# The Randomized Controlled Trial: Methodological Perspectives

**Emmanuel Lesaffre**

## Introduction

A *clinical trial* is any form of planned experiment in medicine, which involves patients and is designed to elucidate the most appropriate treatment for future patients with a given medical condition. In the *randomized clinical trial* (RCT), the subjects are randomly assigned to two or more healthcare interventions. The results from this limited sample of patients are exploited to get insight about what treatment should be given in the general population of patients. In the famous pyramid of evidence-based medicine (see, e.g., Chapter 2 of [1]), the RCT scores the second highest (immediately below meta-analyses of RCTs) with respect to the hard evidence it provides about the tested intervention. In fact, the RCT is the only single study design which allows the researcher to draw causal relationships between a risk factor (absence or presence of experimental treatment) and outcome (improvement of the patient's condition).

RCTs have been widely used in health care starting only in the second half of the twentieth century with the British Medical Research Council trial of streptomycin for treatment of tuberculosis as the landmark study [2]. However, despite the inherent strength of the RCT, its conclusions are only to be trusted when it is set up and conducted properly. In this chapter, we review the essential concepts, steps in setting up, conducting, analyzing, and reporting as related to the RCT.

E. Lesaffre, Dr. Sc. (✉)
Department of Biostatistics, Erasmus MC, Dr. Molewaterplein, 50-60, Rotterdam 3015 GE, The Netherlands

L-Biostat, KU Leuven,
Leuven, Belgium
e-mail: e.lesaffre@erasmusmc.nl



H. Yazici et al. (eds.), *Understanding Evidence-Based Rheumatology*,
DOI 10.1007/978-3-319-08374-2_7

The National Institutes of Health (NIH) classifies the trials into six different types: (1) prevention trials, which aim to prevent people from disease via, e.g., lifestyle changes; (2) screening trials to detect, e.g., diseases; (3) diagnostic trials, which look for better diagnostic procedures; (4) treatment trials to test experimental treatments based on drugs, surgical techniques, etc.; (5) quality-of-life trials, which test new strategies to improve the quality of life of patients; and (6) compassionate use trials that offer experimental (but not yet approved) therapeutics to patients for whom there is no effective therapy and who have no other realistic options. In this chapter, we focus on drug treatment trials in rheumatology, but most of the topics discussed apply also to the other types of interventional studies. Furthermore, what we discuss is not limited to rheumatology.

## Phases of Clinical Research

Drug research is typically classified into the following stages:

- *Preclinical phase*: These are studies on animals to provide Information about efficacy, toxicity, and pharmacokinetics.
- *Phase 0*: This is to find out whether the drug behaves as expected. Subtherapeutic doses are administered to a small number of people (10–20 subjects).
- *Phase I* (Is the drug safe?): These are dose-ranging studies to find the maximally tolerated dose on healthy volunteers or patients (between 20 and 100 subjects), often done in an adaptive manner. In addition, initial information on adverse events is collected, together with pharmacokinetics and pharmacodynamics parameters.
- *Phase II* (Does the drug work?): This is about testing the drug on about 100–300 patients to obtain a better idea of efficacy and safety. This phase determines whether one should move on to phase III studies and are referred to as "proof of concept" studies.
- *Phase III* (Is the drug better than what is on the market?): Here, the formal testing of the therapeutic dose of the drug on patients takes place, involving typically at least 500–1,000 patients. This phase is decisive for the registration of the drug by regulatory agencies like the U.S. Food and Drug Administration (FDA, http://www.fda.gov/) and European Medicines Agency (EMEA, http://www.emea.europa.eu/).
- *Phase IV* (Are there other uses of/problems with the drug?): These are postmarketing surveillance studies to determine infrequent adverse events.
- *Phase V*: This phase is about translational research and is done on already collected data.

Sometimes, a further subdivision into phases IIa, IIb, IIIa, IIIb, etc. is made [3]. In [4] phase 0 to phase II, trials are referred to as *learning* (or exploratory) phase trials, while phase III trials are called *confirmatory* (also pivotal). Nowadays, there is a trend to shorten the entire regulatory process of an experimental drug by combining, especially, phase II and phase III trials in the so-called adaptive designs [5]. From above, it is evident that the aims are different in the different phases of clinical research. The topics discussed in this chapter primarily concern phase II and III trials.

## Asking the Appropriate Scientific Question

A well-formulated scientific question is an essential condition for a successful clinical trial. While it sounds almost as an obvious requirement, in practice, there is always the temptation to verify a great variety of (definitely interesting) clinical hypotheses in one RCT. There is now overwhelming evidence that the identification of an unambiguous *primary scientific question* has enormous benefits for all aspects of the trial. Less important research questions (but still within the scope of the trial) can then be classified as *secondary questions*. By focusing the design and implementation of the trial to address the needs of the primary question, one maximizes the chances of obtaining a definitive answer. Obviously, the nature of the primary and secondary questions depends on the phase of the trial.

The general principles of formulating scientific questions for RCTs of all phases are described below using the *PICO system* (http://www.usc.edu/hsc/ebnet/ebframe/PICO.htm), which specifies that a "well-built" question should identify: (1) the population, (2) the intervention and control treatment, and (3) the outcome.

### *The Population*

A detailed description of the study *population* is a necessary part of the scientific question. The RCT population is defined by *inclusion* and *exclusion criteria*. The inclusion criteria specify what kind of patients one wishes to treat. For instance, the inclusion criteria for patients with systemic sclerosis treated with a disease-modifying intervention might be (1) older than 18 years of age, (2) clinically apparent involvement of the skin on the extremities proximal to the elbows or knees or on the trunk, and (3) disease duration <2 years from the first symptom; see [6]. On the other hand, exclusion criteria aim to reduce the heterogeneity of the population. For instance, an often used exclusion criterion is "drug or alcohol abuse," but also "pregnant women." In reference [6], the investigators excluded also patients with kidney malfunction. Exclusion criteria also address ethical considerations. For instance, by excluding pregnant women, embryos are not exposed to unknown risks. Another typical exclusion criterion is the administration of concomitant medication that might interfere with the trial treatments. This is to avoid adverse events originating from sources other than those from the trial treatment.

Strict eligibility criteria will make the RCT population more homogeneous, which in general will reduce the variability of the outcome measure and hence the necessary study size. The drawback of strict criteria is that they limit the extrapolation of the trial results to the general patient population and thus affect the *generalizability*, also called the *external validity*, of the trial. For instance, by excluding pregnant women from the trial, no claim can be made on the efficacy and safety of the experimental treatment on this subpopulation. Note also that strict eligibility criteria may harden patient recruitment. Therefore, establishing appropriate inclusion and exclusion criteria is often a difficult process balancing between homogeneity and external validity. See also the section on "RCTs versus observational studies" for a further

discussion on external validity. Different eligibility criteria across studies with the same experimental treatment can throw light of the generalizability of the estimated efficacy and safety of the drug. However, when eligibility criteria vary wildly across RCTs with different experimental treatments, the comparability of the effects across the treatments will become difficult. In reference [6], a list of proposed guidelines for specifying eligibility criteria in systemic sclerosis is given.

## *Choice of Intervention and Control Treatment*

The choice of the interventional treatment is often quite clear from the start of the RCT, except for some possibly important details. Indeed, the RCT is set up to evaluate the effect of that intervention. That does not, however, mean that it is a fait accompli. For instance, in drug trials, it may be clear what the experimental medication is, but it still needs to be decided what the mode of delivery (tablet, solution, intravenous, or subcutaneous injection), dose, frequency, and timing of administration of the intervention will be.

Placebo treatment is often the preferred control treatment by regulatory agencies, unless there is an established accepted active treatment for the disease in which case it is unethical to administer a placebo treatment. It goes without saying that placebo treatment does not imply absence of treatment but rather that the standard care has been provided to the patient. Since standard care improves over time, the need for a "placebo" also evolves with time. The choice of the control (active or placebo) arm may impact the type of significance test; see section "Superiority and non-inferiority tests."

More than one intervention or control treatment may be considered. For instance, in phase I *dose escalation studies*, patients are allocated to one of several different doses of a drug with the goal of identifying the dose with an optimal trade-off between a desired biological action and unwanted side effects. Multiple group designs may also arise when combinations of interventions are tested, which occurs in the *factorial design* (see section "Factorial Designs").

## *Superiority and Non-inferiority Tests*

The classical statistical tests described in chapter "A review of statistical approaches for the analysis of data in rheumatology" aim to show that one treatment is superior to the other treatment, either a placebo or an active control. However, in many therapeutic areas, it becomes harder to improve upon current medication and one will be contented if the experimental drug has about the same efficacy as the control drug but shows better properties in other respects. This leads to non-inferiority tests explained below.

For a long time, clinical trials were only *superiority trials*; namely, they were set up to show that the experimental (E) treatment was superior to a control (C) treatment. The statistical tests used to analyze such trials are *superiority tests*. Recall that E is statistically significantly better than C at $\alpha = 0.05$ when $P < 0.05$, or equivalently that the 95 % confidence interval for the true difference (or ratio, odds ratio, etc.) does not include zero (or one in case of a relative measure).

It is, however, increasingly difficult to come up nowadays with new drugs that improve upon the existing ones in efficacy. For instance, thrombolytic agents have been developed over the last five decades to treat patients with an acute myocardial infarction. The initial 30-day mortality rates (percent of patients dying within 30 days after the onset of attack) were around 15 % but then dropped to about 6–7 % a decade later. It became clear that a further reduction in mortality rate may not be hoped for so that the focus turned into improving secondary objectives such as the mode of administration of the thrombolytic agent while preserving the achieved 30-day mortality rates. This requires other types of statistical tests, i.e., equivalence and non-inferiority tests, which will now be illustrated via fictive examples.

An example of a superiority trial in RA could be where one aims to show that $\Delta_S = 10$ % more patients who go into a remission after 6 months for the experimental treatment E compared to the placebo control. In another RCT, the aim might be to show that this experimental treatment has the same efficacy as a standard control treatment C. However, proving that E and C are equally effective is not possible in classical statistics, as we discussed in chapter "A review of statistical approaches for the analysis of data in rheumatology". In fact, we can only show that treatments are practically equivalent, say, that they differ in efficacy by at most, say, 2 %, in absolute value. Such a test is called an *equivalence test* and is used to show that generic drugs have similar properties as the original patented drugs. A one-sided version of such a test is a *non-inferiority test* and the associated trial a *non-inferiority trial*. Namely, a treatment E is called *non-inferior* to treatment C, when it is either better or not much worse than C where "not much worse" is defined by the *non-inferiority boundary* here denoted as $\Delta_{NI}$. This value should be chosen small enough so that it does not create ethical difficulties. Let us assume that for the above RA trial, $\Delta_{NI} = 2$ % is a good choice. Then one way to prove non-inferiority is to show that the 95 % confidence interval does not include $\Delta_{NI} = 2$ %. Since $\Delta_{NI}$ should be considerably smaller than $\Delta_S$, the required sample size with a non-inferiority study is often much larger than that of a superiority trial. The choice of the appropriate non-inferiority boundary is often subjective and requires a balanced choice between ethical, practical, statistical, and regulatory considerations. This may render the comparison of non-inferiority trials with different boundaries hard. Another flaw of the non-inferiority trial is that there is no standard analysis set, in contrast to a superiority trial where the intention-to-treat population is usually the default analysis set (as discussed below in the section on "Intention-to-treat versus per-protocol analysis"). An illustrative example of the difference between a superiority and non-inferiority trial can be found in [7]. Briefly, this study consists of two studies comparing etoricoxib 30 mg qd (ET) and celecoxib 200 mg qd (CE) to placebo (PL), which is the superiority part of the trial. In the second part of the trial, two studies

were conducted to compare the relative performance of ET and CE with a non-inferiority design. The two randomized three-arm double-blinded clinical trials described in [7] each contains a non-inferiority assessment of ET versus CE for the treatment of osteoarthritis of the knee and hip using a time-weighted average (TWA) change from baseline over 12 weeks in (a) the WOMAC pain subscale (WOMACPA), (b) the WOMAC physical function subscale (WOMACPH), and (c) the patient global assessment of disease status (PGADS). All three scales are scored on a visual analogue scale. The experimental treatment CE was defined to be non-inferior to ET when the upper bound of the two-sided 95 % CIs for the difference between CE and ET was not more than 10 mm for the three primary endpoints WOMACPA, WOMACPH, and PGADS. Thus, in order that non-inferiority is shown, all three conditions had to be satisfied. In [7], it is shown that for the two studies, these conditions were satisfied (95 % CIs entirely below upper bound), and the authors' conclusion was therefore that "etoricoxib 30 mg is comparable to celecoxib 200 mg in osteoarthritis." At first glance, the authors used a tough criterion for "non-inferiority," only it is not clear how they chose $\Delta_{NI} = 10$ mm.

We refer to [8] for a more detailed nontechnical introduction to non-inferiority studies, while a more technical and a broader discussion of the subject can be found in [9].

## *Study Outcomes*

The outcome, also called the *endpoint*, is the third component of the PICO system, and its characteristics determine many other aspects of the RCT. That is, the choice of the primary endpoint has a large impact on the size and the conduct of the study. *Hard endpoints*, such as mortality, leave no room for interpretation. However, when we choose for cardiac mortality, subjectivity creeps in since now the clinical judgement of the treating physician is required and this makes it a softer endpoint. *Soft endpoints* suffer from intra- and interobserver variability, and their use will therefore increase the necessary study size. Examples of (relatively) soft endpoints are, e.g., the EULAR response criteria (DAS and DAS28) and the ACR response criteria (ACR20, ACR50, ACR70); see also chapter "Outcome measures in rheumatoid arthritis." The use of many different criteria in European and US clinical trials to measure the rheumatic disease outcomes makes it difficult to compare and combine results in a meta-analysis (chapter "Systematic reviews and meta-analyses in rheumatology"). This was the trigger to establish the OMERACT network in 1992 [10]. Through regular meetings, the network aims to improve the outcome measures in rheumatology.

Clinical considerations may be in conflict with statistical requirements. For instance, it may be clinically more relevant to take the binary endpoint remission defined as DAS28 <2.6. However, from a statistical viewpoint, binarizing the endpoints implies a loss of information and hence a decrease in power. In addition, a statistical comparison between treatments based on DAS28 measurements only at

the study end may suffer a lot from intersubject variability. This variability can be drastically reduced by taking the improvement from baseline as an endpoint instead, as in the ACR criteria.

Prior to the study, it may feel unconformable to bet on one endpoint, so there is often the temptation to select several endpoints and then to choose the one that demonstrates best the efficacy of the experimental treatment. However, this leads to an inflation of the Type I error rate as we discussed in chapter "A review of statistical approaches for the analysis of data in rheumatology". An alternative approach is to make use of a *composite endpoint*, which is a clinical combination of different endpoints. Many of the responses in rheumatology trials are composite. Composite endpoints are also popular when the primary endpoint of interest exhibits a too low frequency, thereby increasing the necessary study size. An example in cardiovascular research is the binary composite endpoint MACE (major adverse cardiac events), which can be 0 or 1. While there are several definitions of MACE, the common definition is an outcome of death, having a myocardial infarction or a stroke. However, interpretation difficulties will occur when, e.g., a better result for MACE is seen under treatment A, while under treatment B, mortality is lower. Finally, we note that multiple endpoints can also be combined in a statistical manner (subject to the same issues as the above clinical composite endpoints) using a factor analysis technique; see chapter "A review of statistical approaches for the analysis of data in rheumatology" and [11].

*Patient-centered outcomes*, i.e., outcomes that represent a tangible benefit or harm to the patient, are especially most relevant in phase III trials. But because it may take too long to record the patient-centered endpoint, it might be necessary to choose for a *surrogate* endpoint, also called *disease-centered outcomes*. Such an outcome represents a measure of the disease process that is believed/hoped to be strongly related to a tangible patient benefit or harm. However, often such a relation is believed to exist purely from lower level studies, e.g., from animal studies. For example, in oncology, progression-free survival (PFS), which is the time to progression of the tumor, is often used in clinical trials as a surrogate outcome for overall survival. While there is a growing use of PFS as a primary outcome, there is no clear evidence of such a strong relationship (see, e.g., [12]), which therefore puts serious doubts on the usefulness of this outcome. We conclude that there are no specific statistical issues involved with using a surrogate endpoint; rather, the problem lies in the clinical interpretation of the study results. See also [13] for considerations on patient- and disease-centered outcomes.

Finally, in some studies, it may be of interest to express the benefit of an experimental treatment by the whole longitudinal profile of the primary endpoint or a summary measure of the profile. For instance, one might be interested in the rate with which DAS28 decreases over time. In that case, the average profiles need to be compared between the treatment arms, or at least the averages of the summary measure. This requires the use of longitudinal models as we saw in chapter "A review of statistical approaches for the analysis of data in rheumatology". In other studies, one might be interested in the time to an event. For instance, one might be interested in the time to remission (DAS28 <2.6). In that case, survival analysis techniques are required; see again chapter "A review of statistical approaches for the analysis of data in rheumatology".

## Randomization and Blinding

Random allocation of patients to treatments together with blinding enables one to draw a causal relationship between the administered treatment and the status of the patient at completion of the RCT. Randomization guarantees balance of the treatment arms with respect to the recorded covariates but also to all unmeasured covariates. Such a balance can never be achieved by any epidemiological study, irrespective of the analysis tricks used (e.g., with regression models).

Several randomization schemes are in use. The simplest randomization technique, e.g., by using a toss of a coin, is only practical for small study sizes, as in phase II studies, but even then, it is rarely used nowadays. Nowadays, the majority of phase III RCTs involve many centers often with a small number of patients from each center. In this case, simple randomization implies too much risk for imbalance in the treatment arms, and this could compromise the conclusions of the study. Therefore, a *blocked randomization* is often used in each center. Block randomization is not a pure stochastic allocation procedure anymore, but rather allocates patients to treatments such that balance is created within blocks of consecutive patients usually sized 6 to 10. To mask the block size (to avoid the investigator can predict the next treatment to administer), the block size is often taken random. Note that any randomization procedure that allocates patients within a center is called a *stratified randomization procedure* with center as stratum.

The with *adaptive randomization*, also discussed in the section on "Adaptive designs," the probability of allocation to one of the treatment arms may change over time. *Minimization* is an example of an adaptive allocation procedure that allocates subjects to treatments such that in a dynamic way, the imbalance of a set of a priori chosen covariates is minimized. For example, when the gender distribution is aimed to be balanced, the next male will be allocated to treatment B when the proportion of males is higher in the treatment arm A. The method is basically deterministic but can be given a stochastic flavor by adding a random component. Adaptive randomization can also be based on the response. In that case, more patients will be dynamically allocated to the winning treatment arm.

Note that randomization guarantees only that there is balance between the treatment groups for large samples. But, there is always the possibility of a random imbalance. Covariate adjustment, via using a regression model containing baseline covariates, can then, besides increasing the power, also remove the random imbalance and thereby improve the interpretability of the results.

Further, note that it does not make sense to statistically compare two randomized treatment groups at baseline with $P$-values since at baseline the patients are only different in the label they received from the trialist (A or B).

In practice, patients are allocated using any of the above procedures in combination with an automated (computerized) allocation system connected to either the Internet or a telephone. For example, the interactive voice response sys-

tem (IVRS) is a multi-language automated telephone system that allocates patients to different treatments, which can accommodate stratified randomization.

While randomization ensures balanced treatment arms at the start, blinding the treatment allocation to all parties of the RCT will avoid bias due to knowledge of which treatment was delivered. The terms "single blind" and "double blind" are often used to indicate that only the patient (single blind) or both the patient and clinician (double blind) are masked, but double blinding often means that basically everyone involved in the study is blinded during the conduct of the study. While double-blinded studies are the gold standard procedure, for some interventions, any blinding may be hard to achieve. For example, suppose that two knee-replacement surgical techniques are compared in one RCT, then blinding the surgeons will be impossible. On the other hand, there is a way out by appointing an evaluator (different from the treating surgeons) who is blinded to the administered treatment.

## Study Designs

In this section, various designs for RCTs are discussed. Focus will be on superiority trials, but what is discussed equally applies to non-inferiority trials.

### *Single-Center Versus Multicenter Studies*

A multicenter trial is a clinical trial conducted at more than one medical center or clinic. Most large clinical trials, particularly phase III trials, are conducted in several clinical research centers. The organizational aspects with single-center studies are considerably simpler than with multicenter studies. A simple illustration of this is that stratified randomization is required for multicenter studies, while simple randomization may readily work for single-center studies. Multicenter studies are recommended whenever it takes too much time for a single center to recruit the necessary number of patients. Such studies may also considerably increase the external validity of the RCT. Statistical methods for multicenter studies are somewhat more involved. While they should incorporate the stratification factor center, in practice, it is often and wrongly ignored. With a small number of centers, a Mantel-Haenszel-type test (see chapter "Methodological issues relevant to observational studies, registries, and administrative health databases in rheumatology") could be used or a regression model with each center as binary covariate (called fixed effects model). With many centers, a mixed effects model may be used with center represented by a random intercept, but there is no consensus on which model is preferable [14].

## *Parallel-Group Versus Cross-Over Designs*

Most popular is the *parallel-group design* whereby patients are randomly assigned to one of two (or more) treatment regimens and are followed up in time. It is a simple design, which is almost always possible to implement. The statistical analysis is often also straightforward involving only standard statistical tests such as the chi-square test for binary outcomes, the unpaired *t*-test for continuous outcomes, a *log-rank test* for survival outcomes, etc.; see chapter "A review of statistical approaches for the analysis of data in rheumatology".

On the other hand, in a *cross-over design* involving treatments A and B, each patient receives more than one treatment in a random order. Namely, one group of patients receives the treatment sequence A-B and the other group receives treatment sequence B-A. More complex allocations with more switches such as ABBA, BABA, etc. and more than two treatment arms are possible. This design has the advantage over the parallel group design in that within-patient treatment comparisons become possible by this method. This, in turn, removes a major portion of the intersubject variability and therefore commonly achieves a higher power than the parallel-arm design (with equal number of patients recruited). However, the cross-over design is only applicable in diseases where the patients return to their initial condition upon withdrawal of the study medication. This happens, for instance, when examining the effect of beta-blockers in treating hypertensive patients. An important issue with cross-over designs is that the effect of the first period treatment may leak into the second period and cause a *carry-over effect* (also called *cross-over effect*). In drug trials, this problem can be solved by inserting a *washout* period between the two treatment periods.

Cross-over designs are typically used in phase II trials, while parallel group designs are regularly applied in phase II and phase III studies. Both designs can be used in a single- and multicenter setting, but single-center cross-over studies are more frequent. Finally, we note that the statistical tests for the analysis of cross-over trials are extensions of the tests seen in chapter "A review of statistical approaches for the analysis of data in rheumatology" for paired data. A comprehensive treatment of cross-over designs can be found in [15].

## *Factorial Designs*

*Factorial designs* aim to examine the effects of two interventions simultaneously. In [16], an RCT with a factorial design was set up to examine the effect of patient-administered assessment tools for pain and disability, on the one hand, and an unsupervised home-based exercise program alone, on the other hand, or their combination on the symptoms of osteoarthritis. In that trial, the rheumatologists were assigned to four groups according to the treatment given to the patient: (1) patient-administered assessment tools, (2) or more exercises, (3) both tools and exercises, or (4) usual care. The aim was to check whether exercises have an impact on the symptoms and

also whether the assessment tools gave a better insight to the necessary treatment to also reduce symptoms. In addition, it was of interest to know whether the two or more interventions work synergistically when combined. Factorial designs are analyzed using 2-way ANOVA approaches when the response is continuous or with logistic regression models for binary or ordinal responses with interaction terms (see chapter "Evidence-based medicine in rheumatology: how does it differ from other diseases?").

## *The Cluster-Randomized Design*

In the abovementioned study [16], the rheumatologists and not the patients were randomized to treatments. In the case of the *cluster-randomized design*, all patients in a center are randomly assigned to the same treatment with the expectation that in another center all patients will be randomized to the alternative medication by other doctors. This design may be needed when it is not practical or ethical to randomize patients within a center, which was the case in [16]. In that study, the cluster-randomized design was chosen because the investigators were convinced that one could not insist that one physician advises one patient to do physical exercises and not give the same advice to other patients. Therefore, each rheumatologist was to enroll four patients with osteoarthritis. Since the response of the four patients assigned to a rheumatologist is more alike than for the patients assigned to another rheumatologist, there is more clustering in the data as compared to what is seen in standard multicenter studies. This almost inevitable clustering must be taken into account at the design stage (increasing the sample size compared to a design without clustering) and at the analysis stage. Specialized statistical methodology has been developed for cluster-randomized designs [17, 18] to account for the correlation between outcomes within a cluster.

## *Group Sequential Designs*

It may be of interest for ethical and/or commercial reasons to evaluate the results of an RCT at an interim time. However, such interim analyses cannot be done ad hoc; there are statistical issues with repeated testing (multiple testing), and it would be impractical not to know in advance when cleaned data ready for inspection should be available. Regulatory authorities require indeed a correction for multiple testing. However, not all interim analyses deserve a statistical penalty. There are three types of interim analyses: (1) administrative interim analyses, (2) interim analyses for safety, and (3) interim analyses for efficacy. Such analyses are typically evaluated by an external committee called the Data Monitoring Committee (DMC) (also called Data and Safety Monitoring Board (DSMB)) consisting of two to four clinicians and one independent statistician. The purpose of an *administrative interim*

*analysis* is to evaluate whether the study, up to that time point, has been conducted according to the plan. If the number of patients enrolled has been too few, the DMC may suggest including more centers in the study or to relax the inclusion and exclusion criteria. With an administrative interim analysis, there is no statistical penalty.

*Interim analyses for safety* are necessary for RCTs where there is a risk for life-threatening adverse events. In such an interim analysis, the DMC reviews safety and other data (e.g., demographic data, data on past medication use, etc.) in a semi-blinded (only the labels A and B are given, not the actual treatments) or unblinded manner. Again, no correction for multiple testing is required, since the treatments are not compared for efficacy.

*Interim analyses for efficacy* involve repeated statistical comparisons between the administered treatments with the aim to see whether the study can be stopped early for efficacy. A correction for multiple testing is required to avoid producing spurious conclusions. Correction for multiple testing is done with dedicated procedures that devote at each interim analysis a part of the overall significance level $\alpha$ (often equal to 0.05) such that together they amount to $\alpha$. The methods look similar to the Bonferroni correction, but here, they capitalize on the staggered data presented at the DMC meetings. They are therefore called *group sequential designs,* but in contrast to Bonferroni correction, their global significance level is exactly the a priori defined $\alpha$. A group sequential design allows stopping the study when the results become convincing enough. In that case, the number of patients needed enroll will be less than originally planned. However, the originally planned (maximal) sample size will be larger with planned interim analyses because the correction for multiple testing inflates the sample size. Pocock's method [19] was one of the first group sequential designs. The procedure specifies an equal, more stringent, significance level at each interim analysis, e.g., for $\alpha = 0.05$ and 5 analyses (4 interim and one final analysis), the study can only be stopped when the $P$-value is smaller than 0.016. Nowadays, the O'Brien-Fleming [14] design is more popular. For this design, a very stringent significance level is used in the early part of the study making it hard to stop early, but is then relaxed towards the end of the study. The timing of the repeated analyses with group sequential designs can be *calendar-driven* or *event-driven*; they must however be specified at the start of the study. A more flexible design was proposed by Lan and DeMets [14], which allows flexible timing and number of analyses, called the *alpha spending approach*. This is now the most popular approach because of its flexibility and has been extended in various ways, e.g., to non-inferiority studies, to cluster-randomized designs, etc. Note that these designs can be also used to monitor safety.

The second type of interim analysis for efficacy checks whether there is a reasonable chance that the study will be positive at the end. Such an analysis, called *futility analysis*, aims to avoid wasting financial resources in a study that has little chance to show a beneficial effect of the experimental treatment. The need for correction for multiple testing is, in this instance, negligible, since now the trial cannot be stopped when at interim the experimental arm shows much better efficacy than the control arm.

### *Adaptive Designs*

*Adaptive designs* are generalizations of the group sequential designs. Examples of adaptive designs are (1) determination of the maximum tolerated dose in a phase I oncology trial, (2) adaptive randomization, (3) sample size reestimation, and (4) adaptive seamless designs. Below, we briefly elaborate on some of these examples but refer to [5, 20] for more details and references. Adaptive designs are sometimes referred to as flexible designs, but the latter incorporate both planned and unplanned features, while the first must be described in detail at the start of the study and must ensure that the probability for a Type 1 error is addressed.

In phase I oncology trials, there is the *continual reassessment method* (CRM) [5, 28], which is a Bayesian approach to determine the maximum tolerated dose of the test drug. It involves assuming a model for the relationship between the dose and the probability of an unacceptable side effect. The maximal dose that a new patient can be administered is determined via the (Bayesian posterior) probability of causing an unacceptable adverse effect.

*Adaptive randomization* is an allocation rule whereby the allocation probabilities depend on covariate imbalance and/or response imbalance (see section "Randomization and blinding" for more details).

Establishing the sample size of a study is not an easy task, always prone to misjudgment. In section "Sample size calculations," we show how sample size estimation is done in practice and indicate possible difficulties in establishing a well-motivated choice. It seems reasonable to roughly guess the sample size of a pilot portion of the trial data and then reestimate the sample size for the whole trial based on this. This is done in a calibrated *internal pilot design*, which is a two-stage design with no interim testing for efficacy but only estimating the nuisance parameters (say common standard deviation for an unpaired *t*-test) from the first-stage data. This approach does not necessitate a correction in the projected Type I error rate.

Traditionally, phase II and phase III trials are set up in two distinct stages. Since this may delay regulatory approval, statisticians have looked for ways to speed up the approval of experimental treatments. One way is to rapidly move from phase II to phase III studies, in fact in a seamless manner. This is done in an *adaptive seamless design* that combines the data of the two stages for the final analysis. For example, one trial could consist in choosing between two doses of a drug in the first stage, while in the second stage, the chosen dose is compared to a control group.

Adaptive designs have recently gained a lot of popularity. However, they are considerably more complex not only from a statistical viewpoint but also from an organizational viewpoint, needing a much more sophisticated clinical trial infrastructure.

## Sample Size Calculations

The sample size calculation is an essential part of any RCT. It attempts to minimize the risk of not detecting the aimed effect (if present) of the experimental treatment vis-à-vis the control treatment. Ultimately, a statistical test determines the necessary

sample size and is a quite technical job that most often requires a computer program. The computation of the sample size for a classical (superiority) unpaired *t*-test goes as follows: (1) fix the overall significance (two-sided) level $\alpha$ (usually = 0.05) and the power (at least 0.80); (2) choose the clinically relevant difference $\Delta_S$ (not the difference that we expect but the difference that we aim for); (3) make an educated guess about the common standard deviation $\sigma$; and (4) the sample size in each treatment arm is then the result of the equation $n=\left(t_{2n-2,a/2}+t_{2n-2,a/2}\right)^2/\Delta_S^2$, with $t_{2n-2,-/2}$ is the $\alpha/2$ quantile of a t-distribution with $2n-2$ degrees of freedom.

These steps illustrate a few important things for computing the sample size:

- Clinicians must have a good idea of the effect they aim to show, i.e., what value for $\Delta_S$ to choose, but they do not need to guess what the true effect might be.
- Extra information to perform the computations is usually required. Here, it is the common standard deviation. For the comparison of two proportions, it is the proportion of the control arm.
- The computation of the sample size is in general quite technical, varies from test to test, and usually requires a dedicated computer program.

Note that for a non-inferiority test, $\Delta_S$ needs to be replaced by $\Delta_{NI}$ and the statistical test needs to be adapted accordingly. For group sequential designs, dedicated programs have been written not only to compute the sample size but also to compute the intermediate significance levels. For more complicated statistical tests, such as for mixed models (see chapter "A review of statistical approaches for the analysis of data in rheumatology") and adaptive designs, often only a simulation computer program may throw light on the required study size. A comprehensive, but technical, reference for sample size calculation is given in [21].

## Intention-to-Treat Versus Per-Protocol Analysis

The eligibility criteria specify which of the screened patients will be included in the statistical analysis. However, during the conduct of the study, a lot of deviations from the initial plan may take place. For instance, it may happen that due to an administrative error, a patient who should have been randomized to treatment A in fact received treatment B, or that a patient violates the protocol (takes forbidden concomitant medication), or even drops out from the study, etc. What to do with such patients? One approach is to take in the analysis only the "pure patient population," i.e., only patients who strictly adhere to the instructions. This set of patients is called the *per-protocol (PP) set* and is preferred by many clinicians because it is believed to express best what the effect of the treatment is on the patients. That is true for the patients still included at the end of the study, but not necessarily for all patients randomized. It is rather the *intention-to-treat (ITT) set* that is the standard in RCTs. The ITT principle states that all patients who have been randomized in the study should be included in the analysis according to the planned treatment irrespective of what happened during the conduct of the trial. This principle may appear

logical at first but may have some unexpected implications. For instance, patients wrongly allocated to B will be analyzed as if they received treatment A; protocol violators are in the ITT analysis set, also patients dropping out the study will be part of the ITT population, etc. FDA and EMEA prefer the ITT analysis in a superiority trial, because it delivers a conservative result in case of the abovementioned problems during the conduct of the study. While the ITT principle is clear, in practice, it may not always be easy to implement and consequently several versions of an ITT analysis exist. For example, it is not immediately clear how to include patients in an ITT analysis with missing values on the primary endpoint. In that case, the ITT analysis cannot include all randomized subjects. But if some values of the primary response are available, then techniques for imputing missing values allow for including such dropouts. Statistical methods that can deal appropriately with missing data are quite important to guarantee the *internal validity* of the RCT, i.e., that the RCT estimates the true treatment effect in an unbiased manner. An imputation technique that was quite popular for many years but now recognized as problematic is *the last-observation-carried-forward (LOCF) approach*. This imputation technique imputes the last observed value for the missing primary outcome. For example, suppose the total treatment period is 2 years and every 6 months the primary outcome is measured. Then, when a patient drops out at year 1, the imputed value with the LOCF method for the primary outcome at years 1.5 and 2 is equal to the value observed at year 1. The problem with the LOCF approach is that it imputes an unrealistic value for the outcome (not taking into account the natural pattern of the disease and/or of the curing process) and it underestimates the natural variability of the outcome. In [22], more appropriate imputation techniques are discussed.

In an equivalence or non-inferiority study, the ITT analysis is not the primary analysis anymore since the ITT analysis will bias the results and the conclusions towards the desired hypothesis (equivalence or non-inferiority). Because also the PP analysis does not guarantee to provide an unbiased estimate, regulatory agencies require that an ITT and a PP analysis are performed in an equivalence/non-inferiority RCT and that they show consistent results.

## RCT and Some Practical Aspects

The *protocol* is the reference manual for the RCT containing the background of the intervention, the reason and motivation for conducting the trial, a review of the phase I and phase II results, the justification of the sample size, the eligibility criteria, and the primary and secondary endpoints. In addition, it contains details of the randomization procedure, the informed consent document, the administration of the interventions, etc.

Furthermore, NIH developed a document, called the *Manual of Procedures* (MOP) (http://www.ninds.nih.gov/research/clinical_research/policies/mop.htm), that transforms a protocol into an operational research project that ensures compliance with federal law and regulations. The MOP typically describes in detail all key ingredients

of the conduct of the study, for instance, how data capture will be done, how the patients will be followed up in order to maximize data collection, etc. For example, a list of all eligible patients is never available at the start of an RCT, so the process by which potential trial participants are identified needs to be explicitly stated at the start. In practical terms, this implies that it needs to be specified which countries and centers will be involved in the RCT and what characteristics the involved centers should have.

The protocol also specifies which statistical tests will be chosen for analysis. This can be tricky since many statistical tests depend on distributional assumptions. For instance, the unpaired *t*-test assumes that there is normality in each of the two treatment arms and that the variances are equal. But, one can only test these assumptions when the results roll in. This rigid requirement does not leave much room for creativity, but is needed to preserve the Type I error rate. As an example, suppose that the protocol dictates to choose the unpaired *t*-test but that this test does not yield a significantly better result for the experimental arm while a nonparametric Wilcoxon rank-sum test (see chapter "A review of statistical approaches for the analysis of data in rheumatology") does. Hand switching from one statistical test to another only on the basis of the obtained *P*-value is an example of a data dredging exercise, which is known to produce many spurious results. In a RCT, all statistical activities should be described in even more detail than discussed in *the statistical analysis plan* (SAP). The SAP is typically finalized prior to locking the database to avoid speculative choices of statistical procedures.

Trial participants must be fully aware of the risks and benefits of participation and therefore must fill in an *informed consent* form. This document is also part of the trial protocol.

Finally, each protocol of a RCT needs to be approved by the *Medical Ethical Committee* of the centers where the study is conducted; they are also called *Institutional Review Boards* in the United States. In addition, in order to avoid difficulties when applying for registration, protocols are nowadays often discussed with the regulatory bodies to obtain approval (not the drug!) prior to the start of the RCT.

## Reporting the Results of a RCT

The statistical analysis plan specifies in detail which statistical tests need to be chosen. No doubt this is accordingly reported in the registration file for the experimental drug, but this is not necessarily the case for the scientific paper written after the study is finalized. Indeed, most referees of medical journals do not check the consistency of the technical report with the submitted paper. Hence, in principle, the reader cannot be sure that the analysis described in the scientific paper is an exact reflection of what has been specified in the protocol. For example, a recently published phase III trial compared pazopanib with sunitinib with respect to progression-free survival in renal-cell carcinoma patients [23]. In that paper, the authors state

that "the results of the progression-free survival analysis in the per-protocol population were consistent with the results of the primary analysis" without providing further details. However, from the technical report, one can infer that the predefined margin of non-inferiority (<1.25) was only met for the ITT population and not for the PP population. This is in conflict with the requirement that in both analysis sets, non-inferiority must be claimed (see also [24]).

*Subgroup analyses* have been a topic of discussion already for many years. Next to the global analysis, clinicians wish to know which patients (if any) may benefit most from the experimental treatment (if any). Therefore, subsequent to a global primary analysis, often the treatments are compared in a variety of subgroups, e.g., within the group of patients (a) below 65 years of age, (b) above 65 years of age, (c) males, (d) females, etc. This is a typical example of data dredging, especially because there is often no strong clinical background why in a particular subgroup the experimental treatment should do much better. Subgroup analyses are sometimes prespecified in the protocol, but that does not alleviate the problem much. Subgroup analyses can be thought provoking, but should always be considered as exploratory analyses for which the conclusions need to be verified with a new study or in a meta-analysis.

## RCTs Versus Observational Studies

In chapter "Methodological issues relevant to observational studies, registries, and administrative health databases in rheumatology" it is seen that the major difference of the observational study with the RCT is that in the observational study, the groups are self-selected. This causes the groups to be different at baseline. The problem is now that there is no way to guarantee that the difference in disease outcome may not be a result of an existing difference at the start of the study. Hence, it is said that an observational study has in general a relatively low internal validity. Regression methods (including the method of propensity scores, see chapter "Methodological issues relevant to observational studies, registries, and administrative health databases in rheumatology") may improve the internal validity by correcting for baseline imbalance, but one can never rule out a residual imbalance caused by unobserved characteristics of the patients. On the other hand, randomization and blinding alone do not guarantee that an RCT has a high internal validity. Indeed, the internal validity can be highly affected by missing values and dropouts. For example, if in one treatment group, patients drop out because of inefficacy of the treatment, while in the other treatment group patients drop out because of safety concerns, then the estimated treatment effect at the end of the RCT is likely not to be a good estimate of the true treatment effect based on all patients who should have been treated.

In an observational study, a heterogeneous group of subjects is included. This is in contrast to an RCT where a homogeneous group of patients is aimed at. This implies that an observational study has a higher external validity than a RCT.

In [25], the factors that cause the low external validity of the RCT are discussed; see also [26]. The author discusses the impact of the general settings of the trial (e.g., the country or countries in which the study is executed), the eligibility criteria of the patients, the difference between the trial protocol and routine practice, etc. Further, the author recommends a thorough consideration of factors which might interfere with the generalizability of the RCT findings to the clinical practice (see also chapter "Limitations of traditional randomized controlled clinical trials in rheumatology").

## The Bayesian Approach to RCTs

In chapter "A review of statistical approaches for the analysis of data in rheumatology" the Bayesian approach to inference was introduced. The main difference with the classical (also called frequentist) approach is that the posterior distribution and its summary measures make up the inference, instead of the *P*-value. For instance, in the frequentist approach, the conservation of the overall Type I error is the motivation to develop the group sequential designs that allow for interim analyses in a calibrated manner. A Bayesian approach in this case consists in repeatedly evaluating the posterior probability that the experimental treatment is better than the control treatment and stops either when the planned number of patients was recruited or that posterior probability was, say, greater than 0.975. An alternative Bayesian approach is to let the stopping rule based on the posterior predictive probability, generate future samples (combined with the already sampled subjects), and determine the predictive probability of a significant result (with a classical test), as was done in [27]. Yet another example of a Bayesian approach is an interim analysis that exploits prior information on the drug (say from phase II and III studies) when monitoring the safety of the drug for a rare event in a phase III study.

The *Bayesian adaptive approach*, i.e., counterpart of the frequentist adaptive approach, is gaining much popularity. While the frequentist approach aims to maintain the overall Type I error rate, the Bayesian approach monitors the more intuitive posterior probabilities (which could be a few in a complex design). We refer to [28] for a recent but a technical review of this topic.

The Bayesian approach has been widely accepted in phase I studies, as mentioned above, and is becoming increasingly popular in phase II studies. Yet in a phase III study, this approach is more used for interim and auxiliary analyses, and there is great resistance against its use for the primary analysis. However, I predict that when combined with non-informative priors, the Bayesian approach will likely become one of the options for a phase III study if the investigator can show its good frequentist properties. Note that in medical device trials, it has now become one of the standard approaches. See the NIH website on guidelines for the use of Bayesian methods for medical devices:http://

www.fda.gov/medicaldevices/deviceregulationand guidance/guidancedocuments/ucm071072.htm

## Conclusions

After its introduction in the 1940s, the RCT remains the only study type that allows for causal relationships between risk factors and disease outcome. However, that does not mean it is the only study type useful for this. The often limited external validity, and the difficulty to keep the internal validity high, requires considering alternative study types more explored in chapter "Methodological issues relevant to observational studies, registries, and administrative health databases in rheumatology" in the context of using registries and administrative data bases.

There are many excellent textbooks on RCTs. For an accessible introduction for clinicians, there is the standard book of Pocock [29] and the more recent book by Senn [30].

## References

1. Lesaffre E, Feine J, Leroux B, Declerck D, editors. Statistical and methodological aspects of oral health research. New York: Wiley; 2009.
2. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. Br Med J. 1948;2:769–82.
3. Spilker BL. Guide to clinical studies and developing protocols. New York: Raven; 1984.
4. Pong A, Chow S-C, editors. Handbook of adaptive designs in pharmaceutical and clinical development. Boca Raton: CRC Press; 2011.
5. Chin R. Adaptive and flexible clinical trials. Boca Raton: CRC Press; 2012.
6. White B, Bauer EA, Goldsmith LA, et al. Guidelines for clinical trials in systemic sclerosis (sleroderma). Arthritis Rheum. 1995;38:351–60.
7. Bingham III CO, Sebba AI, Rubin BR, et al. Efficacy and safety of etoricoxib 30 mg and celecoxib 200 mg in the treatment of osteoarthritis in two identically designed, randomized, placebo-controlled, non-inferiority studies. Rheumatology. 2007;46:496–507.
8. Lesaffre E. Superiority, equivalence and non-inferiority trials. Bull NYU Hosp Jt Dis. 2008; 66(2):150–4.
9. Rothmann MD, Wiens BL, Chan ISF. Design and analysis of non-inferiority trials. Boca Raton: CRC Press; 2012.
10. Tugwell P, Boers M, Brooks P, Simon L, Strand V, Idzerda L. OMERACT: an international initiative to improve outcome measurement in rheumatology. Biomed Cent. 2007;8:38. 1–6.
11. Pillemer SR, Tilley B. Clinical trials, outcome measures, and response criteria. J Rheumatol. 2004;31:407–10.
12. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or measurable. J Clin Oncol. 2012;30(10):1030–3.
13. Lassere MN, Johnson KR, Boers M, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: developments and testing of a quantitative hierarchical levels of evidence schema. J Rheumatol. 2007;34:607–15.

14. Senn S. Some controversies in planning and analyzing multi-centre trials. Stat Med. 1998;17: 1753–65.
15. Byron J, Kenward MG. Design and analysis of cross-over trials. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2003.
16. Ravaud P, Giraudeau B, Logaert I, et al. Management of osteoarthritis (OA) with an unsupervised home based exercise programme and/or patient administered assessment tools. A cluster randomized controlled trial with a 2×2 factorial design. Ann Rheumatol Dis. 2004;63:703–8.
17. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.
18. Murray DM. Design and analysis of group-randomized trials. New York: Oxford University Press; 1998.
19. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Boca Raton: CRC Press; 2000.
20. Kairalla JA, Coffey CS, Thomann MA, Muller KE. Adaptive trial designs: a review of barriers and opportunities, Biomed Cent. 2012;13:145. http://www.trialsjournal.com/content/13/1/145.
21. Julious SA. Sample sizes for clinical trials. Boca Raton: CRC Press; 2009.
22. Molenberghs G, Kenward MG. Missing data in clinical studies. Chichester: Wiley; 2007.
23. Motzer RJ, Hutson TE, Cella D, et al. Pazopanib versus sunitinib in metastatic renal-cell carcinoma. N Engl J Med. 2013;369:722–31.
24. Casper J, Schumann-Binarsch B, Köhne C-H: Letter to the Editor to Motzer RJ, Hutson TE, Cella D, et al. Pazopanib versus sunitinib in metastatic renal-cell carcinoma. N Engl J Med. 2013;369:1969.
25. Rothwell PM. External validity of randomized controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365:82–93.
26. Pincus T, Sokka T. Should contemporary rheumatoid arthritis clinical trials be more like standard patient care and vice versa. Ann Rheum Dis. 2004;63(Suppl II):ii32–9.
27. Buzdar AU, Ibrahim NK, Francis D, et al. Significantly higher pathologic complete remission rate after neoadjuvant therapy with trastuzumab, paclitaxel, and epirubicin chemotherapy: results of a randomized trial in human epidermal growth factor receptor 2–positive operable breast cancer. J Clin Oncol. 2005;23(16):3676–85.
28. Berry SM, Carlin BP, Lee JJ, Müller P. Bayesian adaptive methods for clinical trials. Boca Raton: CRC Press; 2011.
29. Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley; 1987.
30. Senn S. Statistical issues in drug development. 2nd ed. New York: Wiley; 2008.