of 110 of the 206 previously identified reports.

In the 206 published trials, 77% reported either inadequately or unclearly concealed treatment allocation. Additional analyses suggest that non-random manipulation of comparison groups may have occurred.

In the next study, compared with trials in which authors reported adequately concealed treatment allocation, trials in which authors reported inadequately or unclearly concealed allocation yielded larger estimates of treatment effects ($p<0.001$). Odds ratios were distorted by 41% and 33%, respectively. Those associations likely represent bias and are particularly disconcerting in light of the results above from recently published trials. Lack of double-blinding in trials was also associated with larger treatment benefits. However, trials in which authors reported excluding participants after randomisation were not associated with larger treatment effects. That lack of association appeared to be due to incomplete reporting.

The analysis of 110 recently published trials also supported the findings that some of the trials not reporting exclusions may actually have had exclusions. In practice, that incomplete reporting could lead to misinterpretations of trial quality. Moreover, only about half the trials that could have double-blinded actually did so. When investigators attempted double-blinding, only 16% provided any written assurances of successfully implementing blinding and only 6% tested its efficacy.

Table of Contents

Figures and Tables

Preface

I am very grateful to my supervisor, Richard Hayes of the Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, for his invaluable guidance throughout my work. I very much enjoyed this opportunity to work together. He has provided many worthwhile comments and suggestions on drafts of this thesis. Amongst everything else, Richard persevered with me until we identified the most appropriate available model and software for the logistic regression analysis in Chapter 4.

I gratefully acknowledge the many substantive contributions of Iain Chalmers of the UK Cochrane Centre in Oxford and Doug Altman of the Imperial Cancer Research Fund in London. I also am delighted to have had the opportunity to work directly with both of them. They have provided many valuable insights on the design and analysis of my work and many comments and suggestions on drafts of this manuscript. They have remained keenly interested in my work throughout these years and have been very supportive. I spent a great deal of time with Iain in Oxford and am honoured that he appointed me as the first Visiting Research Fellow at the UK Cochrane Centre. Doug, of course, has published many articles on the quality of statistical methods, and I always appreciated his constructive perspective. He also helped me acclimate to cricket from baseball.

I also gratefully acknowledge the contributions, from afar, of David Grimes of the Department of Obstetrics, Gynecology, and Reproductive Sciences at the University of

California, San Francisco. He provided valuable insights into the designs and analyses in Chapters 3 and 5, and made many useful comments on drafts of this manuscript.

I appreciate suggestions contributed by Simon Thompson and Jimmy Whitworth of the Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine on the design of my research. Simon and Jimmy, along with Richard and Doug, comprised the assessment panel for my PhD upgrading process. I also am grateful to Michael Hills and Tom Marshall, also of the Department of Epidemiology and Population Sciences, for kindly reviewing the modelling work in Chapter 4. And my thanks to all the others who have helped during these years, but who are not specifically mentioned here.

My warm appreciation goes to Valerie Beral of the University of Oxford for urging me to come to England for this work and to Stuart Pocock and Peter Smith of the Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine for their help in making it all possible. And of utmost importance, I gratefully acknowledge the generous financial support I have received throughout this whole period from the Centers for Disease Control and Prevention in Atlanta, Georgia. That support made this research a reality.

I reserve my warmest thanks for my family. They played an indispensable part with their enthusiastic interest and encouragement at all stages. My wonderful wife and our 5-year-old son, Susan and Cameron, put up with my long and unsociable working hours with patience and tolerance. I thank them for their love and endurance over these busy years, and Susan, an epidemiologist, for her helpful comments on drafts of this thesis. I also thank my parents, Helen and Ben Schulz, for their support, even though they may not have been too pleased when one of their sons moved five thousand miles away. Indeed, I dedicate this thesis to my family, particularly to my mother, Helen Louise Anderson Schulz. Mom died on January 15, 1994 during the preparation of this manuscript after a long, courageous battle with breast cancer. She might not have fully understood this thesis, but like the thoughtful, devoted Mother she always was, she would have loved it nonetheless.

And finally, Susan, Cameron, and I wish to thank the kind, hospitable, and sociable people of England for making our time in their country so enjoyable. We have made many friends with whom we shall remain in close contact for years to come.

Portions of this thesis have been prepared for dissemination. Excerpts of Chapter 3 were presented at The Second International Congress on Peer Review in Biomedical Publication in Chicago, Illinois and at the meeting of the International Society for Clinical Biostatistics in Cambridge, both during September of 1993. Parts of Chapter 4 have been accepted for presentation at the Society of Clinical Trials meeting in Houston, Texas during May of 1994, and at the International Society for Clinical Biostatistics meeting in Basle during July of 1994. A manuscript distilled from Chapter 3 has been accepted for publication by the ***Journal of the American Medical Association***. I am also preparing manuscripts distilled from other chapters.

Chapter 1

1. Introduction

1.1 Synopsis of background

Comparisons of different forms of health interventions can be misleading unless investigators take precautions to ensure that, prior to receiving care, their study contains unbiased comparison groups with respect to prognosis. In controlled trials of prevention or treatment, randomisation produces unbiased comparison groups by avoiding selection biases. That characterizes randomisation's only unique strength. That strength, however, becomes of crucial importance in the potentially common circumstance where the treatment effects may be of comparable magnitude to the biases that plague most non-randomised comparisons of alternative forms of treatment and prevention.

Even in view of the central importance of randomisation for achieving unbiased comparisons, authors often inadequately detail the steps taken to assign participants to comparison groups in trials. If randomisation indeed prevents bias as suggested, trials that have failed to report adequate approaches to ensuring proper randomisation should yield systematically different estimates of treatment effects compared with those derived from trials that have apparently used adequate approaches. Moreover, trials that have failed to report adequate approaches to other important methodological components of trials, such as blinding and the handling of exclusions, might well also yield systematically different estimates of treatment effects.

Meagre evidence exists, however, in support of those presumptions. One study found that trials that did not take precautions to conceal treatment allocation schedules from those responsible for recruiting and entering participants yielded larger estimates of treatment effects than trials in which allocation had been adequately concealed (Chalmers et al. 1983). Another study sought an association between overall measures of methodological quality of each trial and estimates of treatment effects (Emerson et al. 1990). It did not detect such an association. Methodological aspects of their approach, however, may have handicapped their ability to accurately assess the relationship. Other specific, independent effects, such as blinding and handling of exclusions, do not appear to have been systematically evaluated.

Other secondary analyses from meta-analyses have examined the relationship between study quality and estimates of treatment effects. Apart from a few exceptions, associations between methodological quality of studies and estimates of the size of difference between experimental groups within them usually occur because methodologically inferior studies tend to yield larger differences than those that have employed sounder methods for reducing bias (Chalmers et al. 1989). However, the preponderance of those investigations were subject to post-hoc reasoning and, thereby, largely serve a hypothesis-generating function.

Non-randomised comparisons also tend to yield larger treatment effects than randomised comparisons in controlled trials. That supports the tendencies of the findings alluded to above. Thus, available evidence leans toward suggesting that trials that have failed to report adequate methodological approaches yield larger estimates of treatment effects than those derived from trials that have apparently used adequate approaches.

1.2 Research aims

As one important focus of the research in this thesis, the association was examined between methodological approaches intended to control bias in trials and estimates of treatment effects. That focus was envisaged to lead the research in the direction of at least two potential benefits: (1) if the results supported prior suggestive findings, documentation and dissemination of the importance of adherence to proper design principles in the conduct of RCTs would likely spur greater methodological rigour in the conduct and reporting of primary RCTs; and (2) for future meta-analyses, knowing the effect of methodological quality on treatment effects in primary studies would aid in the establishment of exclusion criteria and in the approach to the quantitative synthesis.

In the examination of the associations between methodological approaches and treatment effects, a database was used of systematic reviews of controlled trials in pregnancy and childbirth (Enkin et al. 1993). Broad entry criteria for primary trials included in the reviews meant that the trials within this analysis varied in quality from well-done RCTs to poorly-done trials where a systematic method of assignment was utilized, such as alternate assignment or assignment by odd or even birth date. That heterogeneity presented an excellent scientific opportunity to study the relationship between methodological quality and effect sizes among trials that otherwise had sufficient similarity to be considered as part of one overview exercise (Sacks et al. 1987; Chalmers et al. 1987a; Chalmers et al. 1987b). While all researchers prefer studies to be of uniformly high quality, in this analysis the net result of heterogeneity in study quality was a better chance to delineate factors associated with that variation.

Quality measures of four primary components were used in the analysis: allocation concealment, exclusions after randomisation, blinding, and generation of the randomisation sequence. Broad agreement appeared to exist that those were the most critical elements in trial design for safeguards against bias (Chalmers et al. 1990; Emerson et al. 1990; Colditz et al. 1989; Miller at al 1989; Gøtzsche 1989; Peto 1987; Collins et al. 1987; Pocock 1982).

As a second important focus of this thesis, the methodological quality of RCTs was evaluated as reported in recently published articles. That work was important in and of itself, but it also served three other important functions relative to the analysis of the association between quality and treatment effects. First, it proved necessary to ensure that useful measures for the adequacy of randomisation had been developed. Second, it served as a valuable resource to put those results in context. And third, it provided more detailed data on the handling of exclusions after randomisation that clarified and supported some of those results from the association analysis. The following section contains descriptions of the temporal progression of the research and its location in this thesis.

1.3 Description of ensuing chapters

Chapter 2 provides background on randomisation, terminology, and previous research. Chapter 3 describes research performed before the analysis of association. The adequacy of the descriptions of randomisation and related topics from current journals of obstetrics and gynaecology were analyzed. That work aided in the development of measures used for the adequacy of randomisation in Chapter 4 and provided a context for those results.

Chapter 4 addresses the analysis of the association between methodological measures to control bias and estimates of treatment effects. The results from that work posed more questions on the adequacy of approaches to blinding and the handling of exclusions. In Chapter 5, the adequacy of information on those topics in current journals of obstetrics and gynaecology was addressed.

Last, in Chapter 6, some overall comments are offered on the findings in this thesis and on further avenues of research.

1.4 Focus on obstetrics and gynaecology

This research focused on obstetrics and gynaecology for at least four reasons. First, and most importantly, the unique database that facilitated the analyses of methodological quality and bias, the ***Oxford Database of Perinatal Trials*** (Chalmers 1992), predominantly contains trials in obstetrics. Second, the subject areas in obstetrics and gynaecology are relatively well-suited for RCTs, particularly in obstetrics where outcome events generally occur more quickly and, consequently, require shorter follow-up times. Third, while perhaps well suited for RCTs, others have reported that of all medical specialties, clinical practice in obstetrics and gynaecology was least likely to be supported by scientific evidence (Cochrane 1979; Office of Technology Assessment 1983). While that dubious distinction has been retracted in England by Cochrane (Chalmers et al. 1989) largely through the activities of the National Perinatal Epidemiology Unit in Oxford, vast improvements in the quality of research remain necessary (Institute of Medicine 1992). Thus, methodological work in obstetrics and gynaecology might yield particularly rewarding scientific gains. And fourth, a comprehensive analysis had not occurred of the methodological quality of RCTs published in the literature from obstetrics and gynaecology.

While this research may be particularly important to obstetrics and gynaecology, it may also have important implications for other medical specialties. In particular, the research addressing the association between the methodological quality of trials and treatment effects has implications for the design, implementation and analysis of all RCTs, as well as for the conduct of meta-analyses of RCTs. Even the research issues addressing the scientific quality of RCTs in the literature of obstetrics and gynaecology can give those in other medical specialties an indication of the scientific rigour they would expect to find in their journals.

Chapter 2

2. Background

2.1 Advantages of randomisation

A comparative clinical trial may be defined as an experiment designed to assess the relative effectiveness of two or more treatments (exposures). Since investigators strive to identify the better treatment if one exists and to convince others of the validity of the results, they need to design and conduct trials carefully to minimize bias.

Until relatively recently in medical and public health history, investigators rarely used formally controlled experimentation. Over the twentieth century, the controlled clinical trial has gained increasing recognition as the best approach to evaluating health care and prevention alternatives, and the number of trials published has increased (Fletcher and Fletcher 1979). A relatively recent development has been the use of randomisation in the assignment of participants to comparison groups.

R. A. Fisher developed randomisation as a basic principle of experimental design in the 1920s. The successful adaptation of randomised controlled trials (RCTs) to health care took place in the late 1940s, largely due to the advocacy and developmental work of Sir Austin Bradford Hill while at the London School of Hygiene and Tropical Medicine (Armitage 1982). His efforts resulted in the first use of random numbers to allocate trial participants (Medical Research Council 1948). Many of the critical design elements espoused in his early writings remain relevant to this day (Hill 1952).

At least three major advantages of randomisation exist. First, it eliminates bias in the assignment of treatments. That is, treatment comparisons will not be prejudiced by selection of patients of a particular kind, whether consciously or not, to receive a particular form of treatment. The concept of avoiding bias includes eliminating it from the acceptance or rejection of participants for the study as well as eliminating it from the assignment of participants, once accepted, to treatment. Proper registration of each participant immediately upon determination of eligibility for the trial, but before the randomised assignment is known, will prevent bias.

Second, random allocation facilitates various devices for blinding the identity of treatments to the investigators, participants, and evaluators, including the possible use of a placebo (Armitage 1982). Those devices enable bias reduction in a trial after assignment. They would be difficult, perhaps even impossible, to implement if investigators assigned treatments in a non-random fashion.

Third, random assignment permits the use of probability theory to express the likelihood that any difference in outcome between treatment groups merely reflects chance. While many view this advantage lightly, it has received greater recognition in the epidemiological community of late (Greenland 1990).

2.2 Terminology

The only unique strength of randomisation as an element in the design of treatment comparisons is that, if successfully accomplished, it eliminates selection bias at the point of trial entry. Successful abolition of selection bias at trial entry depends on successfully fulfilling two interrelated, prior conditions. First, an unpredictable 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 this 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.

Simple (unrestricted) randomisation 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 usually it is achieved by referring to a printed list of random numbers, or to a list of random assignments generated by a computer. Simple randomisation generates unbiased comparison groups, regardless of sample sizes. In trials with large sample sizes, the comparison groups would probably end up being of relatively similar sizes. In trials with small sample sizes, however, simple randomisation could result in comparison groups that would differ relatively more in sizes.

Balanced (restricted) randomisation 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 randomisation 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.

Other approaches to restricted randomisation include the 'biased coin' and 'replacement randomisation' methods (Pocock 1983). Briefly, the 'biased coin' approach alters the allocation probabilities during the course of the trial to rectify imbalances that may be occurring. Replacement randomisation involves the repetition of a simple randomisation allocation scheme until a desired balance is achieved.

To achieve benefits from pre-randomisation stratification, restricted randomisation is used to generate separate randomisation schedules for stratified subsets of participants defined by potentially important prognostic factors (for example, disease severity and study centres). Another good approach called 'minimisation' incorporates both the general concepts of stratification and restricted randomisation (Pocock 1983). It can be used to make small groups closely similar with respect to several characteristics.

A less than optimal, but probably acceptable, restricted randomisation 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.

The restricted shuffled approach is less than optimal for at least three reasons. First, shuffling cards determines the allocation sequence rather than a random number table or generator. Second, while correctly implemented envelopes can be adequate, they are more open to deciphering than other approaches. And third, balance would only usually occur at the end of the trial and not throughout.

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 this may give a falsely optimistic impression. For example, in some populations, the day of the week on which birth occurs is not a matter of chance (Macfarlane 1978). 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 might entail undertaking a separate, more time-consuming study than the primary substantive study planned. If authors report using systematic allocation, readers should be wary of the results. Investigators conducting trials should not expose their work to such scepticism by using systematic methods. Using an appropriate, random approach is easier in the short- and long-run, and reproducible.

An even more important weakness with systematic methods centres on the near impossibility of concealing the assignment schedule. That allows foreknowledge of treatment assignment among those recruiting participants to the trial. For that reason the **British Medical Journal** has recently decided not to publish trials that have used such allocation schemes when randomisation was feasible (Altman 1991).

2.3 Randomisation without foreknowledge

Randomisation encompasses both the approach to generating the random sequence and the approach to concealing those assignments until the point of actual treatment allocation. While generating a random sequence is important, concealment of assignments from everyone involved in the trial may contribute more to producing randomisation without foreknowledge (Peto 1987).

Reducing bias in trials crucially depends upon preventing that foreknowledge. 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 revealed, 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 allocating treatment.

The process of concealing assignments until treatment has been allocated has sometimes confusingly been referred to as 'randomisation blinding' (Chalmers 1983). The main reason for that term being unsatisfactory is that, on the rare occasions that it has been used at all, it has too seldom been distinguished clearly from other forms of blinding (masking) - of patients, physicians, outcome evaluators, and analysts. That 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 (masking) after allocation is not attainable in many instances, such as in trials comparing surgical with medical treatments. Third, control of selection bias is relevant to the trial as a whole, and thus to whatever outcomes are being compared, whereas control of ascertainment bias is often 'outcome-specific', that is, it may be accomplished successfully for some outcomes in a trial, but not for others. Thus, 'blinding' up to the true allocation of treatment and blinding after it are addressing different sources of bias, are inherently different in their practicability, and may apply to different parts of a trial. In light of these reasons for distinguishing the different forms of blinding clearly, the process of concealing assignments will be referred to as **'allocation concealment'** or **'randomisation concealment'**, and the term 'blinding' (masking) will be reserved for measures taken to reduce bias after treatment has been assigned.

Investigators likely achieve allocation concealment if a randomly generated assignment schedule is administered by someone who is not responsible for recruiting participants, for example, someone based in a trial office, or pharmacy. If organising allocation 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. Those include, for example, using numbered or coded bottles, ampoules or other containers, or using serially numbered, sealed, opaque envelopes

(Altman and Doré 1990; Pocock 1983; Mosteller 1980). Simply using an open list ('table' or 'schedule') of random assignments is as open to manipulation as is dependence on one of the systematic methods of assignment.

2.4 Review of previous relevant research

2.4.1 Non-randomised comparisons and bias

The elimination or reduction of bias and the production of comparable treatment groups exemplify good RCT methodology. If that is indeed true, bias should be detectable in non-randomised as compared to randomised studies. That has been illustrated in a number of instances. One example of bias in historical comparisons is provided by studies of anticoagulant therapy in the hospital phase of an acute myocardial infarction for which 18 trials with historical controls and 6 with randomised controls were reviewed (Chalmers et al. 1977; Pocock 1979). The historically controlled trials with over 8,000 patients resulted in a 53% reduction in mortality for anticoagulant therapy compared with no treatment, whereas the RCTs involving almost 4,000 patients showed only a 20% mortality reduction. "Although both types of study confirm that anticoagulant therapy improves survival, it is disturbing that the average bias incurred by using historical controls is of the same order of magnitude as the observed treatment effect" (Pocock 1979).

In another example, the literature was searched to identify therapies studied by both RCTs and historically controlled trials (HCTs). Six therapies were found for which 50 RCTs and 56 HCTs were reported (Sacks et al. 1982). Forty-four of 56 HCTs

(79%), found the therapy better than the control regimen, but only 10 of 50 RCTs (20%) agreed. The same trend emerged when the authors examined each of the six therapies separately. They concluded that the data suggest that biases in patient selection may irretrievably weight the outcome of HCTs in favour of new therapies. While Sacks et al. appeared to examine similar outcome measures between RCTs and HCTs, they did not clarify the degree of similarity. Any potential dissimilarity would be unlikely, however, to account for the large discrepancies observed.

Another study examined controlled clinical trials of the treatment of acute myocardial infarction (Chalmers et al. 1983). Fifty-seven randomised studies were compared with 43 non-randomised studies, which included the use of both simultaneous and historical controls. Differences in case-fatality rates between the treatment and control groups at $p<0.05$ were found in 8.8% of the randomised studies and 58% of the non-randomised studies. In these statistically significant studies, results favoured the treatment group over the controls in 60% and 93% of the studies, respectively. While the large difference in the proportion of statistically significant studies could partially be a function of larger sample sizes in the non-randomised studies as well as a function of bias, bias was clearly evident when the mean differences (± standard error) in case fatality rates were examined: 0.003 ± 0.008 (not significant) for the randomised group and 0.105 ± 0.017 ($p<0.001$) for the non-randomised group. Similarly, other examinations of medical and surgical disciplines have observed that non-randomised studies tend to report larger gains than do the randomised studies (Miller et al. 1989; Colditz et al. 1989). Thus, that randomised versus non-randomised comparisons reduce bias appears well documented and generally accepted by the scientific community.

2.4.2 Methodological rigour and bias in randomised comparisons

Less supportive information exists on methodologically sound randomisation schemes reducing or eliminating bias compared to less rigorous or haphazard schemes. One notable exception is the study that examined the controlled clinical trials of the treatment of acute myocardial infarction (Chalmers et al. 1983). The authors compared 57 papers in which the randomisation process was concealed (more methodologically rigorous) with 45 papers in which it was not concealed (less methodologically rigorous). Differences in case-fatality rate between treatment and control groups at $p<0.05$ were found in 8.8% and 24.4% of the trials, respectively. That large discrepancy could have been partly the result of larger sample sizes as well as bias in the less methodologically rigorous randomised trials. The sample sizes were not provided in that report, however, but they likely approximated those of the more methodologically rigorous trials since all were controlled trials. If anything, less rigorous trials usually tend to have smaller sample sizes.

Bias seems the more likely explanation. The authors examined the mean differences (± standard error) in case-fatality rates: 0.003 ± 0.008 (not significant) for the concealed randomisation group and 0.052 ± 0.016 ($p<0.001$) for the unconcealed randomisation group. The disparity between the mean differences is a clear indication of potential bias in the unconcealed randomisation group. Moreover, in the statistically significant studies, results favour the treatment group over the controls in 60% and 100%, respectively. The authors conclude that the "data strongly suggest that bias in treatment assignment could be a more important determinant of outcome than the treatments under investigation."

While these results appear convincing, one caveat should be considered. Whereas the outcome, death, is the same for all studies included in the analysis, a number of different treatments were aggregated for the analysis and the distributions of these treatments for the concealed randomisation category and the unconcealed randomisation category were different. Thus, different treatments as well as bias could be contributing to the differences between the two randomisation groups (Gillman and Runyan 1984).

Another study addressed this flaw by comparing similar treatments and similar outcome measures (Emerson et al. 1990). The authors examined seven meta-analyses using general linear models, which combined features of analysis of variance and regression, and introduced a categorical variable that distinguished the seven topics from one another. Thus, they essentially controlled for each meta-analysis and compared similar treatments with similar outcome measures. The authors found that an overall measure of RCT quality, used as previously reported (Chalmers at al. 1981), did not have a statistically significant relationship to the primary study's treatment differences within the meta-analyses.

That apparent lack of a relationship in the Emerson et al. study may have at least two potential explanations. First, and most important, the authors used a measure of quality (Chalmers et al. 1981) that quantifies many facets of trial design and analysis, some of which were not likely to be related to bias. Second, they included a measure of study size as an independent variable along with the measure of quality in their linear model. However, study size correlated with quality score so it, through difficulties with multicollinearity, could have obscured the effect of other aspects of quality.

The authors also examined the relationship between the variability of the treatment difference and overall quality score, i.e. attempting to determine whether lower quality studies tend to produce more variable estimates of treatment differences. That lower methodological quality would produce bias in both directions and thereby cause increased variability appears plausible and a worthwhile hypothesis to test. They concluded that no evidence was found for greater variability but, again, the methodology they used may have been insufficiently sensitive. Their paper did not clearly specify whether they had controlled for the individual meta-analyses in this instance.

Chalmers et al. (1983) and Emerson et al. (1990) conducted perhaps the most comprehensive scientific investigations to date of methodological quality and bias in RCTs. The earlier study was persuasive of a relationship whereas the more recent was inconclusive, perhaps partly due to its design.

Secondary analyses from meta-analyses have examined the relationship between study quality and the estimate of the magnitude of the difference between study groups compared. In some, no relationship was found, but in a greater number a relationship was found (Chalmers et al. 1989). One recent meta-analysis of low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery observed the relationship in general surgery but not in orthopaedic (Nurmohamed et al. 1992). Apart from a few exceptions, associations between methodological quality of studies and estimates of the size of difference between experimental groups within them usually occur because methodologically inferior studies tend to yield larger differences than those that have employed sounder methods for reducing bias (Chalmers et al. 1989). However, the preponderance of these investigations were not analyzed statistically and were subject to post-hoc reasoning and, thereby, largely serve a hypothesis-generating function.

The Chalmers et al. (1983) study specifically examined the association between the concealment of the allocation schedule and estimates of treatment effects. Unfortunately, other important, specific, methodological components of trials, such as blinding or the handling of exclusions, appear not to have been similarly examined.

Chapter 3

3. Assessing the Quality of Randomisation and Allocation from Reports of Controlled Trials Published in Obstetrics and Gynaecology Journals

3.1 Summary

3.1.1 Objective. To assess the quality of randomisation, allocation, and other associated methodological components from reports of trials in obstetrics and gynaecology journals.

3.1.2 Methods. Evaluation of all 206 reports of parallel group trials, in which authors stated allocation to have been randomised, published in the 1990 and 1991 volumes of four journals of obstetrics and gynaecology.

3.1.3 Results. Only 32% reported having used an adequate method to generate 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% of the reports of trials published in the obstetrics and gynaecology journals described adequate methods of randomisation. In reports of trials which had apparently used unrestricted randomisation, the differences in sample sizes between treatment and control groups were much smaller than would be expected by chance. In reports of trials in which hypothesis tests had been used to compare baseline characteristics, only 2% of reported tests were statistically significant, lower than the expected rate of 5%.

3.1.4 Conclusions. The generation of unbiased comparison groups in controlled trials requires proper randomisation. Yet, authors usually provided inadequate information or described an inadequate approach. Additional analyses suggested that non-random manipulation of comparison groups and selective reporting of baseline comparisons may have taken place.

3.2 Introduction

Randomisation avoids selection biases in controlled trials of prevention and treatment. Over forty years ago, Austin Bradford Hill (1952) 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".

Methodologically rigorous approaches to randomisation, however, must be used. If investigators improperly implement or inadequately describe the methodology, the results, and thereby conclusions, emanating from those trials should be considered as potentially biased (Mosteller et al. 1980; Altman and Doré 1990). Furthermore, while randomisation allocates treatment without bias, it does not necessarily produce similar groups on important prognostic factors. Chance imbalances can occur. The similarity of baseline characteristics should be established, with hypothesis tests not being the criterion (Altman 1985). Not only should participants be allocated to treatment groups properly, but the sample sizes should be properly planned as well (Freiman et al. 1978).

Unfortunately, investigators often improperly address the process of randomisation in the design and implementation phases of controlled trials, and authors often neglect it in published reports. For example, in 132 reports of trials on cancer topics, only a third of the authors reported how the randomisation had been carried out (Mosteller et al. 1980). Moreover, many of the methods specified were, in fact, non-random, such as using date of birth, using admission number or identification number, or using an odd-even method of assignment.

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 randomisation (DerSimonian et al. 1982).

An analysis by Altman and Doré (1990) of reports of trials published during 1987-8 in the **BMJ**, the **Lancet**, the NEJM, and the **Annals of Internal Medicine** 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 (Altman and Doré 1990). While perhaps an improvement compared with the results from the early 1980s, obviously much capacity remains for further improvement.

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 (Altman and Doré 1990). Overall, 600 tests were reported, of which only 4% were statistically significant at the 5% level. Furthermore, their analysis also illustrated that among trials that used simple randomisation, the sample sizes in the two comparison groups were too often similar (Altman and Doré 1990).

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 (Pignon and Poynard 1990). Apparently, no systematic analysis of randomisation and allocation of journals from obstetrics and gynaecology has been undertaken, although assessments of reports of trials in that field have indicated that the methodological quality may indeed be worse than that of reports published in general medical journals (Tyson et al. 1982; Thacker 1987; Keirse 1988; Grimes and Schulz 1992). I undertook a study, resembling Altman and Doré's, of reports of randomised controlled trials (RCTs) in obstetrics and gynaecology.

The systematic evaluation undertaken was conducted of reports published in the two main American and the two main British journals of obstetrics and gynaecology. The **American Journal of Obstetrics and Gynecology** (AJOG) and **Obstetrics and Gynecology** (OG) emanate from the United States and the **British Journal of Obstetrics and Gynaecology** (BJOG) and the **Journal of Obstetrics and Gynaecology** (JOG) from the United Kingdom.

The reported approaches to treatment assignment and to comparison of baseline characteristics were analyzed. As indicated above, reports in the obstetrics and gynaecology journals were suspected to be of lower quality than reports published in the general medical journals. Furthermore, three primary hypotheses were proposed. **First, the suspicion that the reports published in the BJOG would be of better quality** than those published in other journals of obstetrics and gynaecology. A concerted editorial effort had been made to improve the quality of reporting in that journal, including publication of a series of articles providing reporting standards for different types of studies (Bracken 1989; Wald and Cuckle 1989), including trials (Grant 1989). Second, the notion that, as had been demonstrated using reports of trials published in the general medical journals (Altman and Doré 1990), the numbers of patients in the comparison groups of trials that had apparently used unrestricted randomisation would be more similar than expected by chance. And third, the impression that, as had been suggested by the earlier study (Altman and Doré 1990), the percentage of reported statistically significant differences in characteristics measured at baseline would be less than the expected 5%.

3.3 Methods

3.3.1 Study material

Data were collected from 206 reports of trials published in the 1990 and 1991 volumes of the **AJOG**, the **BJOG**, the **JOG**, and **OG**, using a hand search to try to ensure that all the eligible reports were identified. In addition, both the Oxford Database of Perinatal Trials (Chalmers 1992; Chalmers et al. 1986) and MEDLINE were searched as a cross-check. The study was restricted to reports of parallel group (uncrossed) trials, comparing two or more treatments, in which allocation was stated to have been randomised. 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 randomised trial, even if the actual method of allocation described was non-random. Where possible, the actual method of allocation was recorded. Reports that were not the first publications relating to particular trials were excluded.

The reports were examined and the data collected using methods similar to those used in Altman and Doré's (1990) analysis of general medical journals. The data collection instrument was tested in a pilot study (articles from 1989 in the same journals) before proceeding with the assessments. For consistency of measurement across journals, I did all of the assessments. To examine the reproducibility of items on the questionnaire, David A. Grimes, M.D., Professor of Obstetrics and Gynecology at the University of California, San Francisco, assessed a random sample (random number table) of 15 trials blinded to the results of the initial assessments. We found no notable differences on the main outcome measures. The data were entered interactively into an EPI-INFO questionnaire with on-line editing, skip-pattern, and logic-checking capability (Dean et al. 1990).

3.3.2 Aspects of the analytical approach

The analysis of the differences in numbers of participants reported to have been assigned to comparison groups has been limited to two-group trials which were apparently 'unrestricted'. Trials were categorized as 'unrestricted' if they met all the following criteria: 1) the trial had not been reported to have been restricted; 2) the type of randomisation 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 blocked).

Comparison of baseline characteristics of the treatment groups is an important first step in trial reporting. Although randomisation assigns treatments without selection bias, it does not necessarily produce similar groups on important prognostic factors. **Chance imbalances** can and do occur. The probabilistic argument is that, on average, randomised groups will have the same characteristics. In practice, however, a particular trial may have one or more characteristics unequally split between groups. Large studies generate serious imbalances less frequently, but smaller studies using simple randomisation remain susceptible to substantial covariate imbalances (Lavori et al. 1983).

Such imbalances in baseline characteristics cause concern, however, only when they involve characteristics related to outcome variables of prognostic importance. Then 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 (Rothman 1977; Lavori et al. 1983; Altman 1985). If means or medians for continuous variables are reported, appropriate information about variability should also be reported, e.g. the standard deviation, range, or raw data.

In evaluating presentations of variability in this chapter, if some baseline characteristics within a particular report had variability presented and others had not, the assessment was based on the method used when variability had been presented. 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. Only data were used for which the authors had presented test results.

For the reasons offered in section 2.3, the process of concealing assignments is referred to as **'allocation concealment'** or **'randomisation concealment'**, and the term 'blinding' (masking) is reserved for measures taken to reduce bias after treatment has been assigned.

3.3.3 Statistical methods

Unless otherwise indicated, chi-squared tests were used for comparing nominally scaled variables. The Greenland and Robins approach in EPI-INFO was used to obtain confidence intervals for relative risks (Dean et al. 1990). Because Bartlett's test for homogeneity of variance was typically statistically significant at $p<0.05$, Kruskal-Wallis one-way analysis of variance tests were used to compare continuous variables among journals. Moreover, Kruskal-Wallis is a non-parametric test which does not assume normal distributions, yet it retains most of the power of a parametric test.

To facilitate direct comparison of the reports published in these obstetrics and gynaecology journals with those published in general journals, frequent reference is made to Altman and Doré's (1990) article on the general medical journals in the discussion section. Henceforth, the results from that article are referred to as being from the 'general journals' without constant duplicate referencing.

3.4 Results

3.4.1 Source of reports

Of the 206 reports of trials published in the obstetrics and gynaecology journals, 64 were found in the **AJOG**, 48 in the **BJOG**, 20 in the **JOG**, and 74 in **OG**. Other trials in these journals used systematic methods, but they were not included because they did not purport to be randomised.

3.4.2 Type of randomisation

Over three-quarters (78%) of the reports of trials failed to provide information about the way that the treatment assignments had been generated (Table 3.1). Moreover, 11 reports (5%), about a quarter of those providing any information at all, clearly stated that a systematic method of assignment had been used, despite their claims to be reporting randomised trials.

Only 29 (14%) of the reports described the use of restriction, and of the 23 reports describing 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 minimisation. No reports stated the use of replacement randomisation.

Reports published in the **BJOG** stated the type of randomisation more frequently than reports published in the other journals ($p<0.001$, 3 df). 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 randomisation (p<0.001, 3 df). Only four trial reports stated that simple unrestricted randomisation had been used.

Table 3.1

The type of randomisation stated in the four obstetrics and gynaecology journals, 1990 and 1991

Type of randomisation 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)
Systematic	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)

3.4.3 Stratification

Only nine per cent of the reports of trials in the obstetrics and gynaecology journals reported the use of stratification (Table 3.2), and fewer than half of those reported the use of blocking or minimisation.

Table 3.2

Stratified trials and those stratified trials that report being blocked in the four obstetrics and gynaecology 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 minimisation

3.4.4 Methods for generating random numbers

Only 32% of reports specified an adequate method for generating random numbers, with the rates being similar for the four journals (p=0.27, 3 df)(Table 3.3). A computer random number generator was the most frequently specified method (18%), followed by a random number table (11%). Other random processes used in 4% of the trials primarily included shuffled cards and tossed coins.

Table 3.3

Methods for generating allocation sequences

Method for generating sequence	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 process*	3% (2)	8% (4)	0% (0)	3% (2)	4% (8)
Systematic	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

3.4.5 Treatment allocation methods

Almost half (48%) of the reports of trials did not describe the mechanism used to allocate treatments (Table 3.4). A quarter of the reports described the use of envelopes, but only a quarter of those reports stated that the envelopes had been sequentially numbered, opaque, and sealed. 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 randomisation had been organised. 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 systematic assignment procedure had been used.

Overall, only 23% of the reports published in the obstetrics and gynaecology journals (Table 3.4) reported an adequate approach to randomisation concealment. The proportion of trials in which adequate randomisation concealment appeared to have been achieved varied markedly (p<0.001, 3 df) among the four journals (Table 3.4). 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).

Table 3.4

Allocation concealment methods

Allocation method	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
Numbered or coded containers*	8% (5)	10% (5)	0% (0)	7% (5)	7% (15)
Pharmacy concealed*	6% (4)	10% (5)	0% (0)	8% (6)	7% (15)
Centrally concealed* e.g. telephone	2% (1)	6% (3)	0% (0)	1% (1)	2% (5)
Sequentially numbered, opaque, sealed envelopes*	3% (2)	17% (8)	5% (1)	3% (2)	6% (13)
Envelopes -- other	20% (13)	25% (12)	10% (2)	16% (12)	19% (39)
List, table, or schedule	9% (6)	4% (2)	0% (0)	3% (2)	5% (10)
Systematic	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)

* Adequate allocation concealment method

3.4.6 Overall quality of randomisation and allocation

Fifty-one reports of trials (25%) 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 randomisation concealment. The proportions for each were 15% for the BJOG, 9% for the AJOG, 7% for OG, and 5% for the JOG, but the differences among these proportions were not statistically significant (p=0.46, 3 df).

3.4.7 Relative size of treatment groups in apparently unrestricted trials

In 96 reports of apparently unrestricted trials, the differences in sample sizes between the treatment and control groups were much smaller than would be expected by chance alone. In Figure 3.1, 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). A further indicator of the similarity of group sizes is that 54% of the unblocked trials had differences in group sizes of zero or one. Surprisingly, the blocked trials yielded differences that were less similar overall, with only 36% of the trials having differences in group sizes of zero or one.

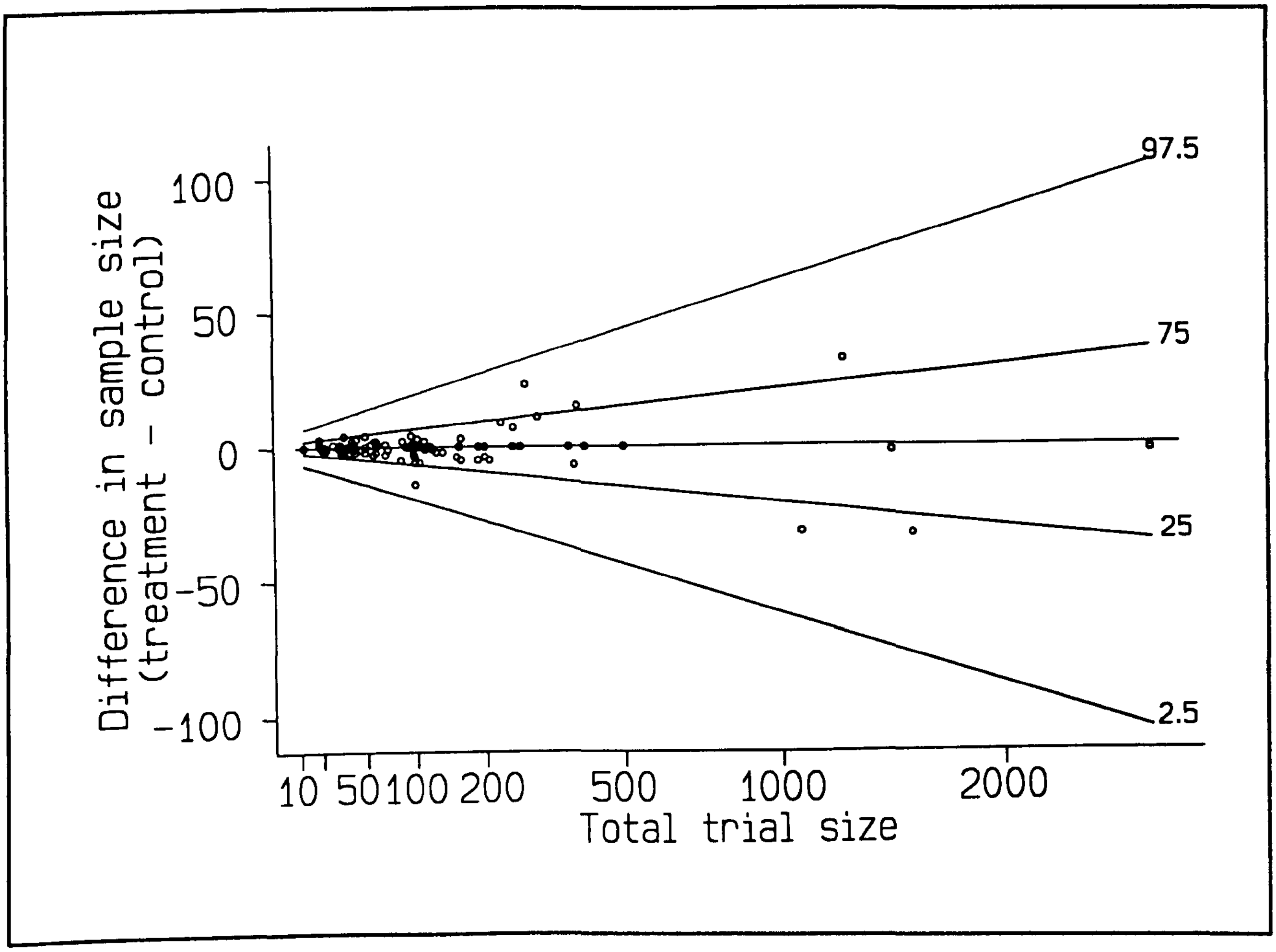

Figure 3.1: The relationship between the difference in sample sizes in the treatment and control groups and total study size for 96 unblocked trials. The straight lines represent the expected distribution due to the play of chance. Total study size is shown on a square root scale to make the confidence interval lines straight. The 95% confidence interval is approximately:

$\pm 2\sqrt{\text{Total Study Size}}$ (Altman and Doré 1990).

3.4.8 Comparisons of baseline characteristics

Comparisons of baseline characteristics were presented in 84% of the reports (Table 3.5). They 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, the differences were not statistically significant (p=0.17; 3 df).

The median numbers of comparisons of baseline characteristics in those reports in which comparisons were presented was 6 (Table 3.5). Reports in the **AJOG** and **OG** tended to present a larger number of comparisons (p=0.008, 3 df).

Comparisons of baseline characteristics presented as continuous variables were reported in 78% of the trials (Table 3.5). In those, 68% were accompanied by appropriate measures of variability. Reports in the **BJOG** were more likely than those in the other obstetrics and gynaecology journals to present appropriate measures of variability, but the differences among the four journals were not statistically significant (p=0.22; 3 df). In 41% of the reports, overall, either authors did not present baseline characteristics, or did not report appropriate measures of variability.

Table 3.5

Baseline characteristics

Baseline characteristics	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
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)
Continuous variables only:					
≥1 Continuous variable reported 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)

* Standard deviation, range, centiles, or raw data

† Standard error or confidence interval

3.4.9 Use of hypothesis tests to compare baseline characteristics

Hypothesis tests were used to compare baseline characteristics in 61% of the reports (Table 3.6). Hypothesis tests were presented far more often in the American journals than in the British (p<0.001, 3 df). Overall, 1,076 hypothesis tests were presented in 125 reports. Only 2% of those were statistically significant at the 5% level; that itself is a statistically significant departure from expectation (p<0.001, z-test).

Table 3.6

Use of hypothesis tests (tests of statistical significance) to compare baseline characteristics

Comparing baseline variables	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
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 (SD)	10.3 (7.4)	6.2 (5.1)	4.0 (2.2)	8.3 (5.8)	8.6 (6.4)
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)

3.4.10 Power calculations

In 50 (24%) of the reports, 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. Trials published in the BJOG 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).

3.5 Discussion

3.5.1 Importance of randomisation

Randomisation is the only reliable way to create comparable comparison groups with respect to unmeasured or imperfectly measured prognostic factors. For that reason, RCTs have been 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 randomisation 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 randomisation can be guaranteed for all trials. By contrast, other measures used to control bias can not 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 randomisation. If investigators cannot conceal allocations up to the point of assignment, they would have difficulty blinding after, and the concept underlying intention-to-treat analyses is simply the preservation of the randomised allocation.

Considering how centrally important randomisation is to any assessment of the validity of a treatment comparison, it is surprising that authors and editors have not been more meticulous in publishing clear reports of the process used to assign participants to comparison groups.

3.5.2 Treatment allocation in the obstetrics and gynaecology journals with a comparison to that in general medical journals

A non-random method accounted for 5% of the 'randomised' trials published in the obstetrics and gynaecology journals. That rate is at the lower end of the range (5-10%) found in earlier surveys of reports of 'randomised' trials (Mosteller et al. 1980; Evans and Pollock 1984; Chalmers et al. 1986), but substantially higher than the rate of 1% found in Altman and Doré's (1990) more recent study of general medical journals.

Blocking was reported in only 11% of trials reported in the obstetrics and gynaecology 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 a third of these reports in both types of journals. Stratification was mentioned explicitly much less frequently in the obstetrics and gynaecology journals (9%) than in the general medical journals (39%). However, in both types of journals, 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 have been stated explicitly because stratification is not effective unless blocking, or some form of restriction, has been used as well. The authors of only four trials reported simple randomisation, but a majority of the trials in which an approach had not been stated explicitly had probably used simple randomisation (Altman and Doré 1990).

The method of generating random numbers was less well reported in the obstetrics and gynaecology than in the general journals. In only about a third of the reports in the obstetrics and gynaecology journals was it concluded that an acceptable approach had been used, as compared to almost half of the reports in the general journals. Moreover, those are generous estimates, since they include processes such as shuffled cards and tossed coins as adequate. Because those methods are subject to human perturbations in the production of allocation schedules and are not reproducible, they are certainly less than optimal (Mosteller et al, 1980), if not unacceptable (Meinert 1986). Random number tables and computer random number generators are recommended not only for being reproducible, but also for being easier and faster.

In the process of allocating treatments such that foreknowledge of the allocation is prevented, allocation concealment is generally more important than generation of the randomised assignments per se (Bradford Hill 1952; Chalmers et al. 1987b), yet only 52% of the reports in the obstetrics and gynaecology journals, and 56% of those in the general medical journals, provided information adequate to assess that aspect of trial design and conduct. Many of those stated approaches were judged to have been inadequate, however, and even with those judged to have been adequate, many reports should have provided further important, clarifying information (see section 6.4.4). In sum, only 23% of the studies reported in the obstetrics and gynaecology journals as compared to 26% of the studies reported in the general journals appear to have used an adequate approach to allocation concealment. Overall, only 9% of the reports of trials published in the obstetrics and gynaecology journals clearly stated that adequate methods of both random number generation and allocation concealment had been used.

While descriptions of the process of treatment assignment were of generally poor quality in both the obstetrics and gynaecology and the general journals, the specialty journals were, on average, somewhat less satisfactory. Remember, however, that the time frames for the assessments were different. The data from the general journals was collected from reports published 3 to 4 years earlier than those from the obstetrics and gynaecology journals. Some of the general journals, **The Lancet** for example, have instituted new statistical review procedures which would have likely produced more favourable results if new data had been collected concurrently with the obstetrics and gynaecology journal data collection. Thus, the actual disparity between these specialist and general journals is probably even greater than the disparities reported.

3.5.3 Descriptions of treatment allocation in the BJOG compared with those in the other three journals

Reports of trials published in the **BJOG** were more informative than those in the other three obstetrics and gynaecology 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 randomisation, 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 three 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**. Editorial efforts similar to those made at the **BJOG** in the mid-1980s are now occurring at **OG** (Grimes 1991), and those too may result in improved quality of reports.

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. Much room for improvement, however, still remains in the **BJOG**. Some rather basic errors of commission and omission continued to be made.

3.5.4 The numbers of participants in the comparison groups were too often similar in trials that had used unrestricted randomisation

Restricting randomisation 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 randomisation 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 this will not usually have an important effect on the power of the study (Altman and Doré 1990).

The relative sizes of comparison groups in trials in which simple randomisation has been used should reflect random variation. In other words, some discrepancy between the numbers in the comparison groups is to be expected. The finding that the reported sizes of the comparison groups tended to be much too similar, however, confirmed the earlier finding from general journals. Not only were the similarities found very unlikely to be due to the play of chance, they were even more similar than those revealed in the earlier study.

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 randomisation; (3) failure to report the use of a restricted shuffled envelope method; (4) failure to report the use of a systematic method of assignment, such as alternation or odd-even date; or (5) 'rectification' of an imbalance in sample sizes by non-random manipulation of assignments or data.

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 particular, since the blocked trials yielded **more** disparate differences than the unblocked trials. Replacement randomisation would also be an acceptable explanation, but no evidence for its use was found. A less acceptable alternative would be that the restricted shuffle approach had been used (Mosteller et al. 1980). Only 5% of the trials that specified a type of randomisation used that method, however, so this seems an unlikely explanation for the similarity in the reported size of the comparison groups. A more likely explanation is the use of systematic allocation. Of the reports in which a method of generation was stated, 14% used this approach. Thus, the unidentified use of systematic allocation may explain at least some of the similarities in the numbers of participants assigned to the comparison groups. Unfortunately, that explanation is not reassuring since systematic allocation is both non-random and unlikely to be concealed.

The last possibility, non-random 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. Indirect evidence now exists from these journals and from the general medical journals that non-random manipulation may have sometimes occurred. Possibly some investigators believed that they would increase the credibility of their trial reports 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 this study. While these results clearly indicate that the set of trials supposedly using unrestricted randomisation are not what they purport to be, one cannot identify any particular trials as suspect, as some trials would be expected to achieve almost equal numbers simply by chance.

3.5.5 The percentage of statistically significant differences in baseline characteristics materialized as less than the expected level of 5%

On strictly theoretical grounds, if randomisation is properly implemented, establishment of comparability at baseline is not necessary. 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, randomisation 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 randomisation underlies significance testing, and that process is independent of prognostic factors, known or unknown (Fisher 1966).

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 randomisation 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 it may also increase precision (Lavori et al. 1983).

Although comparisons of baseline characteristics were presented in a majority of the reports published in the obstetrics and gynaecology journals (84%), 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 higher proportion (48%) of the reports published in the general medical journals were deficient in this respect.

An even more worrying deficiency, which was present in 61% of the reports in the obstetrics and gynaecology 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 randomised trials, it is known already that any differences observed have occurred by chance. As noted by Altman (1985) 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. 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 (Altman 1985).

Although use of hypothesis tests inappropriately addresses the assessment of baseline imbalances in prognostic characteristics (Altman 1985), these tests might, in principle, be used by investigators who are concerned that randomisation may not have been executed effectively in their studies (Altman 1985). Sometimes gross imbalances, quite incompatible with random variation, are revealed in this way (Keirse 1988). Finding several statistically significant differences between the comparison groups may suggest that randomisation has not been achieved; but use of tests in this way will often pose substantial problems of interpretation.

Concern that randomisation 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 which test results were presented: only 2% of the tests reported were statistically significant at the 5% level, a discrepancy from expectation which is unlikely to reflect chance. That observed frequency of statistically significant test results in the obstetrics and gynaecology journals is more extreme in its departure from the expected than the value of 4% reported by Altman and Doré (1990) from general journals.

A plausible explanation for these 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 with the aggregate analysis in this study. Having too few statistically significant results has hurt the credibility of these trials. Investigators should 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 suppress any baseline imbalances.

3.5.6 Conclusions

None of the findings reported in this chapter are particularly 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 (Liberati et al. 1986). Thus, while reporting clearly must be improved, the deficiencies in the design and conduct of trials must also be urgently addressed.

Although, as predicted, descriptions of the process of generating and applying treatment assignments in reports of trials published in obstetrics and gynaecology 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 gynaecology journals (and of comparable quality to those in the best general medical journals), considerable scope for improvement exists.

It was confirmed that the numbers of participants in the comparison groups of trials which have apparently used simple randomisation were too often too similar, and that the observed percentage of statistically significant differences in characteristics measured at baseline was less than the expected value of 5%. Those are disturbing findings in that they suggest the occurrence of some non-random manipulations of comparison groups and some selective reporting of baseline comparisons.

Chapter 4

4. The Association Between Methodological Quality and Treatment Effects in Controlled Trials: An Analysis of 250 Trials from 33 Meta-Analyses

4.1 Summary

4.1.1 Objective. Most reports of randomised controlled trials contain inadequate methodological descriptions. The primary aim was to explore the association between those inadequate descriptions and estimates of treatment effects.

4.1.2 Methods. The methodological quality of the trials was assessed on major dimensions and then the associations between those assessments and estimated treatment effects were analyzed. The analysis used 250 trials from 33 meta-analyses encompassing 62,091 participants and employed multiple logistic regression models.

4.1.3 Results. Compared with trials in which authors reported adequately concealed treatment allocation, trials in which authors reported inadequately or unclearly concealed allocation yielded larger estimates of treatment effects ($p<0.001$). Odds ratios were distorted by 41% or 33%, respectively. Trials in which participants were excluded after randomisation were not, however, associated with larger treatment effects, but that appears to be due to incomplete reporting rather than to the unimportance of adequately handling exclusions. Lack of double-blinding in trials was also associated with larger treatment effects ($p=0.01$), albeit at a lower strength than the lack of adequate measures to conceal assignment. Furthermore, inadequately concealed trials displayed greater variability (heterogeneity) compared with the adequately concealed trials.

4.1.4 Conclusions. Inadequate methodological approaches in controlled trials, particularly those reflecting inadequate or unclear allocation concealment, were

associated with reported larger treatment benefits. Those associations likely represent bias. To be successful at minimising bias, investigators must properly design and meticulously execute their trials.

4.2 Introduction

When conducting a controlled trial, properly implemented randomisation avoids selection biases. Yet, as shown in Chapter 3, many trials in journals of obstetrics and gynaecology may be inadequately executed. Are those trials associated with biased estimates of treatment effects? Presumably, investigators could introduce bias into inadequately executed trials, but meagre quantitative evidence supports that presumption.

While only meagre evidence is available on proper versus improper randomisation, more documentation supports the notion that randomised as compared with non-randomised comparisons reduce bias (Chalmers et al 1977; Pocock 1979; Sacks et al 1982; Chalmers et al 1983; Miller et al 1989; Colditz et al 1989). In those analyses, the trend is for the non-randomised trials to report larger estimates of treatment effects than do the randomised trials. That poorly-executed versus well-executed 'randomised' trials would follow the same trend seems logical.

Indeed, the results of one such analysis did follow that same trend (Chalmers et al. 1983). Apparently, bias may be introduced into trials labelled as 'randomised' if investigators do not take precautions to conceal treatment allocation schedules from those responsible for recruiting and entering participants. As reviewed in section 2.4.2, an analysis published in 1983 found that trials in which the treatment allocation schedule had been inadequately concealed yielded larger estimates of treatment effects than trials in which allocation had been adequately concealed (Chalmers et al. 1983). However, while the larger estimates of treatment effects may have reflected biases resulting from foreknowledge of treatment allocation, they also may have reflected the confounding effects of the different treatments represented among the two categories of trials (Gillman and Runyan 1984).

In an attempt to address that potential source of confounding, a later study restricted an analysis to trials that had evaluated comparable treatments in terms of similar outcome measures (Emerson et al. 1990). They did not, however, detect a relationship. They also suspected that low quality trials might produce bias in both directions thereby causing greater variability in estimates of treatment effects. Again, they did not detect evidence of such.

As reviewed in section 2.4.2, the authors of the Emerson et al. (1990) study, rather than concentrating only on adequacy of the concealment of allocation schedules, sought a relationship between 'quality scores' intended to characterize the overall methodological quality of each trial (Chalmers et al. 1981) and estimates of treatment effects. That and other methodological aspects of their approach may have handicapped their ability to accurately assess the relationship.

Using a database of systematic reviews of controlled trials in pregnancy and childbirth

(Enkin et al. 1993), I conducted a retrospective observational study to explore the associations between estimates of treatment effects and methodological approaches intended to control bias. The hypotheses tested were that estimates of treatment effects would be larger in trials in which (i) inadequate measures had been taken to conceal treatment allocation; (ii) some participants allocated had been excluded; (iii) measures had not been taken to implement double-blinding; or (iv) inadequate measures had been taken to generate the allocation schedule, given that allocation had been adequately concealed. In addition, analyses included whether estimates of treatment effects varied more in trials in which treatment allocation schedules had not been adequately concealed.

4.3 Methods

4.3.1 Derivation of study material

The systematic reviews of controlled trials used in this analysis have been published in printed and electronic forms by the Pregnancy and Childbirth Group of the Cochrane Collaboration (Chalmers et al. 1989; Chalmers 1992; Enkin et al. 1993). Published and unpublished primary trials potentially relevant for inclusion in the reviews were sought and entered in a register, the Oxford Database of Perinatal Trials, using methods that have been described in detail elsewhere (Chalmers et al. 1986; Hetherington et al. 1990; Chalmers et al. 1989). Broad entry criteria were used during this primary phase of data collection for the reviews. Trials were eligible for inclusion in the register if some attempt had apparently been made to create unbiased comparison groups, either by randomisation, or by using a method such as alternation in a consecutive series, case record number, or date of birth.

The methods used in preparing the reviews have been described in detail elsewhere (Chalmers et al. 1989). Reviewers were required to assess the quality of possibly relevant trials using a hierarchy of three dimensions relevant to the control of bias -- concealment of treatment allocation schedule, avoidance of exclusions after trial entry, and blinding of outcome assessment. Because the trials incorporated in most of these reviews varied in quality, the result of each trial was plotted graphically in a hierarchy reflecting its quality assessment. In that way, both those preparing the reviews and those reading them could be alerted to patterns of results which might reflect the influence of bias (Chalmers et al. 1989). The present analysis explores that variation in methodological quality among component trials in systematic reviews.

The database contained a total of more than 500 systematic reviews (Enkin et al. 1993), but about 60% comprised just one or two trials. An appropriate, manageable sample of reviews for this analysis was derived as follows. First, an initial subset of 82 reviews was selected on the basis of the following criteria: each had used meta-analysis to produce a quantitative synthesis of at least 5 trials, and each had a total of at least 25 outcome events among the trial control groups contributing to the review. Second, to ensure that data from a particular component trial would contribute only once to the main analysis, all meta-analyses to which each component trial had contributed were identified, and only the meta-analysis with the most homogeneous grouping of interventions was retained for inclusion in the analysis. For example, a meta-analysis incorporating a specific class of antibiotics given for prophylaxis with caesarean delivery would be included in preference to a meta-analysis which had included trials of **any** antibiotic used in that way. With only minor levels of overlap between two meta-analyses, an overlapping trial was deleted from one of them. That happened with only 6 trials; a random number table determined the deletions. And third, the meta-analyses included in the analysis had to comprise at least one component trial with adequate concealment (see below) of the treatment allocation schedule and at least one trial without.

Of the meta-analyses not used in this study, most exclusions were because of substantial overlap with another meta-analysis, or because no trials had reported adequate allocation concealment. Two meta-analyses were excluded, however, because all of the component trials had used adequate approaches to allocation concealment.

After that selection process had been applied, 33 meta-analyses remained for analysis. Appendix 4.1 provides a brief description of each. They related to care during pregnancy (2), preterm labour and delivery (4), induction of labour (7), labour and delivery (6), prophylactic antibiotics for caesarean delivery (7), the puerperium (3), and the neonatal period (4). Unpublished trials and trials reported in languages other than English were dropped from this analysis because of difficulty in evaluating the necessary information. These 33 meta-analyses included data from 250 primary trials involving a total of 62,091 participants with 12,030 outcome events measured. The trials included were published between 1955 and 1992. Of the 250 trials, 10% were published from 1955-1969, 18% from 1970-1979, 61% from 1980-1989, and 11%

from 1990-1992. The overall effect in all of the meta-analyses suggested that the experimental interventions were more effective than the control interventions in preventing adverse outcomes, thus yielding summary odds ratios of less than 1.0.

4.3.2 Assessment of trial quality

Unfamiliar with the topics, journals, and investigators covered, I evaluated each of the 250 published trial reports contributing to the 33 selected meta-analyses. The trial quality was assessed using the same dimensions that had been used by the reviewers in the Cochrane Pregnancy and Childbirth Group (Chalmers et al. 1989), apart from also assessing the adequacy of sequence generation. For the current analysis, however, assessments entailed special efforts to ensure consistency of quality assessments by using a single observer with a more detailed classification scheme and data abstraction instrument, and by blinding the assessment process to the results of the trials. Blinding was achieved in assessing the descriptions of randomisation and blinding in the articles, because authors embedded those descriptions in the methods sections. Blinding assessments of exclusions after trial entry, however, posed greater difficulties, because authors frequently addressed those descriptions in the results sections. Trials were assessed on the following four major dimensions:

A) **Concealment of treatment allocation schedule** Trials were divided into three groups.

- **Adequately concealed** trials, which were the referent group, deemed to have taken adequate approaches to allocation concealment (central randomisation; numbered or coded bottles or containers; drugs prepared by the pharmacy; serially numbered, opaque, sealed envelopes; or other description that contained elements convincing of concealment).

● **Inadequately concealed** trials in which concealment was clearly inadequate (such as alternation or reference to case record numbers or to dates of birth).

● **Unclearly concealed** trials in which the authors either did not report an allocation concealment approach at all or reported an approach that did not fall into one of the above categories. A majority of the trials in this group probably had randomised allocation schedules, but had inadequate concealment. On the other hand, a few probably had randomised allocation schedules with adequate concealment and a few others probably had used systematic, unconcealed allocation.

B) **Inclusion of all randomised participants** A referent group of trials that reported, or gave the impression, that **no exclusions** had taken place (vast majority not explicit), and a second group of trials that reported having made **exclusions**. The reasons for exclusions, when reported, included protocol deviations, withdrawals, drop-outs, and losses to followup.

C) **Double-blinding** A referent group of trials that reported being **double-blinded**, and a second group that did not report as such, deemed **not double-blinded**. Only meagre information was made available on the approaches used for double-blinding, so the classification was based necessarily on whether the reports purported to be of double-blind trials.

D) Generation of allocation sequences A referent group of **adequate sequence generation** trials which reported adequate approaches (random number tables, computer random number generators, coin tossing, or shuffling), and a second group which did not report one of the adequate approaches, those with **inadequate sequence generation**. This dimension was analyzed only in the subset of 79 trials that had reported adequate allocation concealment. That analytical approach makes sense because having a randomised (unpredictable) sequence would make little difference without adequate concealment. Yet in trials with adequate concealment, particularly those where group assignment becomes known after allocation (such as in unblinded trials), a randomised sequence may become important.

The data were entered interactively into an EPI-INFO questionnaire with on-line editing, skip-pattern, and logic-checking capability (Dean et al. 1990). Basic tabulations were done using EPI-INFO. To examine the reproducibility of items on the questionnaire, Douglas G. Altman (Head, Medical Statistics Laboratory, Imperial Cancer Research Fund) reassessed a sample (selected with the EPI-INFO computer random number generator) of 10 trials blinded to the initial assessments. Information on sequence generation, allocation concealment, blinding, and numbers of participants randomised and analyzed revealed no notable inconsistencies with the initial assessments. Information on whether a trial used 'intent-to-treat' principles, however, revealed some inconsistencies, largely because the reports forced that assessment to be made subjectively by not providing explicit information. Thus, assessments of intent-to-treat were not analyzed.

A recapitulation of two concepts addressed in section 2.3 may be helpful. Allocation concealment can always be implemented, seeks to prevent selection bias, and protects the assignment sequence **before and until** allocation. On the other hand, double-blinding cannot always be implemented, seeks to prevent ascertainment bias, and protects the assignment sequence **after** allocation.

4.3.3 Statistical methods

The intention was to examine, across all 33 meta-analyses, associations between the indicators of trial quality described above and estimates of treatment effects (odds ratios). Multiple logistic regression models were used, including a binary variable for treatment group (experimental or control), indicator variables to control for the effects of each of the 250 trials, and terms for the 'meta-analysis-by-treatment group' interaction (to control for the different overall odds ratios for the treatment effects in the 33 meta-analyses). While all these terms needed to be accounted for in the analysis, estimates of the parameters did not need to be provided. The indicator variables for the different meta-analyses were not included in the models because they were completely confounded with the variables representing the trials.

The effects of all the methodological quality measures were measured at the trial level. Their main effects were, however, completely confounded with variables representing the trials, which were already in the basic model. The terms for 'quality measure-by-treatment' interaction addressed the main question of interest, namely '**On average, do trials judged to have been methodologically inferior yield different odds ratios from trials assigned to the relevant best (referent) category?**' Since all of the treatment effects in these trials were coded in the same direction, a relative odds of less than 1.0 for one of these interaction terms would mean that trials that were methodologically inferior had yielded larger estimates of treatment effects, on average, compared to the referent group. Conversely, a relative odds of larger than 1.0 would indicate association with smaller treatment effects or perhaps even treatment risks.

After analysing inadequately concealed trials in the initial model, they were excluded from further analyses of associations with treatment effects. The further analyses included adding additional terms to the models and the data would have become quite thin with only 21 inadequately concealed trials. Moreover, including them in further analyses makes little theoretical sense. For example, analysing the impact of double-blinding in inadequately concealed trials would be unjustified since double-blinding would likely be impossible under such circumstances. Furthermore, given the initial results, the determination ensued that the most important analyses thereafter should focus on the unclearly and adequately concealed trials.

To analyze the variability (heterogeneity) of the three allocation concealment groups, a separate model was fitted to each of those independent groups of trials. Each model included a binary variable for treatment, indicator variables for the individual trials, and interaction terms for 'meta-analysis-by-treatment.' The deviances derived from those models are approximately chi-squared distributed. Then the remaining heterogeneity in the unclearly and inadequately concealed groups were separately compared to the adequately concealed group, using F-ratio tests.

GLIM 4 (Francis et al. 1993) was used for modelling, using the 'eliminate' command for the trial parameters. The deviances and degrees of freedom (df) for the models appear in Tables 4.2 and 4.3. The changes in deviance associated with adding term(s) to models have an approximately chi-square distribution, with degrees of freedom equal to the difference in degrees of freedom between the two models.

A model yielding a greater deviance than its degrees of freedom indicates heterogeneity (also referred to as 'overdispersion' or 'extra-binomial variation'). For one model (B3/C4) in the analysis, a simple adjustment to the scale parameter took rough account of overdispersion for estimating standard errors (Aitkin et al. 1989). That adjustment yielded a model with a deviance equal to its degrees of freedom.

Separate multiple logistic regression analyses were also performed on the trials in each of the 33 meta-analyses. The basic model for these included parameters for trials and treatment. The primary analytical variables of interest were the terms describing the interactions between allocation concealment and treatment.

4.4 Results

4.4.1 Characteristics of the 250 trials

Steps taken to conceal treatment allocation schedules were adequate in 79 trials, unclear in 150, and inadequate in 21. Overall, 63% did not report any exclusions, but the reported quality of the allocation concealment was inversely related to the proportion without exclusions (Table 4.1). The trials that reported using adequately concealed allocation were the most likely to have reported excluding participants whereas the trials which reported using inadequately concealed allocation were the least likely. The unclearly concealed trials reported exclusions at an intermediate level between those two. Double-blinding and having used adequate sequence generation, however, were much more common in the adequately concealed trials than in the other groups (Table 4.1).

4.4.2 Concealment of treatment allocation schedules and estimates of treatment effects

Trials with inadequate or unclear concealment measures yielded more pronounced estimates of treatment effects than trials that had taken adequate measures to conceal allocation schedules ($p<0.001$; Model A2; Table 4.2). Excluding from analysis inadequately concealed trials, the association for unclearly concealed trials was statistically significant ($p<0.001$; chi-squared=48.6, 1 d.f.; Model B2; Table 4.2) and remained so after accounting for the effects of exclusions and double-blinding (Model B3; Table 4.2). The relative odds was 0.70, which means that the odds ratios in the unclearly concealed trials were, on average, 30% lower than in the adequately concealed trials, i.e., estimating larger treatment effects.

Table 4.1

No exclusions, double-blinding, and allocation schedule generation by the level of allocation concealment for 250 controlled trials

Authors reported:	Adequately concealed	Unclearly concealed	Inadequately concealed	All trials
No apparent exclusions	53%	67%	76%	63%
Double-blinding	73%	39%	14%	48%
Adequate generation of allocation schedule	29%	15%	0%	18%
Total [N]	100% [79]	100% [150]	100% [21]	100% [250]

Table 4.2

Association between judgement of concealment of treatment allocation and estimates of treatment effects

Trials included	Model[1]	Level of concealment of treatment allocation [interaction terms][2]	Relative odds [95% CI]	Deviance [d.f.]
All trials [n=250]	A1	Base model	-	492.09 [217]
	A2	Adequate	1.00	434.18 [215]
		Unclear	0.67 [0.60 - 0.75]	
		Inadequate	0.59 [0.48 - 0.73]	
Adequately and unclearly concealed trials [n=229]	B1	Base model	-	381.69 [196]
	B2	Adequate	1.00	333.13 [195]
		Unclear	0.67 [0.60 - 0.75]	
	B3	Adequate	1.00	325.61 [193]
		Unclear [adjusted][3]	0.70 [0.62 - 0.79]	

[1] Multiple logistic regression models include a binary variable for treatment, indicator variables for the trials, and meta-analysis by treatment interaction terms

[2] All terms involve an interaction with treatment effect

[3] Model includes the exclusion and not-double-blinding by treatment interaction terms which adjusts for their effects in this analysis

The results of the 33 separate logistic regression analyses were consistent with the aggregated modelling results. The direction of the effects for the unclearly concealed trials were toward larger estimates of treatment effects in 27 and smaller in 6. While supportive of the overall trend, this might indicate that the effect of concealment may not have been the same in every meta-analysis. Indeed, an interaction term for 'unclearly concealed-by-treatment-by-meta-analysis' was statistically significant (p=0.01; chi-squared=53.7; 32 d.f.), meaning that the effect of unclearly concealed trials on estimation of treatment effects varied by more than chance among the meta-analyses.

4.4.3 Exclusions after allocation and estimates of treatment effects

Accounting for allocation concealment, trials that excluded participants yielded estimates of treatment effects that, on average, were smaller than those derived from trials that had not excluded any participants. However, this association was not statistically significant (p=0.21, chi-squared=1.54, 1 d.f.; Model C2; Table 4.3), and taking double-blinding into account hardly affected its strength (Model C4; Table 4.3).

4.4.4 Double-blinding and estimates of treatment effects

Accounting for allocation concealment, trials that were not double-blinded yielded estimates of treatment effects that, on average, were larger than those derived from trials that were double-blinded (p=0.01; chi-squared=6.50; 1 d.f.; Model C3; Table 4.3). Controlling for exclusions after allocation did not influence the strength of this association (Model C4).

Table 4.3

Association between assessments of methodological elements to control bias after randomisation and estimates of treatment effects

Trials included	Model[1]	Methodological element [interaction terms][2]	Relative odds [95% CI]	Deviance [d.f.]
Adequately and unclearly concealed trials [n=229]	C1	Base	-	333.13 [195]
	C2	No exclusions	1.00	331.59 [194]
		Exclusions	1.08 [0.96 - 1.23]	
	C3	Double-blinded	1.00	326.63 [194]
		Not double-blinded	0.82 [0.71 - 0.96]	
	C4	No exclusions	1.00	325.61 [193]
		Exclusions	1.07 [0.94 - 1.21]	
		Double-blinded	1.00	
		Not double-blinded	0.83 [0.72 - 0.96]	

[1] All models include a binary variable for treatment, indicator variables for the trials, meta-analysis by treatment interaction terms, and an allocation concealment by treatment interaction term.

[2] All terms involve an interaction with the treatment effect.

4.4.5 Generation of allocation sequence and estimates of treatment effects

In those 79 trials that had used adequately concealed allocation, trials with inadequate sequence generation yielded larger estimates of treatment benefits, on average, than trials with adequate sequence generation (relative odds of 0.75; 95% CI of 0.55 - 1.02; p=0.07). That association was of marginal statistical significance, however.

4.4.6 Allowance for overdispersion

Because statistically significant heterogeneity remained in the model represented in B3 or C4 (Table 4.2, Table 4.3), a separate analysis allowed for overdispersion. The point estimates remained the same, but the 95% confidence intervals for the terms widened slightly: unclearly concealed (0.60-0.82); exclusions (0.91-1.25); and not double-blinded (0.68-1.01).

4.4.7 Heterogeneity

Odds ratios derived from trials that had used inadequately concealed allocation yielded statistically significantly more variability (heterogeneity) than those from adequately concealed trials (Table 4.4). Although the unclearly concealed trials also yielded greater heterogeneity compared with the adequately concealed trials, it was not statistically significantly greater.

Table 4.4

Examination of the variability (heterogeneity) of the unclearly and the inadequately concealed trials relative to the adequately concealed trials

Concealment Status [N]	Model[1]	Deviance[2] [d.f.]	F-ratio for comparison to the adequately concealed trials $[df_1, df_2]$	P-value
Adequate [79]	D	66.32 [46]	–	–
Unclear [150]	E	213.10 [117]	F=1.26 [117,46]	0.19
Inadequate [21]	F	43.07 [7]	F=4.27 [7,46]	0.001

[1] Linear logistic regression models include a binary variable for treatment, indicator variables for the trials, and meta-analysis by treatment interaction terms: separate models for each set of trials

[2] Indicating the residual heterogeneity

4.5 Discussion

4.5.1 General findings

Comparisons of different forms of health interventions can be misleading unless investigators take precautions to ensure that their study contains unbiased comparison groups with respect to prognosis. Random allocation to alternative forms of interventions remains the only way of controlling for selection biases. That characterizes randomisation's only unique strength. That strength, however, becomes of crucial importance in the common circumstances where the treatment effects may be of comparable magnitude to the biases that plague most non-randomised comparisons of alternative forms of health care.

In view of the central importance of randomisation for achieving unbiased comparisons, it is surprising that authors so often inadequately detail the steps taken to assign participants to comparison groups in trials (Mosteller 1980; DerSimonian et al. 1982; Altman and Doré 1990), including authors in the medical specialty represented in this analysis (Chapter 3). If randomisation prevents bias as suggested, trials that have failed to report adequate approaches to ensuring proper randomisation should yield systematically different estimates of treatment effects to those derived from trials that have apparently used adequate approaches. That proposition has been supported by the analyses in this chapter. Specifically, these analyses have confirmed the prior hypothesis that, on average, trials that had not reported adequate measures for allocation concealment yielded estimates of treatment effects that were larger than those from trials that had used adequate approaches to concealment. In addition, these analyses have indicated that estimates of treatment effects were somewhat larger in trials that had not reported double-blinding.

4.5.2 Concealment of treatment allocation and estimates of treatment effects

Trials that reported inadequate concealment methods yielded estimates of odds ratios that were distorted by an average of 41% and trials that reported unclear concealment were distorted by an average of 30% (when adjusted), compared with estimates of odds ratios derived from trials which had apparently taken adequate steps to conceal treatment allocation. Those estimated distortions in odds ratios should not be interpreted as meaning that the distortions in relative risks would be of a similar magnitude. The proportion of outcome events in the control groups of the trials studied was almost 20% overall, and ranged from less than 1% to over 90%. The odds ratio poorly estimates relative risk with proportions over 5%-10%, and gives more divergent estimates as proportions move towards 90%. Thus, a distortion of 30% in odds ratios may translate, for example, to a distortion of only 10%-20% in relative risks. Nevertheless, biases of that magnitude seem considerable and clearly unacceptable.

The effect of unclearly concealed trials on estimation of treatment effects varied by more than chance among the meta-analyses. Therefore, while a distortion of 30% in odds ratios for the unclearly concealed trials appropriately estimates the average association, it should not be interpreted as representing all the meta-analyses. In addition, that odds ratios yielded by inadequately concealed trials were more variable (heterogeneous) than those derived from trials with adequate concealment further illustrates that indeed inadequately concealed trials may unreliably protect against bias.

The associations found between the adequacy of allocation concealment and estimates of treatment effects probably reflect selection biases. Admittedly, poorly executed trials cannot be confirmed as further from the truth than well executed trials since 'truth' eludes recognition. Nevertheless, these results support the primary hypothesis and strongly suggest the presence of bias.

Inadequate concealment of allocation schedules can lead to introduction of bias in many ways, sometimes as the result of deliberate subversions, sometimes as the net effect of subconscious actions. For example, if those responsible for admitting participants to trials have foreknowledge of treatment allocations, they may channel participants with a better prognosis to the experimental group and those with a poorer prognosis to the control group, or vice versa. That could easily be accomplished either by delaying a participant's entry into the trial until the next desired allocation appears, or by excluding eligible participants from the trial, or by encouraging them to refuse entry. Without allocation concealment, biases in both directions become possible, although the clear tendency identified in this study pointed towards larger treatment effects.

Inadequate allocation concealment may also be a surrogate measure for the quality of other aspects of trial design and execution, such as blinding and handling of exclusions. The magnitude of the associations observed may thus also reflect biases other than selection biases at the point of treatment allocation.

In any case, however, the results from this chapter support the policy decision taken by one journal (Altman 1991b) not to publish reports of trials in which foreknowledge of treatment allocation is obviously possible, such as trials which have used alternation, case record numbers, or open lists of random numbers for allocation. These findings also emphasise the importance both of securing adequate allocation concealment in controlled trials, and of ensuring that authors and journal editors publish reports that make those aspects of trial design explicit.

4.5.3 Exclusions after allocation and estimates of treatment effects

Contrary to the prior hypothesis, trials that reported excluding participants after randomisation did not yield inflated estimates of treatment effects compared with trials in which there had apparently been no exclusions. Indeed, if anything, trials with reported exclusions yielded more modest estimates of treatment effects than other trials. That unexpected finding could be the result of some authors inappropriately reporting that they had randomised the same number of participants as they had analyzed, even though some randomised participants had actually been excluded.

The observed results on exclusions probably reflect incomplete reporting. That interpretation receives support from the paradoxical finding that trials using adequately concealed allocation were the most likely to have reported excluding participants and that trials using inadequately concealed allocation were the least likely (Table 4.1). Few of the trials reporting no apparent exclusions explicitly stated that no exclusions had taken place.

Gøtzsche (1989b) has previously detected non-reporting of exclusions in two separate published trials; he speculated that articles in which the issue of exclusions after randomisation has not been addressed explicitly may be less reliable than those in which it has been. We found that authors often omit explicit information about exclusions, and we failed to detect any association between our classification of trials using the information available on exclusions and any differential effects of treatment. That suggests that information on exclusions may currently have little value in assessing trial quality from published information.

Investigators must be urged to report accurately the number of exclusions and, if none took place, to state so explicitly. They must also explicitly state their approach to the analysis, primarily whether they performed an intent-to-treat analysis. The ultimate users of RCTs, including in particular those performing systematic reviews, should be aware of the potentially misleading information provided on exclusions in currently published trials. Minimising exclusions and implementing an intent-to-treat analysis are likely to reduce bias, but available articles appear to report too incompletely to support a reliable examination of our hypothesis on this issue.

4.5.4 Double-blinding and estimates of treatment effects

Trials for which no double-blinding was reported yielded estimates of odds ratios that were inflated by 17%, on average, compared with trials that reported having used double-blinding. Trials that reported double-blinding usually provided little, if any, information on the methods used. Consequently, some trials claiming to be double-blind may not have been, and that measurement error could have been reflected in an underestimate of the effect of not double-blinding. On the other hand, much of the effort to double-blind in these trials may have gone into concealing randomisation. Thus that could have led to a countervailing overestimate of the independent effect of not double-blinding. Whatever the net effects, blinding is clearly of far greater importance to minimizing bias for some outcomes than for others.

4.5.5 Generation of allocation sequence and estimates of treatment effects

In the 79 trials that had used adequate allocation concealment, trials that did not report an adequate sequence generation approach, presumably making prediction of the next allocation easy, found larger treatment effects. The confidence interval for the relative odds, however, just included 1.0. That finding suggests that an unpredictable sequence should help to protect against bias if adequate allocation concealment is used. One could further speculate that the benefit gained from unpredictability would be particularly important when blinding is not instituted. The preparation of an unpredictable sequence could be quite useless if allocation was inadequately concealed. Therefore, given adequate allocation concealment, this result implies the importance of randomly generated, unpredictable allocation sequences.

4.5.6 Overdispersion

Even using the wider confidence intervals based on the adjustment for overdispersion, the interpretation of the results remained essentially unchanged. Ideally, a random effects model, in which the treatment effect is assumed to vary randomly among trials, would be most appropriate to address overdispersion. However, difficulty in identifying a software package that handles both random effects and the requisite large number of parameters in these models led to the use of an approximate adjustment (Aitken et al. 1989), in which the variance of each binomial response is assumed to be inflated by a constant factor (the "scale parameter") chosen so that the deviance is equated to the degrees of freedom. It is unclear how accurate this approximate adjustment is likely to be in the current application, since a random effects model for the treatment effect, and the variable sample sizes of the trials, would imply a more complex variance structure. Nevertheless, the adjustment suggests that the results for allocation concealment would likely remain highly statistically significant regardless of the approach used, while some caution is advisable in interpreting the results for double-blinding.

4.5.7 Test for comparative heterogeneity

Odds ratios yielded by inadequately concealed trials were more heterogeneous than those derived from trials with adequate concealment. To compare the heterogeneity between those two groups, the specified F-tests were used. These tests depend on the validity of the simple overdispersion model referred to in Section 4.5.6, and so should be regarded as approximate. Furthermore, theoretical considerations suggest that the apparent excess variability among inadequately concealed trials might be explained away if these trials had larger sample sizes on average. In fact, however, average sample size was smaller in the inadequately concealed trials than in the adequately concealed trials (93 compared with 203). That suggests that the greater mean deviance among inadequately concealed trials genuinely reflects greater heterogeneity of treatment effects among those trials.

4.5.8 Consideration of results in the light of previous research

These results are consistent with the findings of Chalmers et al. (1983). Whereas their results may have been attributable to confounding by type of treatment, these results cannot be explained in that way because the analysis accounted for the effects of the 33 meta-analyses, each of which investigated similar treatments and similar outcome measures.

Emerson et al. (1990), however, did not find a statistically significant association between methodological quality and treatment effects. Their findings do not necessarily conflict with the findings in this chapter for at least four reasons. First, and most important, the authors used a measure of quality (Chalmers et al. 1981) that quantifies many aspects of trial design and analysis, some of which were unlikely to be related to bias. Second, their inclusion of both study size and their measure of quality as explanatory variables in their model could have obscured the effects of quality relating to bias because study size correlated with quality score. Third, their statistical model may have been less sensitive since a logistic regression package able to handle the requisite number of parameters probably has only recently become available. Fourth, the trials used in their analysis may have been less heterogeneous in terms of quality than those included in this analysis, thus reducing the likelihood that they would have detected any associations that existed.

4.5.9 Conclusions

Some methods used to control biases in trials cannot always be implemented. For example, investigators cannot always blind participants, care-providers, and assessors or include every randomised participant in all of the primary analyses of all the outcomes measured. By contrast, investigators should *always* be able to conceal treatment allocation schedules so that proper random allocation can be assured. That not only addresses selection bias before treatment begins, but also serves as a prerequisite for the success of other measures (such as blinding and 'intention-to-treat' analyses) taken to control assessment and exclusion biases.

Preventing foreknowledge of treatment allocation by effective concealment of allocation schedules emerges from the analyses as crucially important to protecting against bias. Without proper application of measures to achieve concealment, the whole point of randomisation vanishes. The results in this chapter support the comment of Mosteller and his colleagues (1980): "When the randomization leaks, the trial's guarantee of lack of bias runs down the drain".

Appendix 4.1

Description of the interventions and outcome measures in the 33 meta-analyses by general topical area

General topical area	Intervention in the meta-analysis	Outcome measure
Pregnancy	Routine iron administration in pregnancy	Hb < 10-10.5 g/dl at 36-40 wks.
	Antiplatelet agents for IUGR and pre-eclampsia	Baby weight <5th <10th centile
Preterm labour and delivery	Betamimetic tocolytics in preterm labour	Delivery within 48 hours
	Antibiotics for preterm prelabour rupture of membranes	Delivery <1 week after trial entry
	Corticosteroids after preterm prelabour rupture of membranes	RDS (respiratory distress syndrome)
	Prophylactic oral betamimetics in pregnancy	LBW (<2500 g)
Induction of labour	Any prostaglandin vs placebo for induction of labour	Operative delivery
	Vaginal PGE_2 for cervical ripening	Caesarean section
	Prostaglandins vs. mechanical methods for cervical ripening	Caesarean section
	Elective induction of labour at 41+ weeks gestation	Caesarean section
	Any PG vs. oxytocin for induction of labour	Operative delivery
	Endocervical PG for cervical ripening	Caesarean section
	Oestrogen pretreatment before labour induction	Caesarean section
Labour and delivery	External cephalic version at term	Caesarean section
	Vacuum extraction vs. forceps	Significant maternal injury
	Electronic fetal monitoring plus scalp sampling vs. intermittent auscultation in labour	Neonatal seizures
	Birth chair during second stage of labour	Instrumental vaginal delivery
	Prophylactic oxytocics in third stage of labour	Postpartum haemorrhage
	Umbilical vein oxytocin for retained placenta	Manual removal of placenta

General topical area	Intervention in the meta-analysis	Outcome measure
Prophylactic antibiotics for caesarean section	Cephalosporins vs. placebo for Caesarean section	Febrile morbidity /endometritis
	Broad spectrum penicillin vs. placebo for Caesarean section	Febrile morbidity /endometritis
	Metronidazole vs. placebo for Caesarean section	Febrile morbidity /endometritis
	Broad spectrum penicillin + aminoglycoside vs. placebo for Caesarean section	Febrile morbidity /endometritis
	Antibiotic irrigation vs. placebo at Caesarean section	Febrile morbidity /endometritis
	Antibiotic peritoneal irrigation vs. systemic antibiotics for Caesarean section	Febrile morbidity /endometritis
	Broad spectrum penicillin vs. cephalosporins for Caesarean section	Febrile morbidity /endometritis
Puerperium	Oral proteolytic enzymes for perineal trauma	Oedema on the third day
	Polyglycolic acid vs. catgut for perineal repair	Short-term pain
	Stilboestrol for lactation suppression	Continuing lactation 1-week postpartum
Neonatal	Prophylactic phenobarbital in very low birthweight neonates	Any peri/intraventricular haemorrhage
	Natural surfactant extract treatment of RDS	Broncopulmonary dysplasia (BPD) or death
	Prophylactic administration of synthetic surfactant	Neonatal mortality
	Prophylactic vitamin E in preterm infants for BPD	BPD

Chapter 5

5. Blinding and Exclusions After Allocation in Randomised Controlled Trials in Obstetrics and Gynaecology

5.1 Summary

5.1.1 Objective. The objective was to assess the methodological quality of approaches to blinding and to handling of exclusions as reported in controlled trials from obstetrics and gynaecology journals.

5.1.2 Methods. The analysis was based on an evaluation of a random sample of 110 reports of the parallel group trials identified in Chapter 3, in which allocation was stated to have been randomised, published in the 1990 and 1991 volumes of four journals of obstetrics and gynaecology.

5.1.3 Results. Of the 31 trials that reported being double-blind, a mere 45% provided minimally sufficient descriptions of double-blinding, and only 26% provided additional information on the protection of the allocation schedule throughout the trial. Furthermore, only 16% of the reports provided any written assurances of successfully implementing blinding. Investigators tested the efficacy of blinding in only two trials. In the 49 trials in which the authors provided sufficient information for readers to imply that no exclusions after randomisation had taken place, only 22% explicitly stated that no exclusions had taken place and only one (2%) stated that an intent-to-treat analysis had been performed. Moreover, those trials generally provided methodological information of inferior quality to that provided from trials that explicitly reported exclusions after randomisation.

5.1.4 Conclusions. Authors poorly reported methods used to achieve double-blinding. The poor reporting of post-randomisation exclusions, however, causes greater concern. Some of the trials reporting no apparent exclusions may not only have had exclusions, but may have improperly handled exclusions. Conversely, those reporting exclusions may have represented many of the better trials. That situation breeds an insidious problem, because most view trials without exclusions as less biased than those with exclusions. All those who rely upon the results of randomised controlled trials should be aware of this paradox.

5.2 Introduction

Given the importance of randomisation, many consider blinding and handling of exclusions as the next two most important methodological components of controlled trials (Chalmers et al. 1989). While failure to double-blind (double-mask) trials has been found to be associated with biased results, failure to include all randomised participants has not (Chapter 4). In that analysis, however, inadequate information in published trials hampered measurement of the associations.

Double-blinding can be a critical element for reducing bias in trial design for at least a few reasons (Pocock 1983). First, participants who know they have received a new drug might be expected, in many instances, to respond better than untreated controls, even with an ineffective new drug. Second, if those treating a participant know the assignments, they would likely provide different ancillary care that may affect eventual response. Third, if evaluators know the assignments, they may alter their responses in the direction of their biases or overcompensate in the other direction. All too frequently, "evaluators will err towards recording more favourable responses on the new treatment: after all, most trials are conducted in the hope that a new treatment will appear superior and it is only human nature to anticipate such superiority" (Pocock 1983).

Even though investigators might readily acknowledge the importance of double-blinding, they frequently report their study only as 'double-blind' and do not provide any further clarifying information (Mosteller et al. 1980). Trials called double-blind are not always so. In 196 trials of nonsteroidal anti-inflammatory drugs for rheumatoid arthritis that were reported to be double-blind, at least 8% were probably not actually double-blind (Gøtzsche 1989a). Furthermore, some trials become unblinded (Karlowski et al. 1975; Huskisson and Scott 1976). Authors infrequently include information on the method of blinding (Gøtzsche 1989a; Mosteller et al. 1980) and rarely test its adequacy (Gøtzsche 1989a).

Exclusions after randomisation can, for example, be because of eventual discovery of participant ineligibility, protocol deviations, withdrawals, or losses to followup. The inappropriate handling of those exclusions may greatly influence the results of a trial (Chalmers et al. 1981; Pocock 1982; de Jonge 1983; May et al. 1981; Sackett and Gent 1979). The preferred, unbiased approach is to include all randomised participants in the analysis, regardless of compliance with protocol. Methodologists refer to that approach as analysis by "**intention to treat**" (Pocock 1983; Bulpitt 1983).

Trial reports, however, contain insufficient information about exclusions after randomisation (Meinert et al. 1984; DerSimonian et al. 1982; Emerson et al. 1984). In one analysis, at least 42% of the trials provided insufficient information on exclusions after randomisation (Gøtzsche 1989a). Many other trials may give the illusion of adequate information, but the impression provided may be inaccurate (Chapter 4). Indeed, instances have been found in which participants excluded after randomisation were excluded from the entire publication (Gøtzsche 1989b). "Not to mention the existence of patients who withdrew from therapy or otherwise deviated from protocol is a serious failing which can lead to exaggerated claims about treatment efficacy" (Pocock 1983).

The above-mentioned study examining methodological quality and bias (Chapter 4) analyzed trials in pregnancy and childbirth published over the last 40 years. To estimate the current state of reporting on blinding and exclusions in a medical specialty representative of pregnancy and childbirth, I conducted a systematic evaluation of reports of RCTs published during 1990-91 in four obstetrics and gynaecology journals. This probably represents the first systematic analysis of blinding and exclusions undertaken in this medical specialty. The systematic analysis of randomisation and allocation described in Chapter 3 indicated that methodological quality may indeed be worse than that of reports published in general medical journals.

The primary general supposition was that relatively few trials would provide adequate specific information on blinding and exclusions. However, the reports published in the British Journal of Obstetrics and Gynaecology were thought to be potentially of better quality than those published in other journals of obstetrics and gynaecology since a concerted editorial effort had been made to improve the quality of reporting in that journal, including publication of a series of articles providing reporting standards for different types of studies (Bracken 1989; Wald and Cuckle 1989) including trials (Grant 1989). Moreover, its reports contained higher quality descriptions of randomisation (Chapter 3).

5.3 Methods

5.3.1 Study material

Reports were evaluated from the same four journals of obstetrics and gynaecology described in Chapter 3: the **American Journal of Obstetrics and Gynecology** (AJOG), the **British Journal of Obstetrics and Gynaecology** (BJOG), **Obstetrics and Gynecology** (OG), the **Journal of Obstetrics and Gynaecology** (JOG). All 206 reports of randomised trials published in the 1990 and 1991 volumes of those journals were identified for that research. From those, a sample of 110 reports were taken for evaluation by selecting all 20 from the JOG and by taking a random sample (computer random number generator in EPI-INFO) of 30 from each of the other 3 journals. Responses on allocation concealment from the work in Chapter 3 were linked to the records from this study.

A pilot study using articles from 1989 in these same journals aided in the development and testing of the data collection instrument for this study. For consistency of measurement across journals, I did all of the initial assessments. To examine the reproducibility of items on the questionnaire, David A. Grimes, M.D., a Professor of Obstetrics and Gynaecology at the University of California, San Francisco, independently assessed a random sample of 10 trials, blinded to the initial assessments. We did not find any notable differences for specific information on blinding, but did find notable inconsistencies on the handling of exclusions. In particular, our assessments of whether a trial used 'intent-to-treat' principles sometimes revealed inconsistencies, largely because the reports forced this assessment to be made subjectively by not providing explicit information on exclusions and intent-to-treat. Thus, variables that reflected our judgements of 'intent-to-treat' were not analyzed and reported on. We did, however, analyze and report on a variable reflecting an author's explicit statement of adhering to intent-to-treat principles.

The data were entered interactively into an EPI-INFO questionnaire with capabilities for on-line checking and editing (Dean et al. 1990). Chi-squared tests were used for comparing nominally scaled variables.

5.3.2 Terminology

To reiterate concepts presented earlier, allocation concealment can always be implemented, seeks to eliminate or minimize selection bias, and protects the assignment sequence **before and until** allocation (Section 2.3). On the other hand, blinding cannot always be implemented, seeks to eliminate or minimize ascertainment bias, and protects the assignment sequence **after** allocation (Section 2.3).

Comparisons of treatments may be distorted if participants, care-givers, or evaluators know the treatment assignments (Pocock 1983; Altman 1991a). All may respond and observe differently based upon their knowledge of treatment assignments. A **double-blind** trial shields all those individuals from that knowledge for the purpose of avoiding bias. Double-blinding becomes particularly important with subjective outcomes.

In some difficult situations, double-blinding necessitates the use of a technique called **double-dummy**. Briefly, the technique involves the administration to each participant of one of the active treatments and the placebo (dummy) of the alternative active treatment (Altman 1991a).

If either participants, care-givers, or evaluators know the assignments, then the trial is not double-blind. If the individuals in at least one of those categories are blinded to the assignments, then the trial would usually be called **single-blind**. Generally, evaluators are thought to be the most important to blind of those involved in trials.

If possible, blinding is highly desirable. However, in some fields, blinding in general, double-blinding in particular, becomes difficult or impossible. Surgery exemplifies one such field.

5.4 Results

5.4.1 Blinding

Of the trials in this study, 65% involved pharmaceutical interventions. Overall, more than one-quarter of the trials reported being double-blinded, with the JOG reporting the lowest proportion and the other three journals reporting at about the same level (Table 5.1). Double-blinding was not feasible in all of the trials, however. It was judged to be feasible in 59% of the trials overall: 70% of the trials from AJOG, 53% from BJOG, 50% from JOG, and 60% from OG. Thus, of trials that could have been double-blinded, the authors purported only 48% to be double-blind: 43% from AJOG, 63% from BJOG, 20% from JOG, and 56% from OG (p=0.16; 3 d.f.). The BJOG articles reported blinding of the assessor slightly more frequently than did articles from the other journals (Table 5.1).

Of the trials that reported being double-blind, all except one provided information on the approach used (Table 5.2). However, overall only 45% of the reports described similarity of the treatment and control regimens in appearance, taste, administration, or other minimal prerequisites for successful double-blinding: 56% in AJOG, 60% in BJOG, 50% in JOG, and 20% in OG.

Table 5.1

Type of blinding reported by journal

Type of blinding reported by authors	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
Referred to as double-blind; participants, care-givers, and assessors blinded	30% (9)	33% (10)	10% (2)	33% (10)	28% (31)
Those assessing outcomes were blinded	7% (2)	23% (7)	15% (3)	10% (3)	14% (15)
Participants or care-givers blinded, but not assessors	0% (0)	3% (1)	0% (0)	0% (0)	1% (1)
No form of blinding stated	63% (19)	40% (12)	75% (15)	57% (17)	57% (63)
Total	100% (30)	100% (30)	100% (20)	100% (30)	100% (110)

Table 5.2

Approach to double-blinding by journal

Approach to double-blinding	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
Capsules	0% (0)	10% (1)	50% (1)	10% (1)	10% (3)
Tablets	22% (2)	50% (5)	0% (0)	30% (3)	32% (10)
Injections or Intravenous	44% (4)	0% (0)	0% (0)	20% (2)	19% (6)
Double-dummy	11% (1)	0% (0)	0% (0)	10% (1)	7% (2)
Other	11% (1)	40% (4)	50% (1)	30% (3)	29% (9)
No information	11% (1)	0% (0)	0% (0)	0% (0)	3% (1)
Total	100% (9)	100% (10)	100% (2)	100% (10)	100% (31)

Only 16% provided any written assurances of successfully implementing blinding, with small differences among the four journals (Table 5.3). Only about one-quarter of the trial reports provided any additional information, beyond the minimal prerequisite described above, on the methods used to ensure successful blinding (Table 5.3). Those methods included such things as keeping the code in a secure location, or not breaking the code until the end of the study. The BJOG provided this information more frequently than did the other three combined (p=0.03, 1 d.f.).

Table 5.3

Assurances of successful blinding and descriptions of the protection of allocation sequences in double-blind trials by journal

Item	Am J Obstet Gynecol (n=9)	Br J Obstet Gynaecol (n=10)	J Obstet Gynaecol (n=2)	Obstet Gynecol (n=10)	Total (n=31)
Authors stated an assurance of successful implementation of blinding	11% (1)	20% (2)	0% (0)	20% (2)	16% (5)
Authors stated an assurance of protection of allocation sequence[1]	22% (2)	50% (5)	0% (0)	10% (1)	26% (8)

[1] For example, statements such as the 'code kept in a secure location' or 'code not broken until the end of the study.'

Investigators tested the efficacy of blinding in only two trials, once in each of the American journals. Both found substantial deciphering of the assignments. Another trial claiming to be double-blind reported the use of assignment by hospital number, which was judged unlikely to have facilitated double-blinding. Thus, at least three trials indicated violation of blinding. In two reports authors explicitly stated that investigators and participants had been kept blinded to any interim analytical results of their trials. Only one report stated that the investigators analysing the trial results had been kept unaware of the actual treatments represented by the respective groups. Keeping data analysts blinded, however, may not be crucial to preventing bias.

5.4.2 Exclusions after randomisation

Nine of the trials (8%) did not provide sufficient information on the number of participants randomised and analyzed to enable even an inference as to whether exclusions had taken place (Table 5.4). All four journals had about the same proportion of trials reporting no apparent exclusions, but when exclusions were reported, trials in the **BJOG** were more likely to have reported less than 10% exclusions (p=0.002; 1 d.f.). All the trials in **OG** provided sufficient information on the number randomised and analyzed to at least infer the number of exclusions.

Table 5.4

Apparent exclusions after allocation by journal

Apparent exclusions	Am J Obstet Gynecol	Br J Obstet Gynaecol	J Obstet Gynaecol	Obstet Gynecol	Total
None	43% (13)	50% (15)	40% (8)	43% (13)	45% (49)
Less than or equal to 10%	17% (5)	37% (11)	15% (3)	13% (4)	21% (23)
More than 10%	30% (9)	10% (3)	20% (4)	43% (13)	26% (29)
Insufficient information	10% (3)	3% (1)	25% (5)	0% (0)	8% (9)
Total	100% (30)	100% (30)	100% (20)	100% (30)	100% (110)

5.4.3 Trials with no apparent exclusions after randomisation

In 49 trials, authors reported analysing the same number of participants as randomised, thereby implying that no exclusions had occurred. However, only 22% of these explicitly stated that no exclusions had occurred after randomisation: 23% from AJOG, 40% from BJOG, 0% from JOG, and 15% from OG. The BJOG explicitly stated no exclusions more frequently than the other three combined (p=0.05; 1 d.f.). In only one (2%) of those 49 trials (a BJOG trial) did the authors explicitly state that they had performed an intent-to-treat analysis, i.e., they had analyzed all randomised participants in the originally allocated groups (Table 5.5). Furthermore, only 22% reported an adequate allocation concealment method, and only one adjusted for baseline differences in the analysis (Table 5.5). In trials that had explicitly reported no exclusions taking place, 45% (5/11) reported an adequate allocation concealment method; in trials that had just implicitly reported no exclusions, only 16% (6/38) reported an adequate allocation concealment method.

5.4.4 Trials with exclusions after randomisation

Fifty-two trials reported having excluded at least one participant. If authors reported exclusions, they usually reported the reasons for the exclusions. In these 52 trials, 65% provided reasons by treatment group assignment. The remaining 35% provided reasons, but only overall and not by treatment group. Fifty percent of the trials from AJOG, 86% from BJOG, 29% from JOG, and 77% from OG provided information on the group assignments of those excluded.

Of these 52 trials, only four trials from BJOG and two from OG explicitly reported having performed an intent-to-treat analysis. Furthermore, 27% reported an adequate allocation concealment method, and 17% adjusted for baseline differences in the analysis (Table 5.5). Four trials stipulated that the exclusions were made before the assignment code was broken.

Table 5.5

Three indicators of trial quality by reported information on exclusions after randomisation

Indicator	No apparent exclusions (n=49)	At least 1 exclusion (n=52)	Insufficient information (n=9)	Total (n=110)
Reported an intent-to-treat analysis	2% (1)	12% (6)	0% (0)	6% (7)
Adequate allocation concealment	22% (11)	27% (14)	11% (1)	24% (26)
Adjusted for baseline differences in analysis	2% (1)	17% (9)	0% (0)	9% (10)

5.5 Discussion

5.5.1 Blinding

Double-blinding in trials reduces bias (Chapter 4). Obstacles to its implementation abound, however. Indeed, investigators and participants may all attempt to decipher treatments, but effective methodological impediments should frustrate their efforts.

In this study, only 31 reports cited double-blinding. That number of reports seems small; by the reckoning in this study, at least twice as many could have used it. While advising double-blinding, one must acknowledge that in certain trials, particularly those with objective end-points such as death, the anticipated gain in bias reduction may not be worth the additional difficulty and cost. However, having acknowledged that, double-blinding should be used if it is feasible, or the investigator should at least provide a rationale for not doing so.

Distressingly, fewer than half the trials reporting double-blinding provided even minimal information on how it was accomplished. Even fewer reported additional useful information or assurances of successful blinding. Investigators rarely tested double-blinding. Therefore, the suspicion arises that double-blinding may have been ineffective in some of these trials.

The BJOG tended to provide more detailed information on blinding than the other journals. Most differences were modest, however, and most were not statistically significant. The JOG, usually the outlier, tended to provide the least information. In all journals, authors rarely reported the blinding of the investigators and participants to the emerging results of the trial.

Investigators conducting a double-blind trial should provide clarifying information: 1) approach (e.g. capsules, tablets, double-dummy with specifics); 2) characteristics (e.g. that appearance, taste, administration, or anything else minimally necessary for successful blinding were similar for the treatment and control regimens); 3) allocation schedule control (e.g. the location of the schedule during the trial and when it was broken for the analysis, and the special circumstances under which it could be broken for individual cases); 4) assurances of success or reports of failure (appraisals should include specifics of how the blinding failed); 5) assurances of blinding of any interim results that may be produced; and 6) tests of efficacy (any empirical examination of the degree of success attained). Tests of efficacy have their own reliability problems and may not always be feasible. The remainder of the above-mentioned information is easily reported and, along with allocation concealment information (Section ~~2.5.5~~ 6.4), should provide minimal criteria for assessment. If nothing else, these guidelines may begin a dialogue on the requisite elements for implementing and reporting double-blinding. The scientific community should reach a greater consensus on those elements.

Perhaps discussion should also begin on the rather enigmatic term 'double-blinding' ('double-masking'). Some use 'triple-blinding' or 'quadruple-blinding' depending upon the detail of their descriptions. Do they not mean that everyone involved in the trial -- participants, organizers, investigators, assessors, care-givers, and the like -- were all blinded? A more appropriate term might clarify descriptions, such as 'full-blinding'.

Moreover, it would also lead logically to specific enumeration of those blinded and not blinded in 'not-fully-blind' situations; terms such as 'single-blind' are just as enigmatically unspecific as 'double-blind.' As many as possible of the above-named reporting elements should also be addressed in other than fully-blind trials. However, the importance of providing that information has not been validated. While some blinding would intuitively seem better than no blinding, empirical evidence has not as yet confirmed that presumption.

5.5.2 Post-randomisation exclusions

Randomisation, when successfully implemented, eliminates selection bias. Absence of bias persists throughout a trial, however, only if the analysis includes all randomised participants in the originally assigned groups. Thus, handling of post-randomisation exclusions becomes a crucial aspect of trial execution. Unfortunately, most trials in this study provided poor information. Moreover, some authors may have erroneously reported the numbers of participants randomised, by group. Those reported as randomised may have actually been the numbers of participants they had analyzed, after having had exclusions throughout the trial. Readers of reports thereby would mistakenly believe that all randomised patients had been analyzed, creating a false sense of security regarding the quality of the trial.

That observation is consistent with a critique of trials from another discipline (Gøtzsche 1989a). In Chapter 4, an association between handling of exclusions and bias was not found, and that led to the speculation that some authors did not accurately account for exclusions in their reports. Indeed, one investigator documented that practice in two trials (Gøtzsche 1989b).

The findings in this chapter indirectly support that conjecture. Trials that reported any exclusions also more frequently reported adequate allocation concealment, an intent-to-treat analysis, and adjustment for baseline differences than trials that indicated no apparent exclusions. Because those three measures indicate good methodological quality, most methodologists would have expected the opposite, since the lack of exclusions is also usually thought to be an indicator for good quality. Furthermore, the reports of trials that had included explicit statements of no exclusions more frequently reported adequate allocation concealment than the reports of trials that had just implied no apparent exclusions. The deduction from these findings is that many authors reporting no apparent exclusions may in fact have encountered exclusions during the trial, but ignored them in the report.

Reports from the BJOG had better information on exclusions. The authors explicitly stated that no exclusions had taken place over two times more frequently than did the authors from the other three combined. When trials reported exclusions, articles from the BJOG showed small losses (less than 10%) over twice as frequently as those of the other three combined. Reports from the BJOG also more often included explicit descriptions of intent-to-treat analyses. Again, articles from JOG tended to provide the least informative descriptions.

Investigators must provide complete information on exclusions after randomisation including: 1) total number randomised and the numbers allocated to each group; 2) total number analyzed and the numbers analyzed in each group; 3) explicit description of the type of analysis, primarily of adherence to intent-to-treat principles in at least one analysis; 4) the reasons for exclusions by group, if exclusions occurred; 5) the outcomes at least until the time of exclusion, if exclusions occurred, should be included somewhere in the report, or preferably, outcomes through the end of the study; and 6) an explicit statement of the absence of exclusions, if relevant. Only with this minimal set of information can editors and readers judge the merits of a trial.

5.5.3 Conclusions

Authors poorly reported methods used to achieve full-blinding. Furthermore, in the two trials in which authors reported testing of full-blinding, both found that some assignments had been deciphered. Yet, the poor reporting of post-randomisation exclusions causes even greater concern. Some of the trials reporting no apparent exclusions may not only have had exclusions, but may have improperly handled exclusions. That can introduce bias. Conversely, those reporting exclusions may have represented many of the better trials. That situation breeds an insidious problem, because almost everyone views trials without exclusions as less biased. Thus, many biased trials may actually be viewed as 'unbiased' and many unbiased trials as 'biased.' Until authors comprehensively report post-randomisation exclusions, readers, editors, and those conducting systematic reviews should all be wary of this paradox.

Chapter 6

6. Summary

6.1 Review of results

6.1.1 Reporting of randomisation and allocation

The generation of unbiased comparison groups in controlled trials requires proper randomisation, yet authors usually provide inadequate information on the process they used. From reports of trials from obstetrics and gynaecology journals, only 32% reported having used an adequate method to generate random numbers, and only 23% contained information showing that steps had been taken to conceal assignment until the point of treatment allocation. Merely 9% of the reports of trials described adequate methods of both generation and concealment.

Additional analyses suggest that non-random manipulation of comparison groups and selective reporting of baseline comparisons had occurred. In reports of trials which had apparently used unrestricted randomisation, the sample sizes of the treatment and control groups differed by less than would be expected by chance. Furthermore, in reports of trials in which hypothesis tests had been used to compare baseline characteristics, only 2% of reported tests were statistically significant, lower than the expected rate of 5%.

6.1.2 Associations between methodological approaches and treatment effects

Inadequate methodological approaches in controlled trials, particularly those reflecting inadequate or unclear allocation concealment, were associated with reported larger treatment benefits. Compared with trials in which authors reported adequately concealed treatment allocation, trials in which authors reported inadequately or unclearly concealed allocation yielded larger estimates of treatment effects ($p<0.001$).

Odds ratios were distorted by 41% and 30% (adjusted), respectively. Those associations likely represent bias and are particularly disconcerting in light of the results from Chapter 3 in which over three-quarters of recently published trials reported either inadequately or unclearly concealed allocation.

Trials in which participants were excluded after randomization were not, however, associated with larger treatment effects, but that appeared to be due to incomplete reporting rather than to the unimportance of adequately handling exclusions. Lack of double-blinding in trials was also associated with larger treatment effects (p=0.01), albeit at a lower strength than the lack of adequate measures to conceal assignment.

6.1.3 Reporting of blinding and exclusions after randomisation

The results from Chapter 4 indicate that double-blinding protects against bias. Yet, in published trials from recent volumes of obstetrics and gynaecology journals (Chapter 5), only about half the trials that could have double-blinded actually did so. Furthermore, when investigators attempted double-blinding, they poorly reported methods. Of the 31 trials that reported being double-blind, a mere 45% provided at least minimally sufficient descriptions for successful double-blinding, and only 26% provided additional information on the protection of the allocation schedule throughout the trial. Moreover, only 16% of the reports provided any written assurances of successfully implementing blinding and only two tested the efficacy.

The poor reporting of post-randomization exclusions in recent volumes of obstetrics and gynaecology journals, however, causes greater concern. In 49 trials with no apparent exclusions after randomization, only 22% explicitly stated that no exclusions had taken place after randomization and only 2% stated that an intent-to-treat analysis had been performed. Moreover, those 49 trials generally provided methodological information of inferior quality to that provided from trials that reported exclusions after randomization.

These results from Chapter 5 generally support the findings from Chapter 4. Some of the trials reporting no apparent exclusions may not only have had exclusions, but may have improperly handled exclusions. Conversely, those reporting exclusions may have represented many of the better trials. That incongruity fosters a furtive problem, because almost everyone views trials without exclusions as less biased than those with exclusions. All who use the results of randomized controlled trials should be aware of those potential inconsistencies in the reporting of exclusions.

6.2 Implications for meta-analyses

6.2.1 Exclusion of trials from meta-analyses

The results in this thesis suggest some implications for meta-analyses. Trials using inadequately concealed allocation schemes should be the first to be considered for elimination from meta-analyses for quality concerns. Dealing with trials, however, that have used unclearly concealed approaches can be more problematic. They preferably should be excluded, as some do (Peto 1987). If that practice were carried out at the present time, however, meta-analysts in many medical specialties would be left with very few, if any, trials for their reviews. Indeed, the examinations in this thesis of the meta-analyses from pregnancy and childbirth topics revealed that many did not contain even one trial that had used an appropriate allocation concealment scheme. Advocating the exclusion of all inadequately and unclearly concealed trials under such situations might be inappropriate, for that would leave the analyst without any trials to analyze.

6.2.2 Adjustments to the results of presumed biased trials

Some may argue for including inadequately and unclearly concealed trials, but using adjustments for the bias. While that approach may have appeal, it is quite risky because of the great uncertainty in assigning adjustment factors. In estimating the effects of unclearly concealed trials on treatment effects in this thesis, statistically significant heterogeneity existed among the meta-analyses. Thus, using an overall factor based on this study may not even apply to another meta-analysis of perinatal trials let alone another meta-analysis in a totally different medical specialty.

6.2.3 Use of a random effects model

Using a random effects model for analysis in some situations may be helpful because it would account for an additional component of between-trial variance. That might be particularly useful when trials using combinations of adequately, unclearly, and inadequately concealed approaches are included in a meta-analysis because of the likely additional heterogeneity and greater between-trial variance. The confidence interval for the overall odds ratio using a random effects model might be more likely to encompass the true effect than the overall odds ratio using a fixed effect model.

Nevertheless, if an analyst includes biased trials in the meta-analysis, the point estimate of the summary odds ratio would still be biased. That would have to be clearly explicated in the discussion of the results. Of concern also is that the between-trial variance could be relatively small, particularly if all the trials used unclearly concealed allocation. (Remember that the heterogeneity in the unclearly concealed trials was not too different from that in the adequately concealed trials.) With small between-trial variance among unclearly concealed trials, the confidence interval from a random effects model would be very similar to that from a fixed effect model and not of much assistance. Thus, while the use of a random effects model may be helpful in some situations, it is no panacea, and does not alter the need for carefully phrased qualifications to accompany quantitative results based on potentially biased trials. It certainly does not supplant the need for trials that have used adequate allocation concealment.

Moreover, the random effects approach cannot be viewed as a panacea for other reasons. Rather unrealistic assumptions pertain to the heterogeneity between trials being represented by a single variance and the between-trial distribution being Normal (Thompson 1993; Thompson and Pocock 1991). Because of those and other problems, Thompson (1993) suggests using the random effects method as a type of sensitivity analysis, investigating "how much the overall conclusions change as the assumptions underlying the statistical methodology also change." The use of random effects models in general warrant further exploration, but particularly with respect to between-trial variance caused by variations in methodological quality.

6.2.4 Tentative suggestions

Restricting meta-analyses to adequately concealed trials seems prudent. If an insufficient number are available, potentially biased trials may have to be included and sensitivity analyses performed. The results, however, would need to be appropriately qualified.

Another potentially viable option would be to stratify by, and separately analyze, the adequately and unclearly concealed trials. A test of heterogeneity between those two strata could be implemented. In general, however, the issue of proper handling of poor quality (biased) trials in meta-analyses deserves greater attention. Section 6.3.3 contains a recommendation for the next step in providing that attention.

6.3 Further research

6.3.1 Replication

Other investigators certainly should implement analyses on other data sets similar to those contained in this thesis. As the Cochrane Collaboration grows, a great many opportunities will materialize to replicate this work. Indeed, even the Cochrane Collaboration Pregnancy and Childbirth Module has increased in size since this analysis began, and someone could already extend this research to additional meta-analyses within that module.

Beyond replication, suggestions for further research are offered in the next two sections. The following topics represent the two highest priorities.

6.3.2 Unpublished trials and estimates of treatment effects

In Chapter 4, only published trials were used for analysing the association between methodological quality and treatment effects, for obvious reasons. Thus, the results only pertain to published information. But consumers of RCTs, particularly meta-analysts, need to be able to deal with unpublished trials. That becomes particularly important with the potential for publication bias.

Additional research should delve into this realm. The association between unpublished trials and treatment effects should be examined, with the study design resembling that found in Chapter 4. The unpublished trials as a group would represent a quality level that would be compared against the adequately, unclearly, and inadequately concealed trials in the database.

6.3.3 Allocation concealment and heterogeneity in meta-analyses

From Chapter 4, differential treatment effects appeared to be related to allocation concealment. From Chapters 3 and 4, many trials did not report adequate concealment. The suspicion that differing levels of allocation concealment within meta-analyses cause greater heterogeneity appears warranted. Thus, accounting for allocation concealment in a meta-analysis, e.g. by exclusion of inadequately concealed trials or by separate analyses according to allocation concealment strata, may address much of the apparent heterogeneity.

That issue should be explored in a sample of meta-analyses. I anticipate doing so with the data set used in Chapter 4, but hope that other investigators will analyze other data sets with the same intent.

6.4 Comments on randomisation

6.4.1 Shared responsibility for poor reporting

Beyond the discussions in the chapters, a few particulars of randomisation warrant more comment. That such a vast majority of the reports of trials failed to describe adequate methods of randomisation generation and randomisation concealment is disappointing, particularly since inadequate descriptions appeared to be associated with biased results. Those deficiencies in randomisation partly reflect the fact that people with an interest in methodology, such as myself, have failed to provide adequate guidance in the literature.

6.4.2 Generation of the allocation sequence

Nevertheless, having already referred to deficiencies in the literature, available texts generally attend well to the details of generating randomised assignment schedules. One aspect of the process, however, deserves greater attention. If randomisation involves blocking, the block size should be randomly varied to reduce the chances that the assignment schedule will be inferred by those responsible for recruiting participants.

If the trial is not blinded and the block size is fixed, particularly if the size is small (8 or less), the allocation schedule would be too predictable. The block size would invariably be deciphered and selection bias would be introduced, *regardless* of the effectiveness of allocation concealment. Moreover, even if the study is double-blinded, many treatments have obvious side-effects which may eventually lead to the deciphering of the block size. Thus, even in those instances, varying the block size would be recommended as a precautionary measure so as not to jeopardize the effectiveness of the randomisation process.

6.4.3 Need for greater emphasis on allocation concealment

Many of the available texts, on the other hand, contain less useful information on highlighting the importance and mechanics of randomisation concealment. That deficiency needs to be addressed in more detail.

For those involved in implementing a trial that has not incorporated proper procedures for randomisation concealment, the challenge of deciphering a randomisation scheme frequently presents too great a temptation to resist. Succumbing to that temptation may at times be innocent and a reflection of human inquisitiveness and ingenuity rather than scientific malevolence; but, whatever the motivation, the net effects are the same if the introduction of selection bias invalidates the comparisons made in the trial.

At least a few investigators have been sufficiently astute to identify and candid enough to report the deciphering of their schemes (King 1959; Kirkland et al 1960; Wood et al 1981; Lazar et al 1984). Moreover, I have heard many 'off the record' admissions of deciphering. Methodological safeguards must be established to impede investigators from contaminating trials with bias.

6.4.4 The leniency of judgements on concealment

While many readers of this thesis may believe the judgements on the quality of allocation concealment were too harsh, the judgements may actually have been too lenient. For example, sealed envelopes are more susceptible to manipulation through human ingenuity than other approaches. Therefore, some consider them as generally less desirable (Pocock 1982). If investigators use envelopes, every methodological nuance should be addressed to subvert attempts at breaking the randomisation, and, of course, all methods should be reported. Only about a quarter of the reports in the obstetrics and gynaecology journals in which the use of envelopes for concealment was described met the minimum criteria of using sequentially numbered, sealed, opaque envelopes. All of those components are important (Altman and Doré 1990), but, in addition, the trials should have reported that the envelopes had been opened sequentially, and only after the participant's name and other details had been written on the appropriate envelope (Bulpitt 1983). Furthermore, using pressure sensitive or carbon paper inside the envelope transfers such information to the assigned allocation and thus creates a valuable audit trail.

Reports in which authors stated the allocation schedule to have been prepared by the pharmacy were classified as having used an acceptable approach to concealment (Altman and Doré 1990). The compliance of pharmacists with proper randomisation concealment methodology in these trials is unknown, however, and the precautions taken should have been reported, but generally were not. In the past, in certain instances, pharmacists have been responsible for gross distortions of assignment schedules. For example, one large pharmacy allocated all participants to one arm of a two-group trial because they ran out of one drug on the weekend. Another large pharmacy generated the allocation schedule thinking that 'alternate assignment' met the randomisation criteria. Investigators should not assume that pharmacists, and others involved in their trials for that matter, are knowledgeable in RCT methodology and should ensure that their research partners follow proper trial procedures.

The use of numbered or coded containers helps to prevent foreknowledge of treatment assignment, but the criteria in this thesis only required a statement that such precautions had been taken, without requiring further details of how this had been achieved. Assurances that all of the containers were of equal weight and similar appearance, and that some audit trail had been established (such as writing the names of participants on the empty bottles or containers), would help readers to assess whether randomisation was likely to have been concealed successfully. Similarly, although central telephone randomisation was counted as an acceptable approach to allocation concealment, that general criterion might be regarded as having been lenient in not requiring details of the actual procedures used.

With all approaches, the person(s) who prepared the randomisation scheme ideally should not be involved in determining eligibility, administering treatment, or assessing outcome. When considered for a moment, the rationale for that proscription becomes obvious. Regardless of the methodological quality of the allocation generation and concealment process, such an individual would always have access to the allocation schedule and thus the opportunity to introduce bias. Nevertheless, under some extraordinary circumstances, someone may have to prepare the scheme and be involved in the trial. In these instances, the investigators must make sure that the assignment schedule is unpredictable and locked away from even the person(s) who generated it. Thus, who prepared the scheme should have been reported, and the criteria used in this thesis could again be assailed as being too lenient in not requiring the presentation of those details.

6.4.5 Improving the standard of reporting

While improving the standard of reporting is surely a shared responsibility, omission of randomisation details to date has probably been primarily an author-based phenomenon rather than due 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.

Protestations from authors about lack of space does not constitute 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 randomisation process has been omitted. Yet in a well-executed, blinded, randomised controlled trial, many aspects other than randomisation become almost scientifically inconsequential to the treatment comparisons since they would have been applied equally to unbiased comparison groups.

Certainly, I 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. In particular, authors must present more complete information on exclusions after randomisation and include additional information on blinding procedures. Yet, proper reporting of the randomisation procedures should be of the highest priority, and one should have little confidence in those trials failing to provide that information.

6.5 Final thoughts

This thesis research has clearly revealed the vital importance of adherence to proper design principles in conducting trials. Investigators must properly design, meticulously execute, and completely report their trials. In particular, if they do not conceal treatment allocation, bias will likely distort their results. Investigators, editors, and readers need to be aware of that association. Randomised controlled trials supposedly minimize bias; the research community and public should expect nothing less.

References

Aitkin M, Anderson D, Francis B, and Hinde J (1989) ***Statistical Modelling in GLIM***. Oxford University Press, Oxford, UK.

Altman DG (1982) Statistics in medical journals. ***Stat Med*** 1, 59-71.

Altman DG (1985) Comparability of randomised groups. ***Statistician*** 34, 125-136.

Altman DG (1991a) ***Practical Statistics for Medical Research***. Chapman and Hall, London, UK.

Altman DG (1991b) Randomisation: essential for reducing bias. ***BMJ*** 302, 1481-1482.

Altman DG and Doré CJ (1990) Randomisation and baseline comparisons in clinical trials. ***Lancet*** 335, 149-153.

Altman DG and Gardner MJ (1986) Presentation of variability (letter). ***Lancet*** ii, 639.

Armitage P (1982) The role of randomization in clinical trials. ***Stat Med*** 1, 345-352.

Bracken MB (1989) Reporting observational studies. ***Br J Obstet Gynaecol*** 96, 383-388.

Bulpitt CJ (1983) ***Randomised Controlled Clinical Trials***. Martinus Nijhoff, The Hague, Netherlands.

Byar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellenberg JH, Gail MH and Ware JH (1976) Randomized controlled trials: Perspectives on some recent ideas. ***N Engl J Med*** 295, 74-80.

Chalmers I (ed) (1988) ***Oxford Database of Perinatal Trials***. Version 1.2, disk issue 1, Oxford University Press, Oxford, UK.

Chalmers I (ed) (1992) ***Oxford Database of Perinatal Trials.*** Version 1.2, disk issue 8. Oxford University Press, Oxford, UK.

Chalmers I, Adams M, Dickersin K, Hetherington J, Tarnow-Mordi W, Meinert C, Tonascia S and Chalmers TC (1990) A cohort study of summary reports of controlled trials. ***JAMA*** 263, 1401-1405.

Chalmers I, Enkin M and Keirse MJNC (1993) Preparing and updating systematic reviews of randomized trials of health care. ***Millbank Q*** 71, 411-437.

Chalmers I, Hetherington J, Elbourne D, Keirse MJNC and Enkin M (1989) Materials and methods used in synthesizing evidence to evaluate the effects of care during pregnancy and childbirth. In ***Effective Care in Pregnancy and Childbirth. Volume 1: Pregnancy*** (Chalmers I, Enkin M, and Keirse MJNC, eds) Oxford University Press, Oxford, UK, 39-65.

Chalmers I, Hetherington J, Newdick M, Mutch L, Grant A, Enkin M, Enkin E and Dickersin K (1986) The Oxford Database of Perinatal Trials: developing a register of published reports of controlled trials. ***Controlled Clin Trials*** 7, 306-324.

Chalmers TC (1983) The control of bias in clinical trials. In ***Clinical Trials: Issues and Approaches*** (Shapiro SH and Louis TA, eds) Marcel Dekker, New York, USA, 115-127.

Chalmers TC, Berrier J, Sacks HS, Levin H, Reitman D and Nagalingam R (1987a) Meta-analysis of clinical trials as a scientific discipline. II: Replicate variability and comparison of studies that agree and disagree. ***Stat Med*** 6, 733-744.

Chalmers TC, Celano P, Sacks HS and Smith H Jr (1983) Bias in treatment assignment in controlled clinical trials. ***N Engl J Med*** 309, 1358-1361.

Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J and Nagalingam R (1987b) Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials. ***Stat Med*** 6, 315-325.

Chalmers TC, Matta RJ, Smith H Jr and Kunzler AM (1977) Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. ***N Engl J Med*** 297, 1091-1096.

Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D and Ambroz A (1981) A method for assessing the quality of a randomized control trial. ***Controlled Clin Trials*** 2, 31-49.

Cochrane AL (1979) 1931-1971: A critical review, with particular reference to the medical profession. In ***Medicines for the Year 2000***. Office of Health Economics, London, UK, 1-11.

Colditz GA, Miller JN and Mosteller F (1989) How study design affects outcomes in comparisons of therapy. I: Medical. ***Stat Med*** 8, 441-454.

Collins R, Gray R, Godwin J and Peto R (1987) Avoidance of large biases in the assessment of moderate treatment effects: The need for systematic overviews. ***Stat Med*** 6, 245-250.

Dean AG, Dean JA, Burton AH and Dicker RC (1990) ***Epi Info, Version 5: a Word Processing, Database, and Statistics System for Epidemiology on Microcomputers***. Centers for Disease Control and Prevention, Atlanta, GA, USA.

De Jonge H (1983) Deficiencies in clinical reports for registration of drugs. ***Stat Med*** 2, 155-166.

DerSimonian R, Charette LJ, McPeek B and Mosteller F (1982) Reporting on methods in clinical trials. ***N Engl J Med*** 306, 1332-1337.

Detsky AS, Naylor CD, O'Rourke K, McGeer AJ and L'Abbe KA (1992) Incorporating variations in the quality of individual randomized trials into meta-analysis. ***J Clin Epidemiol*** 45, 255-265.

Emerson JD, Burdick E, Hoaglin DC, Mosteller F and Chalmers TC (1990) An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. ***Controlled Clin Trials*** 11, 339-352.

Emerson JD, McPeek B and Mosteller F (1984) Reporting clinical trials in general surgical journals. ***Surgery*** 95, 572-579.

Enkin MW, Keirse MJNC, Renfrew MJ and Neilson JP (1993) ***Cochrane Database of Systematic Reviews***: Published through 'Cochrane Updates on Disk', Update Software, Oxford, UK.

Evans M and Pollock AV (1984) Trials on trial: a review of trials of antibiotic prophylaxis. ***Arch Surg*** 119, 109-113.

Fisher RA (1966) ***The Design of Experiments***. 8th ed. Oliver and Boyd Limited, Edinburgh, UK.

Fletcher RH and Fletcher SW (1979) Clinical research in general medical journals. A 30-year perspective. ***N Engl J Med*** 301, 180-183.

Francis B, Green M, Payne C, Swan T, Gilchrist R, Bradley M, Clarke M, Green P, Reese A, Hinde J, Stalewski A, O'Brien C (1993) ***GLIM 4: The Statistical System for Generalized Linear Interactive Modelling***. Oxford University Press, Oxford, UK.

Freiman JA, Chalmers TC, Smith H and Kuebler RR (1978) The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. ***N Engl J Med*** 299, 690-694.

Gehan EA and Freireich EJ (1974) Non-randomized controls in cancer clinical trials. ***N Engl J Med*** 290, 198-203.

Gillman MW and Runyan DK (1984) Bias in treatment assignment in controlled clinical trials (letter). ***N Engl J Med*** 310, 1610-1611.

Gøtzsche PC (1986) Enalapril, atenolol, and hydrochlorothiazide in hypertension. ***Lancet*** ii, 38-39.

Gøtzsche PC (1989a) Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. ***Controlled Clin Trials*** 10, 31-56.

Gøtzsche PC (1989b) Multiple publication of reports of drug trials. ***Eur J Clin Pharmacol*** 36, 429-432.

Gøtzsche PC (1989c) Patients' preference in indomethacin trials: an overview. ***Lancet*** i, 88-91.

Grant A (1989) Reporting controlled trials. ***Br J Obstet Gynaecol*** 96, 397-400.

Green SB (1982) Patient heterogeneity and the need for randomized clinical trials. ***Controlled Clin Trials*** 3, 189-198.

Greenland S (1990) Randomization, statistics, and causal inference. ***Epidemiology*** 1, 421-429.

Grimes DA (1991) Randomized controlled trials: "it ain't necessarily so". ***Obstet Gynecol*** 78, 703-704.

Grimes DA and Schulz KF (1992) Randomized controlled trials of home uterine activity monitoring: A review and critique. ***Obstet Gynecol*** 79, 137-142.

Hetherington J, Dickersin K, Chalmers I and Meinert CL (1990) Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians. ***Pediatrics*** 84, 374-380.

Hill AB (1952) The clinical trial. ***N Engl J Med*** 247, 113-119.

Huskisson EC and Scott J (1976) How blind is double blind? And does it matter? ***Br J Clin Pharmacol*** 3, 331-332.

Institute of Medicine (1992) ***Strengthening Research in Academic OB/GYN Departments: Committee on Research Capabilities of Academic Departments of Obstetrics and Gynecology*** (Jessica Townsend, ed) National Academy Press, Washington, DC, USA.

Karlowski TR, Chalmers TC, Frenkel LD, Kapikian AZ, Lewis TL and Lynch JM (1975) Ascorbic acid for the common cold. A prophylactic and therapeutic trial. ***JAMA*** 231, 1038-1042.

Keirse MJNC (1988) Amniotomy or oxytocin for induction of labor: Re-analysis of a randomized controlled trial. ***Acta Obstet Gynecol Scand*** 67, 731-735.

King AG (1959) Prevention of puerperal breast engorgement with large doses of long-acting estrogen. ***Am J Obstet Gynecol*** 78, 80-85.

Kirkland JA, Greenberg BG and Flowers CE (1960) Suppression of lactation: a double-blind hormone study. ***Obstet Gynecol*** 15, 292-298.

Lavori PW, Louis TA, Bailar JC and Polansky M (1983) Designs for experiments -- parallel comparisons of treatment. ***N Engl J Med*** 309, 1291-1299.

Lazar P, Gueguen S, Dreyfus J, Renaud R, Pontonnier G and Papiernik E (1984) Multicentred controlled trial of cervical cerclage in women at moderate risk of preterm delivery. ***Br J Obstet Gynaecol*** 91, 731-735.

Liberati A, Himel HN and Chalmers TC (1986) A quality assessment of randomized controlled trials of primary treatment of breast cancer. ***J Clin Oncol*** 4, 942-951.

May GS, DeMets DL, Friedman LM, Furberg C and Passamani E (1981) The randomized clinical trial: bias in analysis. ***Circulation*** 4, 669-673.

Macfarlane AJ (1978) Variations in numbers of births and perinatal mortality by day of week in England and Wales. *BMJ* 2, 1670-1673.

McCullagh P and Nelder JA (1983) ***Generalized Linear Models.*** Chapman and Hall, London, UK.

Medical Research Council (1948) Streptomycin treatment of pulmonary tuberculosis. *BMJ* 2, 769-782.

Meinert CL (1986) ***Clinical Trials: Design, Conduct, and Analysis.*** Oxford University Press, New York, USA.

Meinert CL, Tonascia S and Higgins K (1984) Content of reports on clinical trials: a critical review. ***Controlled Clin Trials*** 5, 328-347.

Miller JN, Colditz GA and Mosteller F (1989) How study design affects outcomes in comparisons of therapy. II: Surgical. ***Stat Med*** 8,455-466.

Mosteller F, Gilbert JP and McPeek B (1980) Reporting standards and research strategies for controlled trials: agenda for the editor. ***Controlled Clin Trials*** 1, 37-58.

Naylor CD (1989) Meta-analysis of controlled clinical trials. ***J Rheum*** 16, 424-426.

Nurmohamed MT, Rosendaal FR, Buller HR, Dekker E, Hommes DW, Vandenbroucke JP and Briet E (1992) Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. ***Lancet*** 340, 152-156.

Office of Technology Assessment (OTA), U.S. Congress (1983) ***The Impact of Randomized Controlled Trials on Health Policy and Medical Practice.*** U.S. Government Printing Office, Washington, DC, USA.

Peto R (1987) Why do we need systematic overviews of randomized trials? ***Stat Med*** 6, 233-240.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson C, Peto J and Smith PG (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I: Introduction and design. ***Br J Cancer*** 34, 585-612.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson C, Peto J and Smith PG (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II: Analysis and examples. ***Br J Cancer*** 35, 1-39.

Pignon JP and Poynard T (1990) Statistics in clinical trials (letter). ***Lancet*** 335: 614.

Pocock SJ (1979) Allocation of patients to treatment in clinical trials. ***Biometrics*** 35, 183-197.

Pocock SJ (1982) Statistical aspects of clinical trial design. ***Statistician*** 31, 1-18.

Pocock SJ (1983) ***Clinical Trials: A Practical Approach***. Wiley, Chichester, UK.

Pocock SJ, Hughes MD and Lee RJ (1987) Statistical problems in the reporting of clinical trials. ***N Engl J Med*** 317, 426-432.

Preston DL and Lubin JH (1993) ***GMBO version 1.8w: Epicure***. HiroSoft International, Seattle, USA.

Rothman KJ (1977) Epidemiologic methods in clinical trials. ***Cancer*** 39 (suppl 4), 1771-1775.

Sackett DL and Gent M (1979) Controversy in counting and contributing events in clinical trials. ***N Engl J Med*** 301, 1410-1412.

Sacks HS, Berrier J, Reitman D, Ancona-Berk VA and Chalmers TC (1987) Meta-analyses of randomized controlled trials. ***N Engl J Med*** 316, 450-455.

Sacks H, Chalmers TC and Smith H Jr (1982) Randomized versus historical controls for clinical trials. ***Am J Med*** 72, 233-240.

Thacker SB (1987) The efficacy of intrapartum electronic fetal monitoring. ***Am J Obstet Gynecol*** 156, 24-30.

Thompson SG (1993) Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. ***Statistical Methods in Medical Research*** 2, 173-192.

Thompson SG and Pocock SJ 1991 (1991) Can meta-analyses be trusted? ***Lancet*** 338, 1127-1130.

Tyson JE, Furzan JA, Reisch JS and Mize SG (1983) An evaluation of the quality of therapeutic studies in perinatal medicine. ***J Pediatr*** 102, 10-13.

Wald N and Cuckle H (1989) Reporting the assessment of screening and diagnostic tests. ***Br J Obstet Gynaecol*** 96, 389-396.

Williamson JW, Goldschmidt PG and Colton T (1986) The quality of medical literature. In ***Medical Uses of Statistics*** (Bailar JC III and Mosteller F, eds) NEJM Books, Waltham, MA, USA, 370-391.

Wood C, Renou P, Oats J, Farrell E, Beischer N and Anderson I (1981) A controlled trial of fetal heart rate monitoring in a low-risk obstetric population. ***Am J Obstet Gynecol*** 141, 527-534.