11

CONTROLLED STUDIES

OBJECTIVES

After reading this chapter, you should be able to:

1. Design a controlled trial to produce a valid and efficient evaluation of an intervention, paying special attention to:
    a. the statement of objectives of the trial
    b. the definition of the study subjects
    c. the allocation of subjects to the interventions
    d. the identification and definition of appropriate outcome variables
    e. ethical considerations in the design and implementation of the trial.

2. Conduct a controlled trial efficiently, while paying special attention to:
    a. masking as a procedure to reduce bias
    b. following all intervention groups adequately and equally
    c. developing and using appropriate data-collection methods and instruments
    d. proper assessment of the outcomes being measured
    e. correct analysis and interpretation of the results
    f. clear and complete reporting of methods and results.

3. Design and conduct a valid controlled trial of a vaccine, or prophylactic, against an infectious agent.

11.1 Introduction

A randomised controlled trial (RCT) is a planned experiment carried out on subjects in their usual environment. Particular care must be taken in the design and execution of these studies because they involve people, and their size and scope make it very difficult to replicate them for the purpose of validating the findings. Despite the importance of RCTs, "Overwhelming evidence shows the quality of reporting of RCTs is not optimal. Without transparent reporting, readers cannot judge the reliability and validity of trial findings nor extract information for systematic reviews. Recent methodological analyses indicate that inadequate reporting and design are associated with biased estimates of treatment effects" (Ocana *et al*, 2011). Such systematic error is damaging to RCTs, which are considered to be the gold standard for evaluating interventions because of their ability to minimise or avoid bias (Moher *et al*, 2010; Schulz *et al*, 2010). However, as Boutron *et al* (2005) indicate, a single RCT is rarely sufficient to answer questions about complex interventions. Thus, a group of scientists and editors developed the Consolidated Standards of Reporting Trials (CONSORT) statement to improve the quality of reporting of RCTs. It was first published in 1996, updated in 2001 and again in 2010 (see Section 11.12). There is evidence that reporting (and design) has improved since these standards were published (Kane *et al*, 2007; Moher *et al*, 2005); nonetheless, room for improvement remains (Berwanger *et al*, 2009). Rather than cite the recent extensive publications which still find bias, errors, and inadequate reporting, we will assume that we can do better. To that end, we will use the main headings from CONSORT as our template in this chapter; the actual checklist is in Table 11.1

RCTs are especially useful for the evaluation of interventions that can be easily manipulated, such as therapeutic or prophylactic products, diagnostic procedures, and health programmes. Most trials are conducted to assess one specific intervention and, indeed, this is their forte. The outcome might include a specific health parameter (*eg* clinical disease), or a measure of quality of life or longevity. The study groups are formed by random assignment of the intervention(s) being evaluated and can be composed of individuals or groups. Lavori and Kelsey (2002) edited a comprehensive review of clinical trials; this provides an excellent overview of trial design, analysis and interpretation. A special issue of Statistics in Medicine (Vol 21, Issue 19, 2002) was devoted to a discussion of the design of long-term clinical trials. In the same year, the editors of *The Lancet*, Schulz and Grimes, wrote a series of articles on the design of RCTs (2002a; 2002b; 2002c; 2002d; 2002e).

The term **clinical trial** is often used synonymously for controlled trial. However, some authors restrict its use to trials of therapeutic products and/or trials carried out in a clinical setting. Similarly, the term **field trial** is often used for studies which are carried out in a general population setting. We will use 'randomised controlled trial' to refer to all planned experiments designed to evaluate products or procedures in subjects outside the laboratory. Because RCTs can be used to investigate a wide range of products/programmes, we will refer to the factor being investigated (*eg* treatment) as the **intervention**, and to the effect of interest as the **outcome**. People or groups participating in the trial will be referred to as **subjects** or **participants**.

RCTs are by far the best way for evaluating health interventions because they allow much better control of potential confounders than observational studies, as well as reducing bias due to selection and misinformation (..."the randomised controlled trial is at present the unchallenged source of the highest standard of evidence used to guide clinical decision-making" (Lavori, P. and Kelsey, 2002).) In the absence of evidence as to the efficacy and safety of health products and procedures derived from controlled trials, clinicians would be left in the unenviable position of making decisions about their use based on extrapolation of data from studies carried out under artificial (laboratory) conditions or based on their own limited and uncontrolled experience. Having said this, the results of many trials have been criticised for being "of limited relevance to answering questions about whether an intervention does work under usual circumstances" (Treweek and Zwarenstein, 2009; Zwarenstein *et al*, 2006). This issue has led others to describe how to design trials to investigate practical problems (Treweek *et al*, 2006) and to develop a specific 'tool' to help researchers prepare high-quality research proposals for clinical trials.

In order to encourage (or force) transparency regarding what clinical trials have been done and what the original design was, "as of 2005, the International Committee of Medical Journal Editors required investigators to register their trials prior to participant enrolment as a precondition for publishing the trial's findings in member journals" (Mathieu *et al*, 2009). Unfortunately, many trials are not registered until late in their implementation, and often the reported outcomes do not match the registered outcomes indicating considerable selective reporting (Ocana and Tannock, 2011). There is an International Clinical Trials Registry Platform (ICTRP) operated by WHO. However, at present, registration appears to be voluntary.

11.1.1 Phases of clinical research

While controlled trials are valuable for assessing a wide range of factors affecting health, one of their most common uses is to evaluate pharmacological products (therapeutic and preventive). Consequently, a brief review of the phases of research used in the development and evaluation of these products is warranted.

Before pharmaceuticals companies start clinical trials on a drug, they conduct extensive pre-clinical studies. These involve in vitro (test tube or cell culture) and in vivo (animal) experiments using wide-ranging doses of the study drug to obtain preliminary efficacy, toxicity, and pharmacokinetic information. Clinical pharmaceutical research can be divided into 4 or 5 phases.

Phase 0 A recent designation for exploratory, 'first-in-human' trials. These studies, conducted with a few subjects (n<15) establish whether the drug or agent behaves in human subjects as was expected from preclinical studies. Drug dosages are typically subtherapeutic.

Phase I trials (sometimes referred to as formulation trials) are studies carried out in a small group of 20–100 healthy volunteers. This phase is designed to assess the safety and pharmacodynamics of a drug. The subject who receives the drug is usually observed until several half-lives of the drug have passed. People are paid an inconvenience fee for their time spent in the volunteer centre.

Phase II trials are the first evaluation of the drug in larger groups (100–300) and are designed to assess how well the drug works. Some Phase II trials are designed as case-series, others as RCTs. Ocana *et al* (2012) note that 'synergy' is often cited in Phase I and II trials, but the term is often misused. Lara and Redman (2012) warn against over-interpretation of Phase II data; their predictive value of success in Phase III trials is unknown or poor (at least in the area of cancer control).

Phase III trials are usually large-scale RCTs, but some are observational studies, designed to determine the efficacy of a drug in a typical clinical population, to monitor side effects, and to compare the drug with other available treatments. Phase III RCTs are often conducted in mutiple centres using large patient groups (300–3,000 or more depending upon the disease/medical condition studied) (see http://www.accessdata.fda.gov/). The comparison intervention usually is the current gold-standard treatment. Korn *et al* (2012) discuss Phase II and III trials and recommend combining them into a randomised adaptive PhaseII/III design. Pitrou *et al* (2009) indicate that reporting of adverse effects was quite variable in the 133 reports of the RCTs they reviewed.

Phase IV trials/studies are non-mandatory post-registration trials designed to evaluate the most effective way of using a product and assessing its long-term safety. These may be conducted as RCTs in settings different from the Phase III trials, or are conducted using observational techniques. They provide the most reliable information about the efficacy of a product in the context of everyday real-world activities. The safety surveillance is designed to detect any rare or long-term adverse effects over a much larger patient population and longer time period than was possible during the Phase I–III clinical trials.

11.1.2 Key design elements

An important feature in the design of a controlled trial is the development of a detailed study protocol which covers all elements of the study design and execution. This ‘road map’ includes: stating the objectives, defining the source population in which the study will be conducted, allocation of subjects, specifying the intervention, masking (blinding), follow-up and compliance, specifying and measuring the outcome, analysis of trial results, and ethical considerations. These aspects of the trial design are related to the features that should be reported when describing the results of a trial (see Table 11.1). Each of us should be aware of the common biases that can impact the design and reporting of trial results in order to minimise these (Gluud, 2006).

11.2 Background, objectives, and summary trial design

The objective(s) of the trial must be stated clearly and succinctly. The objective should describe the intervention being investigated, the allocation design (factorial, cross-over, *etc*), and the primary outcome(s) to be measured. In addition, the title of any RCT proposal should include the term (RCT or similar) to enhance ‘searchability’ of pertinent research. As a general rule, each trial should have a limited number of objectives (see Example 11.1 for a straightforward trial of an evaluation of a prostate cancer screening programme with time to diagnosis of prostate cancer and time to death as the 2 outcomes of interest). Some trials might also include a small number of secondary outcomes. Increasing the number of objectives may unnecessarily complicate the protocol and might jeopardise compliance and other aspects of the trial. A trial with a very simple design might be able to include a much larger sample size within a given budget, thus enhancing the power of the study. Lancaster *et al* (2010) provide a thorough description of the main elements of a RCT, especially those RCTs conducted in a primary care setting where the interventions can become quite complex and where cluster randomisation is common. Peduzzi *et al* (2010) focus their discussion of clinical trial design on comparative research.

This chapter will focus on controlled trials that contrast 2 groups, the intervention and the comparison (sometimes referred to as 2-arm studies) (*eg* Example 11.2), although the principles

Example 11.1 A randomised controlled trial of prostate cancer screening

A total of 38340 men aged 55–74 years was randomly assigned to the intervention arm and 38345 men to the control (usual care) arm at 10 screening centres in the USA between 1993 and 2001 (Andriole *et al*, 2012). Men who were randomly assigned to the intervention arm were offered screening with annual prostate specific antigen (PSA) tests for 6 years and digital rectal examination (DRE) for 4 years; screening was completed in October 2006. A positive test was defined as a PSA value greater than 4 ng/mL, or a suspicious DRE. Usual care sometimes included opportunistic screening when a test was requested by a participant or recommended by a doctor. Follow-up was through December 31, 2009, or to 13 years from trial entry. Tumour stage was categorised according to the fifth edition of the *American Joint Committee on Cancer (AJCC) Cancer Staging Manual*. Gleason grade was determined using the biopsy Gleason score (range 2–10); high-grade cancer was defined as a Gleason score of 8–10, and non–high-grade cancer as a Gleason score of 2–7. Subjects completed a baseline questionnaire near the time of enrolment, which enquired about demographics, medical history, and past screening practices.

The primary analysis was an intention-to-screen comparison of prostate cancer–specific mortality rates between the 2 trial arms. Event rates were defined as the ratio of the number of events (deaths or diagnoses) in a given time period to the person-years at risk for the event; measured from randomisation to the date of diagnosis, death or date of censoring (whichever came first) for incidence rates. Incidence rate ratios (IRs) were derived as the ratio of event rates in the 2 arms.

also apply to studies with more than 2 ‘arms’ (Freidlin *et al*, 2008). The latter may require a more complex design, and a larger sample size, although some efficiency in this area can be obtained through the use of factorial designs (see Section 11.7.2). The 2 groups might be a comparison of an intervention with a placebo, with no treatment, with the usual treatment, or with a different dose of the same product. The trial can be active (concomitant groups) or historical for the control group. An introduction to the design of these trials is available (D’Agostino *et al*, 2003). Placebos are ideal when there is no established alternative intervention, and where possible, a placebo should be used in preference to ‘no treatment’. Recent results in psychology have questioned our assessment of the ‘placebo effect’ and its importance in medical research (Kirsch, 2009). Often, it is unethical to include a placebo or no-treatment group when a standard effective treatment is available. Furthermore, the decision as to whether to use a positive control (existing therapy) or negative control (placebo) might have profound effects on the subjects available for inclusion in a trial, as well as the results. As the ‘comparison’ treatment level is often the current standard treatment, a frequently used trial design is the non-inferiority trial—an active controlled trial to investigate whether a new intervention is at least not inferior to the existing best intervention (Laster *et al*, 2006; Siqueira *et al*, 2008). The main difference between an non-inferiority and the usual RCT design is that in the non-inferiority trial, the null hypothesis is that the new intervention differs from the standard intervention by at most an amount (δ) that is not clinically important and in that sense it is not inferior. The challenge is to design a trial with sufficient power to reject the null if it is not true; a key issue in this is determining the appropriate value of δ.

11.3 PARTICIPANTS: THE STUDY GROUP

When designing a trial, you should be able to specify the target population and the source population. The target population is the one to which you want the results of the trial to apply (see Chapter 2). It is useful to state this explicitly as it can help ensure that the results are practical and applicable (Ahmad *et al*, 2009). The source population should be representative of

Example 11.2 A sequential trial of creatine in amyotrophic lateral sclerosis (ALS) survival

From June 2000 onward, ALS patients visited the outpatient clinic for neuromuscular diseases at the University Medical Center in Utrecht, and the Academic Medical Center of the University of Amsterdam, where they were neurologically examined before entry into the study (Groeneveld *et al*, 2003). The outcome of interest was survival. Sample size estimates indicated that at least 190 patients would be needed, assuming a cumulative survival percentage in the placebo group of 60% after 16 months and a 20% difference in the cumulative survival percentages.

Patients (n=175) were allocated such that the 2 treatment groups were evenly balanced for the prognostic factors age, site of symptom onset (bulbar, spinal), percentage predicted vital capacity (VC %) at onset and maximum voluntary isometric contraction (MVIC) arm strength at onset. The 2 treatment groups were creatine monohydrate and placebo (designated A and B).

An independent physician, ignorant to treatment assignment, instructed the research pharmacist to prepare trial medication A or B. Patients were seen 1 and 2 months after inclusion, and subsequently every 4 months until the end of the study at 16 months.

Reasons for withdrawal included serious adverse events and withdrawal of patient consent. Patients who stopped taking trial medication were not excluded from the intention-to-treat analysis.

Analysis was conducted by an independent biostatistician. The analysis was repeated when each new set of data based on the occurrence of an event using a sequential version of the log-rank test. The estimate for the hazard ratio was adjusted for the fact that the data were analysed sequentially.

the target population and represents the subjects who are eligible for the trial. The settings and location of the source population should be described. The study group is the collection of subjects who fit the inclusion/exclusion criteria and agree to participate in the trial. If the study group is not randomly obtained from the source population, it should be representative of it. Usually, the study group is obtained by seeking volunteer participants either by contacting them directly (*eg* personally, via letter, or the media) or by asking practitioners to nominate some of their clients who meet the eligibility criteria. While the use of volunteers is unavoidable, how well the study group (participants or study subjects) represent the source and target populations (see Section 2.1.3) must be taken into consideration when extrapolating the study results.

The period during which potential participants will be recruited should be stated (and any deviations from this explained).

11.3.1 Unit of concern

When stating the objective of the RCT, an early issue is to decide the level of organisation at which the intervention will be applied (*eg* individuals (Example 11.2), or groups/schools/communities (Example 11.3)). This also relates to the design of the trial because, if an intervention can only be applied at a group level and the outcome is measured at the group level, it is a group-level study. If the outcome is measured at the individual level, it is a cluster randomised study—Section 11.7.2). McRae *et al* (2011) list the characteristics that help define who the research subjects are in clustered randomised trials. They include those intervened upon by researchers, either directly or by deliberate manipulation of their environment; those who interact with researchers to provide data; or those who provide identifiable private information. They have created a website to discuss this and other ethical

Example 11.3 A cluster randomised controlled trial of an adolescent smoking cessation interventions

This was a randomised controlled trial of an adolescent smoking cessation intervention for students aged 15−21 at 22 continuation schools in Denmark (Dalum *et al*, 2012). Randomisation of intervention was done by flipping a coin. In order to minimise the effects of geographical and school-type differences the randomisation process was blocked in such a way that each county contained both intervention and control schools, and that both the intervention and the control group consisted of nine commercial and 2 social and health schools. At each school a coordinator was responsible for collecting data from students. Self-administered questionnaires were collected in week 5/2005 (baseline), week 11/2005 (first follow-up, short term) and week 11/2006 (second follow-up, long term). The intervention was based on open events where all smokers and ex-smokers were offered the following: (1) carbon monoxide measurement; (2) a personal and short counselling on motivation for quitting; (3) self-help materials with a combination of quit-guides; (4) referral to a smoking cessation programme, and (5) referral to telephone counselling.

Assessment of smoking status was based on self-reports. Short- and long-term quitters were defined as students who were daily smokers at baseline and had stopped smoking within the last 30 days at first follow-up (short term) and/or had been smoke-free for at least the last 30 days at the 2nd follow up (long term).

Analyses were performed as intention to treat and therefore all baseline smokers were included at both intervention and control schools when determining intervention effects. Ordinary logistic regression was used in the analysis of results with school as a random factor.

issues of controlled cluster trials (http://crtethics.wikispaces.com) (see also Weijer *et al*, 2011).

11.3.2 Eligibility criteria

- Adequate records should be available to document the subject's past history
- For trials of therapeutic agents, clear case definitions for the disease being treated must be developed to determine which cases are eligible for inclusion
- For trials of prophylactic agents, healthy subjects are required and procedures for documenting their health status at the start of the trial might be required
- Subjects in a trial need to be able to benefit from the intervention. Restriction of a trial to subjects that are most likely to benefit will increase the power of the trial but might limit the generalisability of the results
- Avoid subjects with high risks for adverse effects.

Eligibility criteria, must be stated clearly, and applied to all study subjects. A narrow set of eligibility criteria likely will result in a more homogenous response to the intervention and this might increase the statistical power of the study, but it might reduce the generalisability of the results. A broad set of eligibility criteria will result in a much larger pool of potential participants, but there could be a large background variation in study subjects (this can be advantageous for detecting variation in response to the intervention in subgroups of subjects), but can have negative effects on the power of the trial. Balancing these 2 considerations must be done case by case, while adhering to the objectives of the study. In general, we suggest using eligibility criteria that reflect the breadth of subjects who might receive the intervention in the future if it is shown to be effective (Zwarenstein *et al*, 2006).

Once the participants have been selected, it is important to collect and document the baseline characteristics of the study group that are germane to the trial. These data will be of considerable benefit when analysing and interpreting the trial results. The study period, especially the last date during which follow-up of study subjects will occur, needs to be specified also.

11.4 Specifying the intervention

The nature of the intervention, and how it is administered or implemented, must be clearly defined. They can vary from medical interventions (Examples 11.2, 11.5), surgical techniques (Example 11.4), type of lens (Example 11.6), to a screening programme (Example 11.1), or a smoking cessation programme (Example 11.3). A fixed intervention (one with no flexibility) is appropriate for assessing new products (particularly in phase III trials). A more flexible protocol might be appropriate for products that have been in use for some time and for which a body of clinically applied information exists. When possible, the initial treatment assignment should remain masked so that clinical decisions are not influenced by knowledge of group allocation. Clear instructions about how the intervention needs to be administered, or implemented, are essential, particularly if participants are going to be responsible for some or all of the interventions (*eg* instructions for how and when to take medication). In addition, a system of monitoring the intervention process should be put in place.

Example 11.4 A 2X2X2 factorial randomised trial of 3 Caesarean-section surgical techniques

In this trial, women >15 years of age were eligible if they were undergoing delivery by their first Caesarean section (Caesar_Study_Collaborative_Group, 2010). A telephone randomisation service was employed to allocate the interventions using a minimisation algorithm to ensure comparability between women with respect to 3 prognostic factors: participating centre; in labour or not in labour; single or multiple pregnancy. The 3 treatments were single- versus double-layer uterine closure; closure of the peritoneum, and liberal versus restricted use of a subsheath drain.

The primary outcome was maternal infectious morbidity, defined as having one or more of the following: (i) antibiotic use for maternal febrile morbidity during the postnatal hospital stay; (ii) endometritis; (iii) wound infection treated with antibiotics. Secondary outcomes included the individual components of the primary outcome

The estimated sample size was 3,500 women to demonstrate a change in the incidence of the primary outcome between any pair of arms from 12% to 9%, with 80% power and a 2-sided significance level of 5%.

Data were collected from the hospital notes and women were sent a questionnaire to complete 6 weeks after the date of Caesarean section. In this questionnaire, they were asked to record whether antibiotics or additional painkillers had been prescribed during the postnatal period and, if so, to give the reason. Patients were analysed in the groups to which they were assigned, regardless of deviation from the protocol or treatment received. Statistical analysis entailed the calculation of the risk ratios (*RR*) plus the 95% confidence interval (CI) for the primary outcome and 99% CI for the secondary outcomes to take account of multiple comparisons. Pairwise interactions between the different interventions were examined.

11.5 Measuring the outcome

A controlled trial should be limited to 1 or 2 primary outcomes (*eg* disease occurrence in a trial of a prophylactic agent) and a small number (1–3) of secondary outcomes (*eg* longevity). Having too many outcomes can lead to a problem of 'multiple comparisons' in the analysis (see Section 11.9.1). If multiple outcomes are measured, the intervention could have a different effect on each outcome. Whether or not to combine multiple outcome events into a single composite measure (*eg* a global measure of health by combining scores or occurrences of several diseases) has been the subject of much debate (Ferreira-Gonzalez *et al*, 2007) but for our purposes, we prefer designs based on a limited number of primary and secondary hypotheses.

The outcome in most medical trials is measured on a dichotomous scale, but continuous outcomes are common also. Time to event measures are very common; thus, it is important that the actual time when disease occurred should be as accurate as possible. Accuracy of measurement should be stressed; however, Korn *et al* (2010a) note that non-differential errors in the time of event usually do not have a major impact on the assessment of treatment effect. When selecting outcomes, those that can be assessed objectively are preferred to subjective outcomes, but the latter cannot always be avoided (*eg* occurrence of self-reported clinical disease). If the outcome is not assessed by a 'near-gold-standard' procedure, the impact of the intervention on the true outcome may differ from the surrogate outcome (Kassai *et al*, 2005).

In general, outcomes should be clinically relevant. Outcomes might also be measured at a single point in time, or assessed multiple times for each subject (longitudinal data). Intermediate outcomes, (*eg* antibody titres in a vaccine trial) might be useful in determining why an intervention might or might not produce the desired outcome, but should not be a replacement for a primary, clinically relevant outcome related to the objectives of the study (*eg* occurrence of clinical disease). Clinically relevant outcomes include the following:

- diagnosis of a particular disease—requires a clear case definition
- mortality—objective but still requires criteria to determine the cause (if relevant) and time of death
- clinical signs scores for assessing the severity of disease—difficult to develop reliable scales
- objective measures of clinical disease—(*eg* rectal temperature, blood samples to assess the extent of dehydration, *etc*)
- measures of health—(*eg* quality of life).

Example 11.5 A split-plot design randomised controlled trial of managing acute shoulder pain in primary healthcare

This pragmatic split-plot, randomised trial was conducted in general practices in 5 centres across the United Kingdom (Watson *et al*, 2008). Physicians in 91 practices (whole plot) were randomised to receive additional training in diagnosing and injecting rotator-cuff problems or no additional training; 215 patients with acute shoulder pain were then randomised to receive either a corticosteroid or lignocaine injection (split plot). The main outcome was score on the British Shoulder Disability Questionnaire (BSDQ).

Example 11.6 A multicentre cross-over randomised controlled trial to compare the effectiveness of 2 generations of progressive lenses for presbyopia

This controlled trial was performed in a primary-care setting (5 optical dispensaries) (Boutron *et al*, 2008). Two categories of progressive lenses were compared: a new-generation lens and an older-generation lens (which was considered to be the reference lens.)

The *a priori* estimated sample size was 68 patients. The authors thought that about 40% patients would not provide informative data (*ie* would have no preference) and a rate of lost to follow-up of about 15%, so they aimed for a sample of 130 patients. A total of 127 patients was randomised to wear one generation of progressive lens for 4 weeks, then cross over to wear the other lens for 4 weeks, without knowing the sequence of lenses. Inclusion criteria were age 43–60 years wearing progressive lenses with a correction of ≤3 dioptres. Patients and the statistical analyst were blinded and all equipment was assembled in one laboratory. The primary outcome was patient preference for one progressive lens at week 8. Secondary outcomes were subjective measures of bifocal visual performance.

11.6 Sample size

11.6.1 Allocating interventions to individuals

We begin this discussion by assuming that we are designing a trial with a fixed sample size—the most common approach used for clinical trials. The size of the study is determined through appropriate sample size calculations (see Chapter 2), with attention paid to the estimated effect of the intervention and both Type I and Type II errors (Schulz and Grimes, 2005a). The magnitude of the effect to be detected (or estimated) should be clinically meaningful. When computing the power of the study (1-Type II error), it is common to set the power to 90%. The sample sizes do not need to be equal in both arms of the trial (Schulz and Grimes, 2002c).

The sample size required for qualitative (*eg* dichotomous) outcomes is often much larger than that required for outcomes measured on a continuous scale. Obviously, the choice of outcome and its measurement should reflect the study objectives. The basic formulae for sample-size calculation, where the individual subject is the unit of randomisation, and the outcome is either binary or continuous, are presented in Chapter 2. Here we mention a few important issues that impact on sample size. Auleley *et al* (2004) discuss planning the sample size and how it is impacted by choice of the outcome measure(s), the scale of the outcome (*ie* continuous, binary, or time to event), and the occurrence of missing values. Barthel *et al* (2006) also discuss sample size issues when the outcome is time to event (*ie* survival), and provide a very flexible computer program for performing the calculations required for planning sample size in complex designs that allows for adjustments for missing data, non-proportional hazards and censoring (see Chapter 19). If there are multiple outcomes of approximately equal merit, using a generalised approach to power based on the probability that important changes in all outcomes will be observed has been described by Borm *et al* (2007b). Korn and Freidlin (2006) update the approach to determining sample size if historical controls will be used.

11.6.2 Sample size for the allocation of clusters of subjects

Cluster randomised trials are those in which all subjects within a group (*eg* all members of a family) are allocated to the same intervention (see Section 11.7.2 and Example 11.3). In planning such trials, we need to account for the intra-cluster correlation coefficient (ρ=*ICC*) and the cluster size (*m*) (see Chapter 20). As noted in Section 2.11.6, the sample size for a study needs to be increased by a factor of $(1 + \rho(m\text{-}1))$, so even if ρ is small, the overall sample size can become very large if the cluster size is large. With respect to cluster size, it has been shown that the power of a study does not increase appreciably once the number of individuals within each cluster exceeds $1/\rho$ (Campbell *et al*, 2007). Giraudeau *et al* (2008) discuss sample size estimation for continuous outcomes in cluster allocated cross-over designs. Teerenstra *et al* (2008) describe sample size estimation for 3-level cluster randomisation and elaborate on this for proposed GEE analyses (Teerenstra *et al*, 2010). de Hoop *et al* (2012) describe a formal method known as best balance to allocate treatment to clusters; it is particularly effective when the number of clusters is small.

When trying to decide on the unit (*ie* individual or group) to randomise, as the ICC increases, random allocation of individuals becomes much more efficient statistically, especially if the cluster size is large. However, if the intervention is allocated to clusters, and the number of clusters available is small, a matched design (*eg* matching on strong cluster-level confounders) may be used. When feasible, employing a cluster cross-over design can add to the efficiency particularly when the number of clusters is small (Turner *et al*, 2007).

11.6.3 Sample size for sequential and adaptive designs

A **sequential design** trial (also called a 'monitored' study) incorporates 'a method allowing hypothesis tests to be conducted on a number of occasions as data accumulate through the course of a trial' (Todd, 2007). Thus, the sample size is not fixed in advance of the trial, rather sequential designs have specified stopping rules. Typically, the planning of these studies is more complex than the fixed trial design, and there is the potential for bias in that the researchers might alter the implementation of the trial after learning the results of the interim analyses. Zou *et al* (2005) describe sequential methods for cluster randomisation. Lavori and Dawson (2007) describe how to improve the efficiency of adaptive designs with emphasis on variance estimates using G-computations (not discussed in this text). Subsequently, Dawson and Lavori (2008) developed a sequential method for making causal inferences.

Barthel *et al* (2009) describe multistage trials that contain an explicit interim analysis and allow for stopping the trial if lack of benefit is evident. If the trial is not stopped, more recruitment of study subjects proceeds. Bassler *et al* (2008; 2010) caution against stopping trials early because of interim evidence that the intervention works; these 'stopped' trials tend to overestimate the efficacy of the intervention (Schulz and Grimes, 2005b); although Freidlin and Korn (2009) and Korn *et al* (2010b) indicate that the bias is not large. As a contrast, Royston *et al* (2011) developed a multiarm adaptive trial with numerous points of interim analysis with the intent of stopping the trial for lack of benefit at the earliest point possible.

Adaptive design studies are ones in which the design may change as the study progresses. Consequently, they are more flexible than sequential designs (Golub, 2006). The most common 'adaptation' is modification of the sample size of the second stage based on the predicted power of the trial at the end of the first stage. However, adaptive designs also include dropping or adding treatment arms, changing the primary endpoint, and even changing objectives (for example, switching from non-inferiority to superiority (Todd, 2007). Outcome adaptive designs strive to ensure that the majority of subjects get the benefit of the best therapy available. The allocation of subjects is influenced by the experience of previous subjects in the trial. One example is ‘play the winner’ allocation in which subjects continue to be allocated to an intervention level as long as that treatment is producing beneficial results. As soon as it fails, the allocation switches to the other treatment. These procedures are only suitable if the result of the intervention is clearly identifiable in a very short period after treatment.

11.6.4 Other sample size issues

Another issue to consider when planning the size of the trial is the time to recruit study subjects (2 days in Example 11.3; 5 months in Example 11.6). The length of time it will take to recruit the required number of study subjects can be a serious problem for studies on therapies for relatively rare conditions. If season of treatment is likely to influence the results, then the recruitment period should span at least one full calendar year.

Loss of subjects from the study might happen for a variety of reasons and can lead to bias. Some subjects might be lost to follow up (*eg* subject moved away) while others might be non-compliers (*eg* participants who do not comply with the protocol). Finally, some subjects might be lost due to competing risks (*eg* die from other diseases while still on the trial). Once a sample size has been estimated, it is wise to compute the expected power of the study based on different estimates of the potential losses to the study, and adjust the sample size accordingly.

11.7 Allocation of study subjects

It is important to remember that controlled trials are based on volunteer subjects and participants must agree to receive either of the interventions (treatment versus comparison) as determined by the allocation process. Once a subject has been enrolled, the allocation should be carried out close to the time at which their participation in the study is scheduled to start.

It is clear that a formal randomisation process is the best method for allocating subjects to study groups; indeed, without this formal allocation procedure, bias is very likely to distort the findings (Gluud, 2006). The use of propensity scores (Section 13.8) is one potential approach to dealing with the problem of inadequate randomisation. However, propensity scores are generally used in observational studies of interventions in situations in which a randomised control trial is not feasible. Before proceeding with formal randomisation procedures we will discuss some alternatives.

11.7.1 Alternatives to randomisation

Historical control trials are ones in which the outcome after an intervention is compared with the level of the outcome before the intervention (before/after comparison). For a historical control trial to have any validity, 4 criteria must be met.

1. The outcome being measured must be predictable,
2. There must be complete and accurate databases on the disease of interest,
3. There must be constant and specific diagnostic criteria for the outcome, and
4. There must be no changes in the environment of the subjects in the study.

Rarely are all of these criteria met. An additional limitation of historical control trials is that it is impossible to use blinding techniques.

Systematic assignment of individuals to treatment groups (*eg* alternating assignment) can be a reasonable alternative to formal randomisation under field conditions (people arriving at a vaccine clinic where half will get the vaccine and half the placebo). Systematic assignment might make it harder to keep participants and study personnel blind as to the intervention identity, but aside from this, it is often just as effective as random allocation (provided outcome assessment is done blindly). If half the subjects are to be allocated to receive the treatment, the initial subject allocation should be random and thereafter, every second subject would receive the allocated intervention. Do not apply the intervention to the first (or last) half of the subjects and the comparison treatment to the remainder. When feasible, true randomisation is preferred.

11.7.2 Random allocation

As indicated, formal randomisation is the preferred method of allocation. It must be noted that random allocation does not mean 'haphazard' allocation and a formal process for generating random intervention assignments (*eg* computer-based random-number generator, or even a coin toss—Example 11.3) must be employed. Random allocation should be carried out as close as possible to the start of the study to reduce the possibility of withdrawals after allocation.

Simple randomisation (*ie* a completely randomised design) involves each subject being assigned to an intervention level (*eg* vaccine or not, treated or not) through a simple random process without any further considerations, this is the most frequently used allocation process in field trials (Example 11.1). Stratified randomisation (*eg* randomisation within age categories) helps ensure that a potential confounder (age) is equally distributed across study groups. Hofmeijer *et al* (2008) propose a method of block allocation to gain efficiency in small trials. It adjusts the assignment of the next subject depending on the imbalance in treatment allocation that exists at the time.

Cross-over studies
In a cross-over study, each subject gets both of the interventions (in sequence) (Example 11.4). However, the first intervention administered is still assigned randomly. This process is only suitable for the evaluation of therapies where the condition of the subject is stable, and the duration of the intervention effect is relatively short-lived. A 'wash-out' period might be required between interventions. It has the advantage that it increases the power of the study since the same subject receives both levels of the intervention.

Factorial designs
This design is particularly well-suited to trials investigating 2 or more interventions, especially if the interventions might produce synergism or antagonism (Example 11.5). Here, all possible combinations of the treatments (*eg* neither treatment, treatment 1 only, treatment 2 only, both treatments) are assigned to the study subjects. Because the design is usually balanced, the treatment effects are not confounded (*ie* they are unrelated, or orthogonal, to the intervention) and the analyses are straightforward. Normally, one should not attempt to assess more than 2–3 interventions as the possible interactions become difficult to interpret.

Cluster randomisation
There are a number of reasons why a cluster of people should be allocated to an intervention group rather than individual people. In some instances, it might be the only feasible method. For example, if the intervention is one which is always given at the group level (*eg* medication in the drinking water), then there is no choice. Even if the intervention could be administered at the individual level, it might be impossible to keep track of individuals within the group, or the

intervention in some subjects could influence events (*eg* through spread of a live vaccine) in non-intervention subjects housed with them (see Section 11.11) so assignment of the whole group to one intervention would be appropriate (Example 11.3). Cluster randomisation is also appropriate if there is potential for physical spread of a treatment to the control group or the potential for the effects of the intervention to impact the non-intervention groups as in herd immunity (see Example 11.7). Recent developments in the design and analysis of cluster randomised trials have been reviewed (Campbell *et al*, 2007).

Cluster randomised trials are much less statistically efficient than trials with random allocation of individuals and the clustering of individual subjects within the groups needs to be taken into account in analysis (see Chapters 20–23). In a cluster randomised trial, the best scenario for follow-up is if all individuals can be monitored for the duration of the study. If this is not possible, following a randomly selected cohort would be the most statistically powerful approach. If it is not possible to follow individuals, the investigator will have to carry out repeated cross-sectional samplings throughout the follow-up period (Campbell *et al*, 2007). Donner and Klar (2004) review the advantages and pitfalls of using cluster randomisation; Donner *et al* (2007) also comment on 'breaking the matches' to gain some statistical efficiency in the analysis of matched-cluster randomised trials.

Split-plot designs

A final elaboration of allocation discussed here is a split-plot design. This design is used if there are 2 or more interventions, 1 of which needs to be applied at the group level and the other(s) assigned to individuals (Example 11.6). The analysis must take account of the different degrees of freedom to assess intervention effects at the different levels (*ie* group versus individual).

Multicentre trials

If an adequate number of subjects is not available at a single site, a multicentre trial might have to be planned (Fedorov and Jones, 2005) (see Example 11.6). A key feature is that the within-centre and between-centre variances need to be accounted for in the design and analysis. Although a multicentre trial complicates the protocol and the implementation of the trial, it can enhance the generalisability of the results (because of the usually larger geographic area covered by the trial) and also increases the opportunity to identify interaction effects (*eg* different responses by centre). One key for statistical efficiency in multicentre trials is to try and maintain approximately the same number of subjects per centre (Dragalin and Fedorov, 2006).

11.7.3 Masking (blinding)

'Blinding' (or 'masking') in randomised trials refers both to the general methodological principle of withholding information from individuals with the aim of preventing bias, and to a group of procedures used to withhold information from specific groups of individuals, *eg* patients, treatment providers, and data analysts) (Hrobjartsson and Boutron, 2011). Unfortunately, the usage of the terms single, double, and triple blinding is not consistent. Often the terms to describe who is blinded are ambiguous, and the specific mechanisms for masking need to be described. In large trials, it is beneficial to pilot test the masking procedures. For our purposes, a single-blind study means that the participant/patient is unaware of the identity of the intervention applied to them. This should help reduce response bias, prevent the placebo effect; prevent differential attrition and non-compliance, and reduce co-intervention bias and follow-up bias. A double-blind study means that both the participants and selected members of the study team (*ie* people administering the interventions and those assessing the outcomes) are unaware of intervention assignment. Keeping the hands-on members of the study team blind should

Example 11.7 Immunity conferred by killed oral cholera vaccines in Bangladesh

This individually randomised, placebo-controlled trial evaluated the effectiveness of killed oral cholera vaccines (Ali *et al*, 2005; Hudgens and Halloran, 2008). The interest was in determining whether the level of vaccine coverage in a residential area, called a bari, was related to the incidence of cholera in individual vaccine recipients or placebo recipients residing in the bari. The target population was divided into groups by level of vaccine coverage. Hudgens and Halloran used the data from 2 groups; Group A with more than 50% and Group B with less than 28% coverage as their example. The effects of vaccination are estimated based on differences in the risk of cholera during the first year of follow-up of the trial. The pertinent data are:

- R_{nvB} = 7.01 (risk per 1,000 in non-vaccinates in population B)
- R_{vB} = 2.66 (risk per 1,000 in vaccinates in population B)
- R_{nvA} = 1.47 (risk per 1,000 in non-vaccinates in population A)
- R_{vA} = 1.27 (risk per 1,000 in vaccinates population A)
- R_B = 4.13 (overall risk/1,000 in population B)
- R_A = 1.34 (overall risk/1,000 in population A).

The direct effects are estimated by comparing the risk between vaccinated individuals and unvaccinated individuals within each group. The estimated relative direct effect in group A (the more highly vaccinated group) was:

$$VE_d = \frac{R_{nvA} - R_{vA}}{R_{nvA}} = \frac{1.47 - 1.27}{1.47} = 0.14$$

The relative direct effectiveness of vaccination within group B was:

$$VE_d = \frac{R_{nvB} - R_{vB}}{R_{nvB}} = \frac{7.01 - 2.66}{7.01} = 0.62$$

which was higher than in group A. The difference in the 2 estimates illustrates one of the challenges in making comparisons directly within groups when indirect effects are present. If an analysis (or indeed a full trial) was limited to group A only, the evidence would suggest that the vaccine has little effect. In group B, the vaccine lowered the risk of disease by 62%.

The **indirect effect** of vaccination is the effect due to the level of coverage. It can be estimated by comparing the risks in the unvaccinated in the 2 groups or the risks in the vaccinated in the 2 groups. The estimated indirect effect in the unvaccinated is (7.01-1.47)/7.01=0.79 in relative terms. In the vaccinated the indirect effect is (2.66-1.27)/2.66=0.52. Note that the estimate in the unvaccinated is greater than the estimated direct effect in either of the groups, highlighting the importance of looking beyond direct effects.

The relative **total effect** of vaccination is the effect of not being vaccinated in the group with lower coverage (B) compared with being vaccinated in the group with higher coverage (A). In this instance, the estimated relative total effect is (7.01-1.27)/7.01=0.82. Note the total effect {in absolute terms (B-A)} estimate equals the direct effect estimate in group A plus the indirect effect estimate in the unvaccinated (B-A).

The relative **overall effect** is the average effect of being in the group with higher coverage compared with being in the group with lower coverage. It can be estimated using the difference in risk between the 2 groups, that is, (4.13-1.34)/4.13=0.68 in relative terms. In this instance the vaccine reduced the frequency of disease in group A by 68% relative to the level in group B.

reduce response bias and placebo effect (impact of patient–provider interaction); equalise attrition, non-compliance and co-intervention as well as preventing selective decisions; and referral that could arise if the intervention status was known. Keeping those responsible for assessing the outcome blind should reduce observer and diagnostic bias. In a triple-blind study, those who are analysing the data also are unaware as to which group received which treatment. This feature is designed to ensure that the analysis is conducted in an unbiased manner. Of course it is advisable to design the trial so that as many errors as possible are prevented and this can reduce the impact of any differential errors that arise. It is recommended that the success of blinding be evaluated and not taken for granted (Boutron *et al*, 2007; Hrobjartsson *et al*, 2007).

In many cases it is necessary to use a placebo to ensure that the relevant individuals remain blind. A placebo is a product that is indistinguishable from the product being evaluated and which is administered to people in the groups designated to receive the comparison treatment. In many drug trials, the placebo is simply the vehicle used for the drug, but without any active ingredient (see Example 11.2). One concern with the use of a placebo is that, even though it might not contain the active ingredient being investigated, it could still have either a positive or negative effect on the study subjects. For example, a placebo vaccine that does not contain the antigen of interest might still induce some immunity as a result of adjuvant in the placebo. These issues should be discussed and settled prior to conducting the trial.

In some cases, using a placebo might not be adequate to ensure blinding. Nonetheless, masking the intervention should be used whenever possible.

11.8 Follow-up/compliance

The practical issues involved in managing and conducting a controlled trial have been well described by Knatterud (2002). One important item is to ensure that all groups are followed rigorously and equally. This is a simpler process if the observation period following the intervention is short, but this time period must be long enough to ensure that all outcomes of interest have been observed and recorded. Regardless of the effort expended on follow-up, it is inevitable that some individuals will be lost to the study through drop-out or lack of compliance. Thus, for studies with long follow-up periods, the status of all study subjects should be ascertained at regular intervals throughout the follow-up period. The CONSORT statement suggests using a flow diagram to outline participant numbers at key points (allocation, intended intervention, completed the protocol, and had outcome status evaluated) in the trial (see Section 11.12).

A major factor in minimising losses from the study is regular communication with all participants. Incentives to remain in the study might also be provided. These might include provision of information which the participants might not otherwise have or public recognition of their efforts (provided confidentiality concerns have been addressed). For those participants that do drop out, information about study subjects might still be available through routine databases if the participant is willing to provide access. This can be used to either provide some follow-up information or to compare general characteristics of the study subjects withdrawn from the study with those that remained in the study. Nonetheless, because participants in a trial should always have the opportunity to withdraw from a trial, procedures for evaluating those withdrawals should be put in place. This should include methods of documenting the reason for the withdrawal and, potentially, procedures to collect samples from all subjects being withdrawn before their departure. In any event, any losses should be recorded, at specified time points, throughout the conduct of the trial.

Effort needs to be expended, in addition to maximising retention in a study, to determine if study subjects are complying with the protocol. This might be evaluated through interviews at periodic visits or through collection of samples to test for levels of the drug being investigated. Indirect assessment might be carried out by methods such as collecting all empty containers (pill boxes, vials, *etc*) from products used in a trial.

11.9 Statistical methods and analysis

As noted, outcomes might be measured and then analysed on a continuous scale, or as categorical data (often dichotomous), or as time-to-event measurements (*eg* time to the occurrence of a disease). Studies based on time-to-event data might have greater power than a study based on simple occurrence—or not—of an event in a defined time period. Regardless of the analysis conducted, results reported should include both the effect size and its precision. If the outcome is dichotomous, the results should be in both absolute (*eg* risk difference) and relative terms (*eg* risk ratio).

Analysis can be carried out either on an **intent-to-treat** (ITT) basis or a **per-protocol** basis. In an intent-to-treat analysis, data from all subjects assigned to a specific intervention are included in that intervention regardless of whether they completed the study, or complied with the protocol. Such an analysis will provide a conservative estimate of the effect of the intervention, as it is recommended to be used. However, the ITT result reflects the expected response when the intervention is used in another population with characteristics similar to the study population. In a per-protocol analysis, only subjects which complied and completed the study as outlined in the protocol are included in the analysis. This approach might provide a good measure of response given that the intervention is used as intended but will likely produce a biased estimate of the intervention effect in future use for 2 reasons. First, non-compliance is not likely a random event and non-compliers probably are not representative of all participants assigned to that intervention so the estimate of effect may be biased (see Section 12.2). Second, there will always be some non-compliance in future use of the intervention, so estimating an effect under an assumption of 100% compliance would be unwise. Stating the numbers of participants in the groups, and whether or not they complied is essential to interpreting the results. Hernan and Hernandez-Diaz (2011) suggest that if there are considerable losses from the trial or a lack of adherence to the study protocol, that analyses based on inverse probability weighting be used to reduce the potential bias.

An analysis usually starts with a baseline comparison of the characteristics of the groups as a check on the adequacy of the randomisation procedures. This should not be based on an assessment of the statistical significance of the difference among groups, but rather an assessment of their comparability. Differences among the groups, even if not statistically significant, should be noted and taken into consideration in the analyses (see below) (Austin *et al*, 2010). Hernandez *et al* (2004; 2006) suggest adjusting for predictive variables (*ie* including them in the analysis) whenever possible as it increases the power (or decreases the sample size for a given power).

The specific procedures for analysing data from controlled trials will not be covered in this chapter as most are discussed in more detail elsewhere in the book. However, a few specific issues will be touched on.

While randomisation is designed to equally distribute potentially confounding factors across the intervention groups, it might not remove all potential **confounding**, especially with small sample sizes (hence the rationale for examining this as noted above). When the outcome is dichotomous, adjustment for covariates is recommended. The best approach is to identify strong predictors *a priori*, the next best option is to control for covariates that are predictive of the outcome in the trial data (Hernandez *et al*, 2004). Adjusted results should be less biased if the adjustment procedure has removed any residual confounding (particularly a concern in small trials). Adjustment for non-confounders does little harm, provided they are not intervening variables (see Chapter 13). If the outcome is continuous, control of other factors might improve the precision of the estimate of the intervention effect by substantially reducing the unexplained variance.

When measurements are made **before and after** the intervention is administered, it is often useful to adjust for the baseline (pre-intervention) level in each subject when evaluating the response to the intervention. This can either be done by subtracting the pre-intervention value from each post-intervention measurement (*ie* analysing the change in the outcome) or by including the baseline level as a covariate in an analysis of the post-intervention values. Either approach will result in a gain in power for the study, particularly if the correlation between the baseline and the post-intervention measurement is >0.5 (Borm *et al*, 2007a).

Subgroup and ancillary analyses should be avoided unless there were specific plans for them in the original design. Brookes *et al* (2004) demonstrated that greatly increased sample sizes (>4-fold) are needed for reasonable power of interaction tests. Subgroup analyses based on seeing the data often lead to spurious conclusions.

Many controlled trials involve repeated assessments of subjects throughout the study period (**longitudinal data**). Analysis of longitudinal data presents some unique challenges. For a starting point the investigator needs to determine if they are most interested in an average effect following intervention, a change in the effect over time or a total effect. Methods of dealing with repeated measures data are covered in Chapter 23. Twisk and de Vente (2008) review methods for dealing with repeated measurements in RCTs. They suggest that if GEE (Chapters 20 and 23) or a similar approach is used for analysis and if the outcome is measured on a continuous scale, only the first follow-up should be adjusted for the baseline (pre-intervention) level of the outcome. Using their approach we would do the following:

- First, perform a linear regression analysis between the first follow-up measurement and the baseline value
- Second, calculate the difference between the observed value at the first follow-up measurement and the predicted value from that regression analysis. This difference is called the 'residual change'
- Third, use this 'residual change' in place of the actual first outcome value in the subsequent GEE analysis.

Longitudinal data often have **missing values** for some of the observations. The problem of missing data is briefly introduced in Section 15.5 and more detailed discussion of the issue can be found in Peduzzi *et al* (2002) and Auleley *et al* (2004). If more than a few observations are missing, the analysis and interpretation will have to take this into account.

Finally, if study subjects are maintained in groups (**clustered data**), it is important to account for the effects of the groups. This is particularly important in cluster randomised trials, but might also be important in trials in which randomisation occurred within the group. Procedures for analysing clustered data are presented in Chapters 20–22.

11.9.1 Multiple comparisons and assessments

Controlled trials often give rise to analyses in which **multiple comparisons** are made. There are 3 ways in which multiple comparisons can arise in the analysis of RCTs: examining multiple outcomes, examining multiple subsets of the data, and performing periodic interim analyses during the trial. The problem with multiple comparisons is that the experiment-wise error rate is often much larger than the error rate applied to each single analysis (usually 5%; see Section 15.8.2). This can result in the declaration of spurious effects as significant.

There are many procedures for adjusting the analyses to account for these multiple analyses (Korn and Freidlin, 2008). One of the simplest ways to retain an appropriate experiment-wise error rate is the **Bonferroni adjustment**. This requires that each analysis be carried out using an α/k Type I error rate, where α is the normal error rate (often 0.05) and k is the number of comparisons made. However, this results in a very conservative estimate of the statistical significance of each evaluation. Other, less conservative, procedures can be found in standard statistical texts.

The problem of **subgroup analyses** deserves special attention (Brookes *et al*, 2004). While it is tempting to evaluate a wide range of subgroups within a trial to determine if an intervention had an effect in them, only analyses planned *a priori*, should be carried out. Otherwise, there is serious danger of identifying spurious associations. Many researchers recommend that findings from unplanned subgroup analyses be reported as exploratory. Furthermore, the recommended approach to ascertain if the intervention effect differs by subgroup is to conduct one overall test of interaction between the intervention and the subgroup identifier. Bear in mind that the sample size of the study usually was based on a single overall test of significance not on a per-subgroup basis and in many instances subgroup analyses will have insufficient power to detect meaningful effects. Brookes *et al* (2004) also describe a method to determine the appropriate sample size required to investigate such interactions reliably. As a guideline, effects sizes of at least twice the magnitude of the assumed overall effect have a similar power of detection to that of the overall intervention effect.

Sequential design studies are those in which, by design, planned periodic analyses of the data are carried out throughout the trial (see Section 11.6.3). These analyses are carried out so the trial can be stopped if there is:

- clear (and statistically significant) evidence of the superiority of one intervention over another, or
- convincing evidence of harm arising from an intervention (regardless of the statistical significance of that finding), or
- little likelihood that the trial will produce evidence of an effect, even if carried to completion. (This concern is not relevant if the goal of a trial is to demonstrate that a new product/procedure has the same efficacy as an existing standard therapy.)

While sequential designs seems like a logical approach, they tend to lack power (on a per-subject basis), and hence their usage should be restricted to those situations where the benefits are clear. Interim analyses should not be conducted unless the trial is designed to accommodate them. Methods for interim analyses and for adjusting the sample size to accommodate the procedures are beyond the scope of this text but are reviewed in Todd (2007).

11.10 Conclusions

This section should include an overall interpretation of the study and its importance for the target population. Be careful when doing this as Boutron *et al* (2010) have noted that many authors tend to overinterpret the possible value of interventions despite the non-significant evaluation of the trial data. Any limitations such as sources of bias, and unintended impacts, should be explained. The registration number, sources of funding and the role (if any) of funding groups in the trial should be stated. Finally, if the full trial protocol can be accessed, how to do this should be specified.

11.11 Clinical trial designs for prophylaxis of communicable organisms

The standard designs discussed thus far need to be modified when the intervention is a prophylactic against a communicable organism (*eg* a vaccine). Here we will explain why this modification is needed and make some suggestions about trial designs. See Chapter 27 for a discussion of issues related to infectious disease epidemiology.

When estimating the 'protective ability' of a prophylactic against communicable organisms, we need to consider whether we are measuring protection at the individual or at the population level. Furthermore, we need to recognise that the protection we observe can be strongly influenced by:

- the baseline level of transmission of the agent in the population of interest,
- the effectiveness of the vaccine (this is of course what we want to estimate), and
- the percentage of the population we chose to vaccinate in our evaluation of the vaccination strategy.

In a population, disease-causing organisms spread from subject to subject, either directly or via vehicles contaminated with the organism of interest. The rate of transmission depends on the number of adequate contacts a susceptible subject makes with an infected subject or contaminated vehicle per time period (*eg* per day) and the level of susceptibility of the subjects that are contacted. (See Section 27.3 for a discussion of infectious disease transmission). The key difference between designing a trial for a therapeutic agent and a vaccine is that an effective vaccine has effects on the vaccinated as well as on the non-vaccinated; in other words, the study subjects are not independent. This effect is called **interference** by Hudgens and Halloran (2008). Given a reasonable limit to the number of contacts each susceptible subject makes per day, if some of these contacts are with vaccinated subjects, and if vaccinated individuals are completely or partially protected against infection, the rate of spread of the disease through the population is decreased. In general, the number of adequate contacts each individual makes and the baseline transmission level depend on the characteristics of the study groups. Consequently, "2 different randomised, double-blind, placebo-controlled studies taking place in sites that differ by the level of transmission would report different estimates of vaccine efficacy even if the level of individual (direct) protection conferred by the vaccine to a specified challenge to infection is the same in both studies" (Hudgens and Halloran, 2008; Struchiner and Halloran, 2007). In addition, in order to understand disease spread, it is helpful to know whether transmission of an agent within subunits of the population (*eg* family) is density or frequency dependent. In density-dependent transmission, disease transmission is the same among units of different sizes when the proportion of initially infected subjects is the same. In frequency-dependent transmission, transmission increases with the number of initially infected individuals (so larger families or groups would have greater transmission even if they had the same proportion initially infected).

Prophylaxis can have a number of benefits. First, it can prevent infection given exposure. Second, it can prevent clinical disease or reduce the severity of infection among the infected and this can lower the onward transmission of the agent. Whether infection or disease is the chosen endpoint often depends on the context and on the incubation period of the disease—if short, disease is often the endpoint; if long, infection is usually the endpoint. The ability to reduce the severity or duration of disease among those receiving the prophylactic may have a larger impact on the transmission probability in the population than the ability to protect against infection in individuals. The key is that the protective effect of prophylaxis can differ depending on the endpoint evaluated.

As an example, the usual measure of vaccine efficacy (*VE)* (for simplicity, we will not differentiate between infection vs disease as outcomes) at the individual level is typically measured as:

$$VE_d = \frac{(I_{nv} - I_v)}{I_v} \qquad \textbf{\textit{Eq 11.1}}$$

where I_{nv} and I_v are the incidence rates of the outcome in non-vaccinated and vaccinated individuals, respectively (Halloran, 2006). We have added the subscript 'd' to denote that this is the direct efficacy of the vaccine. Of course, to ascertain the true VE_d, we would like to compare counterfactuals (see Section 1.7); *eg*, the incidence of the outcome in the vaccinated subjects contrasted to what the incidence would have been if the subjects were non-vaccinated. Since we cannot observe these events, we estimate the *VE* by randomly assigning half (or some other proportion) of the study subjects to receive the vaccine and half to get a placebo; both vaccinated and non-vaccinated subjects are free to intermingle in the population. Unfortunately, the measure of *VE* we obtain from using this design is likely to be confounded by the proportion of the study population that is vaccinated. We will explain the rationale for this statement subsequently.

Because the direct *VE* measure is often biased and may only be a small proportion of the total efficacy, epidemiologists are more interested in population-based measures of vaccine effectiveness. The total effect of prophylaxis is a population measure and consists of 2 components: the direct or individual level vaccine efficacy (VE_d) noted above and the indirect (VE_{ind}) vaccine efficacy. The indirect vaccine efficacy is a population-based measure and is found by comparing the frequency of the outcome in non-vaccinated people from the randomised study area with the higher level of vaccination (here designated population A) to the frequency in non-vaccinated people from a similar population of non-vaccinated people in an area with lower (or no) level of vaccination (designated here as population B) as follows:

$$VE_{ind} = \frac{I_{nvB} - I_{nvA}}{I_{nvB}} \qquad \textbf{\textit{Eq 11.2}}$$

where I_{nvA} and I_{nvB} are the incidence rates (or risks) in populations A and B, respectively. This indirect effect often is a major component of what is referred to as **herd immunity**. The phenomenon of herd immunity provides protection to non-vaccinated, susceptible individuals by interfering with transmission of the agent beyond the direct protective effects in vaccinated individuals. For example, in Section 27.5.2, it is shown that if an infected person typically contacts 5 susceptible people, a vaccine that is 80% effective will be expected to stop all transmission of the agent (as a result of herd immunity). Since achieving 100% vaccination coverage can be very difficult and knowing that vaccination levels below 100% can be effective in eliminating disease agents, ascertaining the critical level of vaccination that is required to eliminate a specific disease agent (*eg* measles virus) is a key component of research on population disease control (Longini *et al*, 1998).

The overall effect of the prophylaxis (VE_{tot}) is a weighted combination of VE_d and VE_{ind} and can be estimated using:

$$VE_{tot} = \frac{I_B - I_A}{I_B} \qquad \textit{Eq 11.3}$$

Knowing the overall effect of a vaccine provides much more useful information in terms of disease control than does the usual direct measure of vaccine efficacy.

11.11.1 Design and analysis issues for estimating vaccine efficacy

A number of different trial designs can be used to obtain estimates of vaccine efficacy. For example, we can employ a cluster randomised trial design in which we compare the disease frequencies in fully vaccinated versus non-vaccinated populations. Riggs and Koopman (2005) developed a model of transmission with group randomisation, and they noted that if cluster randomisation is used, it increases the power of the study if the majority of transmission is from within the cluster, but decreases the power if most transmission comes from outside the cluster. They also note that, when using cluster randomisation, it is advantageous to sample study subjects and determine their natural level of immunity (*ie* prior to vaccination). This allows for the adjustment for natural immunity prior to assessing vaccine induced immunity. While this is perhaps the best approach to obtain valid estimates of *VE*, it becomes very expensive and it does not extend easily to the situation where natural, stable groupings of study subjects are not available. Nor does this approach reflect what might happen in populations where it is unlikely that 100% vaccine coverage will be obtained. Furthermore, because of the indirect effects of a vaccine, there is a critical level of vaccination, often considerably below 100%, that will protect the population and potentially lead to eradication of the organism. In order to estimate this, we would have to assign different levels of vaccination (say 25%, 50%, and 75%) to groups without exceeding the critical fraction vaccinated that would eliminate disease in the non-vaccinated subjects (see Hudgens and Halloran (2008); their data form the basis of Example 11.7. The theorems behind the calculations are beyond the scope of this test).

As noted, for disease control, the total effectiveness of the vaccine in the population is of more interest than VE_d. A suggested approach is to use a design that will allow the estimation of the direct, indirect, and total vaccine efficacies described above. To implement this, we need at least 2 comparable populations of subjects (more than 2 populations would provide much better estimates of vaccine efficacies but these could become prohibitively expensive). We would randomly assign a proportion of individuals in one population, denoted population A, to receive the vaccine and the remainder to receive a placebo. Subjects in the other similar population, denoted population B, would all remain non-vaccinated (Halloran, 2006), or be allocated to a lower level of vaccination (Hudgens and Halloran, 2008). Ensuring exchangeability (*ie* that the populations are similar in all important characteristics that affect the outcome) is a difficult task. Perhaps the most important characteristic they should share is the same level of transmission as this greatly affects the indirect efficacy. In addition, it is important that the 2 populations are fully separated from each other so there is no intermixing of subjects. The building blocks for the calculations are the outcome frequencies in each of 2 populations that have a different proportion of vaccinated subjects. Since the total effectiveness is a population level measure, several populations are needed for statistical evaluation. Nonetheless, this design is feasible under selected circumstances, and this concept is the basis for interpreting population effects of vaccines (Glezen, 2006). Glezen notes that although the direct effects may be small, the impact on population levels of disease can be very marked (see Example 11.7). If the disease frequency is judged to be stable, information on the level of disease in the non-vaccinated population can be supplemented with data on the level of disease (in population A) prior to the vaccine trial.

An alternative design that can be used when obtaining these 2 'similar' populations is difficult, is the natural clustering of subjects within the population of concern (*eg* children within schools). In this situation we would randomly assign vaccination to half the subjects (*ie* children) in the larger geographical area and then subsequently investigate the spread of infection/disease within the clusters (*ie* schools) noting the proportion of individuals in the cluster that were vaccinated (Longini, *et al*, 1998). However, one would need to ensure some stability to the population of these subsets over the duration of the prophylactic trial. Glezen *et al* (2010) investigated the direct and indirect effects of vaccination against influenza. They noted that the direct effect was impacted by the percentage vaccinated. The indirect effect was based on patients presenting to clinics with no history of current influenza vaccination; 51.1% were culture positive at the intervention clinics, compared with 55.7% at the comparison clinics (*IR*=0.92; difference not significant). The overall efficacy was 1-*IR*=1-0.89=0.11 or 11%. Longini *et al* (2002) describe how to assess vaccine efficacy when communities (or schools) are assigned to receive different levels of vaccination; the method allows researchers to estimate the effectiveness of vaccination at levels that differ from those used in the trial.

Detailed consideration of power is discussed by Riggs and Koopman (2005) but is beyond the scope of this text. Kong *et al* (2006) describes the design of vaccine non-inferiority trials.

11.11.2 Estimation of vaccine effects on post-infection outcomes

Hudgens and Halloran (2006) note that "The effects of vaccine on post-infection outcomes, such as disease, death, and secondary transmission to others, are important scientific and public-health aspects of prophylactic vaccination." Thus, it might seem straightforward to estimate the efficacy against severe infection or against transmission but as they point out, it isn't. Evaluation of post-infection vaccine effects must condition on being infected. Because the set of individuals who would become infected if vaccinated is likely not identical to those who would become infected if given the control, comparisons that condition on infection do not have a causal interpretation. Thus these authors propose selection models that identify the causal estimand and closed-form maximum likelihood estimators (MLEs) are derived under these models.

11.12 Reporting of clinical trials

Poor quality of reporting of trials remains a problem (Berwanger *et al*, 2009), although the standard of reporting has improved following the release of the CONSORT statement (Kane *et al*, 2007). Recently, the CONSORT statements were modified (Moher *et al*, 2010; Schulz *et al*, 2010). These reporting standards should serve as guidelines to help ensure that critical issues in study design, implementation and eventual reporting are addressed during the planning of the study.

Table 11.1 CONSORT 2010 check list to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item
Introduction	1a	Identification as a randomised trial in the title
	1b	Structured summary of trial design, methods, results, and conclusions
Background and objectives	2a	Scientific background and explanation of rationale
	2b	Specific objectives or hypotheses
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons
Participants	4a	Eligibility criteria for participants
	4b	Settings and locations where the data were collected
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed
	6b	Changes to trial outcomes after trial commenced, with reasons
Sample size	7a	How sample size was determined
	7b	When applicable, explanation of any interim analyses and stopping guidelines
Randomisation:		
Sequence generation	8a	Method used to generate the random allocation sequence
	8b	Type of randomisation; details of any restriction (such as blocking and block size)
Allocation mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions
Blinding	11a	If done, who was blinded after assignment to interventions and how (for example, participants, care providers, those assessing outcomes)
	11b	If relevant, description of the similarity of interventions
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for primary outcome
	13b	For each group, losses and exclusions after randomisation, together with reasons
Recruitment	14a	Dates defining the periods of recruitment and follow-up
	14b	Why the trial ended or was stopped
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group

Section/Topic	Item No	Checklist item
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether it was by originally assigned groups
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group and the estimated effect size and its precision (such as 95% confidence interval)
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory
Harms	19	All important harms or unintended effects in each group
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses
Generalisability	21	Generalisability (external validity, applicability) of the trial findings
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence
Other information		
Registration	23	Registration number and name of trial registry
Protocol	24	Where the full trial protocol can be accessed, if available
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders

*We strongly recommend reading this statement in conjunction with the CONSORT 2010 Explanation and Elaboration for important clarifications on all the items. If relevant, we also recommend reading CONSORT extensions for cluster randomised trials, non-inferiority and equivalence trials, non-pharmacological treatments, herbal interventions, and pragmatic trials. Additional extensions are forthcoming: for those and for up-to-date references relevant to this check list, see www.consort-statement.org .

References

Ahmad N, Boutron I, Moher D, Pitrou I, Roy C, Ravaud P. Neglected external validity in reports of randomized trials: the example of hip and knee osteoarthritis. Arthritis and Rheumatism. 2009;61(3):361-9.

Ali M, Emch M, von Seidlein L, Yunus M, Sack DA, Rao M, et al. Herd immunity conferred by killed oral cholera vaccines in Bangladesh: a reanalysis. Lancet. 2005;366(9479):44-9.

Andriole GL, Crawford ED, Grubb RL, 3rd, Buys SS, Chia D, Church TR, et al. Prostate Cancer Screening in the Randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: Mortality Results after 13 Years of Follow-up. J Nat Cancer Res. 2012;104(2):125-32.

Auleley GR, Giraudeau B, Baron G, Maillefert JF, Dougados M, Ravaud P. The methods for handling missing data in clinical trials influence sample size requirements. J Clin Epidemiol. 2004;57(5):447-53.

Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. J Clin Epidemiol. 2010;63(2):142-53.

Barthel FM, Babiker A, Royston P, Parmar MK. Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. Stat Med. 2006;25(15):2521-42.

Barthel FM, Parmar MK, Royston P. How do multi-stage, multi-arm trials compare to the traditional two-arm parallel group design--a reanalysis of 4 trials. Trials. 2009;10:21.

Bassler D, Montori VM, Briel M, Glasziou P, Guyatt G. Early stopping of randomized clinical trials for overt efficacy is problematic. J Clin Epidemiol. 2008;61(3):241-6.

Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. J Am Med Assoc. 2010;303(12):1180-7.

Berwanger O, Ribeiro RA, Finkelsztejn A, Watanabe M, Suzumura EA, Duncan BB, et al. The quality of reporting of trial abstracts is suboptimal: survey of major general medical journals. J Clin Epidemiol. 2009;62(4):387-92.

Borm GF, Fransen J, Lemmens WA. A simple sample size formula for analysis of covariance in randomized clinical trials. J Clin Epidemiol. 2007a;60(12):1234-8.

Borm GF, van der Wilt GJ, Kremer JA, Zielhuis GA. A generalized concept of power helped to choose optimal endpoints in clinical trials. J Clin Epidemiol. 2007b;60(4):375-81.

Boutron I, Ravaud P, Giraudeau B. Inappropriateness of randomised trials for complex phenomena: single trial is never enough evidence to base decisions on. BMJ (Clin Res). 2005;330(7482):94; author reply 5.

Boutron I, Guittet L, Estellat C, Moher D, Hrobjartsson A, Ravaud P. Reporting methods of blinding in randomized trials assessing non-pharmacological treatments. Plos Medicine. 2007;4(2):e61.

Boutron I, Touizer C, Pitrou I, Roy C, Ravaud P. The VEPRO trial: a cross-over randomised controlled trial comparing 2 progressive lenses for patients with presbyopia. Trials. 2008;9:54.

Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically non-significant results for primary outcomes. J Am Med Assoc. 2010;303(20):2058-64.

Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol. 2004;57(3):229-36.

Caesar_Study_Collaborative_Group. Caesarean section surgical techniques: a randomised factorial trial (CAESAR). Int J Obstet and Gynecol. 2010;117(11):1366-76.

Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. Stat Med. 2007;26(1):2-19.

D'Agostino RB, Sr., Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. Stat Med. 2003;22(2):169-86.

Dalum P, Paludan-Muller G, Engholm G, Kok G. A cluster randomised controlled trial of an adolescent smoking cessation intervention: Short and long-term effects. Scand J Public Health. 2012 Mar;40(2):167-76.

Dawson R, Lavori PW. Sequential causal inference: application to randomized trials of adaptive treatment strategies. Stat Med. 2008;27(10):1626-45.

de Hoop E, Teerenstra S, van Gaal BG, Moerbeek M, Borm GF. The "best balance" allocation led to optimal balance in cluster-controlled trials. J Clin Epidemiol. 2012;65(2):132-7.

Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. Am J Public Health. 2004;94(3):416-22.

Donner A, Taljaard M, Klar N. The merits of breaking the matches: a cautionary tale. Stat Med. 2007;26(9):2036-51.

Dragalin V, Fedorov V. Design of multi-centre trials with binary response. Stat Med. 2006;25(16):2701-19.

Fedorov V, Jones B. The design of multicentre trials. Stat Meth Med Res. 2005;14(3):205-48.

Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, Bryant DM, Montori VM, Alonso-Coello P, et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. J Clin Epidemiol. 2007;60(7):651-7; discussion 8-62.

Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: some design considerations. Clin Cancer Res. 2008;14(14):4368-71.

Freidlin B, Korn EL. Stopping clinical trials early for benefit: impact on estimation. Clinical Trials. 2009;6(2):119-25.

Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. Stat Med. 2008;27(27):5578-85.

Glezen WP. Herd protection against influenza. J Clin Virol. 2006;37(4):237-43.

Glezen WP, Gaglani MJ, Kozinetz CA, Piedra PA. Direct and indirect effectiveness of influenza vaccination delivered to children at school preceding an epidemic caused by 3 new influenza virus variants. J Inf Dis. 2010;202(11):1626-33.

Gluud LL. Bias in clinical intervention research. Am J Epidemiol. 2006;163(6):493-501.

Golub HL. The need for more efficient trial designs. Stat Med. 2006;25(19):3231-5; discussion 313-4, 326-47.

Groeneveld GJ, Veldink JH, van der Tweel I, Kalmijn S, Beijer C, de Visser M, et al. A randomized sequential trial of creatine in amyotrophic lateral sclerosis. Annals of Neurology. 2003;53(4):437-45.

Halloran ME. Overview of vaccine field studies: types of effects and designs. J Biopharm Stat. 2006;16(4):415-27.

Hernan MA, Hernandez-Diaz S. Beyond the intention-to-treat in comparative effectiveness research. Clinical Trials. 2011.

Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol. 2004;57(5):454-60.

Hernandez AV, Eijkemans MJ, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power? Annals Epidemiol. 2006;16(1):41-8.

Hofmeijer J, Anema PC, van der Tweel I. New algorithm for treatment allocation reduced selection bias and loss of power in small trials. J Clin Epidemiol. 2008;61(2):119-24.

Hrobjartsson A, Forfang E, Haahr MT, Als-Nielsen B, Brorson S. Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. Int J Epidemiol. 2007;36(3):654-63.

Hrobjartsson A, Boutron I. Blinding in randomized clinical trials: imposed impartiality. Clin Pharm and Therap. 2011;90(5):732-6.

Hudgens MG, Halloran ME. Causal vaccine effects on binary postinfection outcomes. J Am Stat Assoc. 2006;101(473):51-64.

Hudgens MG, Halloran ME. Toward causal inference with interference. J Am Stat Assoc. 2008;103(482):832-42.

Kane RL, Wang J, Garrard J. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. J Clin Epidemiol. 2007;60(3):241-9.

Kassai B, Shah NR, Leizorovicza A, Cucherat M, Gueyffier F, Boissel JP. The true treatment benefit is unpredictable in clinical trials using surrogate outcome measured with diagnostic tests. J Clin Epidemiol. 2005;58(10):1042-51.

Kirsch I. Antidepressants and the placebo response. Epidemiologia e psichiatria sociale. 2009;18(4):318-22.

Knatterud GL. Management and conduct of randomized controlled trials. Epidemiol Rev. 2002;24(1):12-25.

Kong L, Kohberger RC, Koch GG. Design of vaccine equivalence/non-inferiority trials with correlated multiple binomial endpoints. J Biopharm Stat. 2006;16(4):555-72.

Korn EL, Freidlin B. Conditional power calculations for clinical trials with historical controls. Stat Med. 2006;25(17):2922-31.

Korn EL, Freidlin B. A note on controlling the number of false positives. Biometrics. 2008;64(1):227-31.

Korn EL, Dodd LE, Freidlin B. Measurement error in the timing of events: effect on survival analyses in randomized clinical trials. Clinical Trials. 2010a;7(6):626-33.

Korn EL, Freidlin B, Mooney M. Bias and trials stopped early for benefit. J Am Med Assoc. 2010b;304(2):157-8; author reply 8-9.

Korn EL, Freidlin B, Abrams JS, Halabi S. Design Issues in Randomized Phase II/III Trials. J Clin Oncology. 2012.

Lancaster GA, Campbell MJ, Eldridge S, Farrin A, Marchant M, Muller S, et al. Trials in primary care: statistical issues in the design, conduct and evaluation of complex interventions. Stat Meth Med Res. 2010;19(4):349-77.

Lara PN, Redman MW. The hazards of randomized phase II trials. Annals of Oncology. 2012;23(1):7-9.

Laster LL, Johnson MF, Kotler ML. Non-inferiority trials: the 'at least as good as' criterion with dichotomous data. Stat Med. 2006;25(7):1115-30.

Lavori P, Kelsey J. Clinical trials. Epidemiol Rev. 2002;24:1-90.

Lavori PW, Dawson R. Improving the efficiency of estimation in randomized trials of adaptive treatment strategies. Clinical Trials. 2007;4(4):297-308.

Longini IM, Jr., Sagatelian K, Rida WN, Halloran ME. Optimal vaccine trial design when estimating vaccine efficacy for susceptibility and infectiousness from multiple populations. Stat Med. 1998;17(10):1121-36.

Longini IM, Jr., Halloran ME, Nizam A. Model-based estimation of vaccine effects from community vaccine trials. Stat Med. 2002;21(4):481-95.

Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. J Am Med Assoc. 2009;302(9):977-84.

McRae AD, Weijer C, Binik A, White A, Grimshaw JM, Boruch R, et al. Who is the research subject in cluster randomized trials in health research? Trials. 2011;12:183.

Moher D, Schulz KF, Altman D, %9 CG. The CONSORT Statement: revised recommendations for improving the quality of reports of parallel-group randomized trials 2001. Explore. 2005;1(1):40-5.

Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol. 2010;63(8):e1-37.

Ocana A, Amir E, Seruga B. Clinical research: show us the data. J Clin Oncology. 2011;29(9):1099-100.

Ocana A, Tannock IF. When are "positive" clinical trials in oncology truly positive? J Nat Cancer Res. 2011;103(1):16-20.

Ocana A, Amir E, Yeung C, Seruga B, Tannock IF. How valid are claims for synergy in published clinical studies? Annals of Oncology. 2012.

Peduzzi P, Henderson W, Hartigan P, Lavori P. Analysis of randomized controlled trials. Epidemiol Rev. 2002;24(1):26-38.

Peduzzi P, Kyriakides T, O'Connor TZ, Guarino P, Warren SR, Huang GD. Methodological issues in comparative effectiveness research: clinical trials. Am J Med. 2010;123(12 Suppl 1):e8-15.

Pitrou I, Boutron I, Ahmad N, Ravaud P. Reporting of safety results in published reports of randomized controlled trials. Archives Int Med. 2009;169(19):1756-61.

Riggs T, Koopman JS. Maximizing statistical power in group-randomized vaccine trials. Epidemiol and Inf. 2005;133(6):993-1008.

Royston P, Barthel FM, Parmar MK, Choodari-Oskooei B, Isham V. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. Trials. 2011;12:81.

Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002a;359(9306):614-8.

Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet. 2002b;359(9307):696-700.

Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. Lancet. 2002c;359(9310):966-70.

Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet. 2002d;359(9308):781-5.

Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet. 2002e;359(9305):515-9.

Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet. 2005a;365(9467):1348-53.

Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. Lancet. 2005b;365(9471):1657-61.

Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. J Clin Epidemiol. 2010;63(8):834-40.

Siqueira AL, Whitehead A, Todd S. Active-control trials with binary data: a comparison of methods for testing superiority or non-inferiority using the odds ratio. Stat Med. 2008;27(3):353-70.

Struchiner CJ, Halloran ME. Randomization and baseline transmission in vaccine field trials. Epidemiol and Inf. 2007;135(2):181-94.

Teerenstra S, Moerbeek M, van Achterberg T, Pelzer BJ, Borm GF. Sample size calculations for 3-level cluster randomized trials. Clinical Trials. 2008;5(5):486-95.

Teerenstra S, Lu B, Preisser JS, van Achterberg T, Borm GF. Sample size considerations for GEE analyses of three-level cluster randomized trials. Biometrics. 2010;66(4):1230-7.

Todd S. A 25-year review of sequential methodology in clinical studies. Stat Med. 2007;26(2):237-52.

Treweek S, McCormack K, Abalos E, Campbell M, Ramsay C, Zwarenstein M, et al. The Trial Protocol Tool: The PRACTIHC software tool that supported the writing of protocols for pragmatic randomized controlled trials. J Clin Epidemiol. 2006;59(11):1127-33.

Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. Trials. 2009;10:37.

Turner RM, White IR, Croudace T, Group PIPS. Analysis of cluster randomized cross-over trial data: a comparison of methods. Stat Med. 2007;26(2):274-89.

Twisk JW, de Vente W. The analysis of randomised controlled trial data with more than one follow-up measurement. A comparison between different approaches. Europ J Epidemiol. 2008;23(10):655-60.

Watson J, Helliwell P, Morton V, Adebajo A, Dickson J, Russell I, et al. Shoulder acute pain in primary healthcare: is retraining effective for GP principals? SAPPHIRE--a randomized controlled trial. Rheumatology. 2008;47(12):1795-802.

Weijer C, Grimshaw JM, Taljaard M, Binik A, Boruch R, Brehaut JC, et al. Ethical issues posed by cluster randomized trials in health research. Trials. 2011;12:100.

Zou GY, Donner A, Klar N. Group sequential methods for cluster randomization trials with binary outcomes. Clinical Trials. 2005;2(6):479-87.

Zwarenstein M, Oxman A, Pragmatic Trials in Health Care S. Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol. 2006;59(11):1125-6.