Chapter 3
A Focus on Clinical Trials

Stephen P. Glasser

A researcher is in a gondola of a balloon that loses lift and lands in the middle of a field near a road. Of course, it looks like the balloon landed in the middle of nowhere. As the researcher ponders appropriate courses of action, another person wanders by. The researcher asks, 'Where am I?' The other person responds, 'You are in the gondola of a balloon in the middle of a field.' The researcher comments, 'You must design clinical trials.' 'Well, that's amazing, how did you know?' 'Your answer was correct and precise and totally useless.' (ANON)

Abstract The spectrum of evidence imparted by the different clinical research designs ranges from ecological studies through observational epidemiological studies to randomized control trials (RCTs). This chapter addresses the definition of clinical research, the major aspects of clinical trials e.g. ethics, randomization, masking, recruitment and retention of subjects enrolled in a clinical trial, patients/subjects lost to follow-up during the trial etc. Although this chapter focuses on the weaknesses of clinical trials, it is emphasized that the randomized, placebo-controlled, double blind clinical trial is the design that yields the greatest level of scientific evidence.

Keywords Generalizability/external validity. Internal validity • Superiority testing • Equivalence/noninferiority testing • Randomization • Intention to treat • Missing

---

Less than 50 % of this chapter is taken from "Clinical trial design issues: at least 10 things you should look for in clinical trials" [1] with permission of the publisher.

S.P. Glasser, M.D. (✉)
Division of Preventive Medicine, University of Alabama at Birmingham,
1717 11th Ave S MT638, Birmingham, AL 35205, USA
e-mail: sglasser@uabmc.edu

S.P. Glasser (ed.), *Essentials of Clinical Research*, DOI 10.1007/978-3-319-05470-4_3,

data • Eligibility • Efficacy/effectiveness • Blinding/masking • Subgroup analysis • Surrogate endpoints • Composite endpoints • Primary and secondary endpoints

The differences in clinical research designs and the different weights of evidence imparted by different clinical research designs, are exemplified by the post-menopausal hormone replacement therapy (HRT) controversy. Multiple observational epidemiological studies had shown that HRT was strongly associated with the reduction of atherosclerosis, myocardial infarction risk, and stroke risk [2–4]. Subsequently, 3 RCTs suggested that HRT was not beneficial, and might even be harmful [5–7]. This latter observation raises a number of questions, including: why can this paradox occur? What can contribute to this disagreement?; and, why do we believe these 3 RCT's more than so many well-done observational trials? The reasons for this are many (also see Chap. 2), but include: concerns about the generalizability of clinical trial results to the general population, and the reproducibility of the results; and, RCTs are increasingly involving thousands of patients form many sites, and from multiple countries making them challenging to design and difficult to execute and monitor [8]. Also, some clinical trials have been criticized by regulatory agencies due to apparent high dropout rates and patients lost to follow up, which has led to new FDA guidelines emphasizing the importance of patient retention and innovative site monitoring [9]. In support of this latter issue, is a post hoc analysis of the Efficacy of Vasopressin Antagonism in Heart Failure: Outcome Study with Tolvaptin (EVEREST) in which the authors evaluated the relationship between the number of patients enrolled in each site with trial outcomes. They found that the high enrolling sites had better clinical outcomes and more protocol completion rates compared to the lower enrolling sites [10]. Of course, there are a number of explanations for this observation from EVEREST, and as was pointed out in the discussion of this trial, the use of block randomization (see below) within each center should have equally distributed patients between the sites of potentially differing quality who were on or off study drug; none-the-less, the point is one worthy of further research [10]. Participant differences based on geographic disparities have been well described, but differences related to participant volume have not.

Frequently, there is confusion about the difference between clinical research and clinical trials. In general usage experimental design is the design of any information-gathering exercises where variation is present, whether under the full control of the experimenter or not. Other types of study are opinion polls and statistical surveys (which are types of observational study), natural experiments and quasi-experiments. In the design of experiments, the experimenter is often interested in the effect of some process or intervention (the "treatment") on some objects (the "experimental units"), which may be people, parts of people, groups of people, plants, animals, materials, etc.

A clinical trial is a type of experimental study undertaken to assess the response of an individual (or in the case of group clinical trials-a population) to interventions introduced by an investigator. Clinical trials can be randomized or non-randomized,

un-blinded, single-blinded, or double-blinded; comparator groups can be placebo, active controls, or no treatment controls, and RCTs can have a variety of designs (e.g. parallel group, crossover, etc.). That being said, the RCT remains the 'gold-standard' study design and its results are appropriately credited as yielding the highest level of scientific evidence (greatest likelihood of causation). However, recognition of the limitations of the RCT is also important so that results from RCTs are not blindly accepted. As Grimes and Schultz point out, in this era of increasing demands on a clinician's time it is 'difficult to stay abreast of the literature, much less read it critically. In our view, this has led to the somewhat uncritical acceptance of the results of a randomized clinical trial' [11]. Also, Loscalzo, has pointed out that 'errors in clinical trial design and statistical assessment are, unfortunately, more common that a careful student of the art should accept' [12].

What leads the RCT to the highest level of evidence and what are the features of the RCT that renders it so useful? Arguably, one of the most important issues in clinical trials is having matched groups in the interventional and control arms; and, this is best accomplished by randomization. That is, to the degree that the two groups under study are different, results can be confounded by any difference, while when the two groups are similar, confounding is reduced (see Chap. 17 for a discussion of confounding). It is true that when potential confounding variables are known, one can relatively easily adjust for them in the design or analysis phase of the study. For example, if one believes that smoking might confound the results of the success of treatment for hypertension, one can build into the design a stratification scheme that separates smokers form non-smokers, before the intervention is administered and in that way determine if there are differential effects in the success of treatment (e.g. smokers and non-smokers are randomized equally to the intervention and control). Conversely, one can adjust after data collection in the analysis phase by separating the smokers from the non-smokers and again analyze them separately in terms of the success of the intervention compared to the control. The real challenge of clinical research, is not how to adjust for **known** confounders, but how to have matched variables in the intervention and control arms, when potential confounders are **not** known. Optimal matching is accomplished with randomization, and this is why randomization is so important. More about randomization later, but in the meanwhile one can begin to ponder how un-matching might occur even in a RCT. In addition to randomization, there are a number of important considerations that exist regarding the conduct of a clinical trial, such as: is it ethical? What type of comparator group should be used? What type of design and analysis technique will be utilized? How many subjects are needed and how will they be recruited and retained?

Finally, there are issues unique to RCTs (e.g. intention-to-treat analysis, placebo control groups, randomization, equivalence testing) and issues common to all clinical research (e.g. ethical issues, blinding, selection of the control group, choice of the outcome/endpoint, trial duration, etc.) that must be considered (Table 3.1). Each of these issues will be reviewed in this chapter. To this end, both the positive and problematic areas of RCTs will be highlighted.

Table 3.1 Issues of importance for RCTs

Ethical considerations
Randomization
Eligibility criteria
Efficacy vs. effectiveness
Compliance
Run-in periods
Recruitment and retention
Masking
Comparison groups
Placebo
'Normals'
Analytical issues
ITT
Subgroup analysis
Losses to follow-up
Equivalence vs. traditional testing
Outcome selection
Surrogate endpoints
Composite endpoints
Trial duration
Interpretation of results
Causal inference
The media role in reporting RCT results

Ethical Issues

Consideration of ethical issues is key to the selection of the study design chosen for a given research question/hypothesis. For RCTs ethical considerations can be particularly problematic, mostly (but by no means solely) as it relates to using a placebo control. A full discussion of the ethics of clinical research is beyond the scope of this book, and for further discussion one should review the references noted here [13–15]. (There is also further discussion of this issue under the section entitled "Traditional vs. Equivalence Testing" and Chaps. 4 and 7). The opinions about when it is ethical to use placebo controls are quite broad. For example, Rothman and Michaels are of the opinion that the use of placebo is in direct violation of the Nuremberg Code and the Declaration of Helsinki [15], while others would argue that placebo controls are ethical as long as withholding effective treatment leads to no serious harm and if patients are fully informed. Most would agree that placebo is unethical if effective life-saving or life-prolonging therapy is available or if it is likely that the placebo group could suffer serious harm. For ailments that are not likely to be of harm or cause severe discomfort, some would argue that placebo is justifiable [14]. However, in the majority of scenarios, the use of a placebo control is not a clear-cut issue, and decisions need to be made on a case-by-case basis. One prevailing standard that provides a guideline for when to study an intervention against placebo is when one has enough confidence in the intervention that one is

comfortable that the additional risk of exposing a subject to the intervention is low relative to no therapy or the 'standard' treatment; but, that there is sufficient doubt about the intervention that use of a placebo or active control ('standard treatment') is justified. This balance, commonly referred to as *equipoise*, can be difficult to come by and is likewise almost always controversial. Importantly, equipoise needs to be present not only for the field of study (i.e. there is agreement that there is not sufficient evidence of the superiority of an alternative treatments), but equipoise also has to be present for individual investigators (permitting individual investigators to ethically assign their patients to treatment at random).

Another development in the continued efforts to protect patient safety is the Data Safety and Monitoring Board (DSMB-see Chap. 9). The DSMB is now almost universally used in any long-term intervention trial. First a data and safety monitoring plan (DSMP) becomes part of the protocol, and then the DSMB meets at regular and at 'as needed' intervals during the study in order to address whether the study requires early discontinuation. As part of the DSMP, stopping rules for the RCT will have been delineated. Thus, if during the study, either the intervention or control group demonstrates a worsening outcome, or the intervention group is showing a clear benefit, or adverse events are greater in one group vs. the other (as defined within the DSMP) the DSMB can recommend that the study be stopped. But, the early stopping of studies can also be a problem. For example, in a recent systematic review by Montori et al., the question was posed about what was known regarding the epidemiology and reporting quality of RCTs involving interventions stopped for early benefit [16]. Their conclusions were that prematurely stopped RCTs often fail to adequately report relevant information about the decision to stop early, and that one should view the results of trials that are stopped early with skepticism [16].

Randomization

Arguably, it is randomization that results in the RCT yielding the highest level of scientific evidence (i.e. resulting in the greatest likelihood that the intervention is causally related to the outcome). Randomization is a method of treatment allocation that is a distribution of study subjects at random (i.e. by chance). As a result, randomization results in all randomized units (e.g. subjects) having the same and independent chance of being allocated to any of the treatment groups, and it is impossible to know in advance to which group a subject will be assigned. The introduction of randomization to clinical trials in the modern era can probably be credited to the 1948 trial of streptomycin for the treatment of tuberculosis [17]. In this trial, 55 patients were randomized to either streptomycin with bed rest, and were compared to treatment with bed rest alone (the standard treatment at that time). To quote from that paper, 'determination *of whether a patient would be treated by streptomycin and bed rest (S case) or bed rest alone (C case), was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each center by Professor Bradford Hill; the details of the series were unknown*

to any of the investigators or to the co-coordinator and were contained in a set of sealed envelopes each bearing on the outside only the name of the hospital and a number. After acceptance of a patient by the panel and before admission to the streptomycin centre, the appropriate numbered envelope was opened at the central office; the card inside told if the patient was to be an S or C cases, and this information was then given to the medical officer at the centre'. Bradford Hill was later knighted for his contributions to science including the contribution of randomization.

With randomization the allocation ratio (number of units-subjects- randomized to the investigational arm versus the number randomized to the control arm) is usually 1:1. But a 1:1 ratio is not required, and there may be advantages to unequal allocation (e.g. 2:1 or even 3:1). The advantages of unequal allocation are: one exposes fewer patients to placebo, and one gains more information regarding the safety of the intervention. The main disadvantage of higher allocation ratios is the loss of power.

There are three general types of randomization: simple, blocked, and stratified. Simple randomization can be likened to the toss of an unbiased coin (i.e. heads group A, tails group B). This is easy to implement, but particularly with small sample sizes, could result in substantial imbalance (for example if one tosses a coin 10 times, it is not improbable that one could get 8 heads and 2 tails. If one tosses the coin 1,000 times it is likely that the distribution of heads to tails would be close to 500 heads and 500 tails). Blocked randomization (sometimes called permuted block randomization) is a technique common to multi-center studies. Whereas the entire trial might intend to enroll 1,000 patients, each center might only contribute 10 patients to the total. To prevent between center bias (recall each sample population has differences even if there is matching to known confounders) blocked randomization can be utilized. Blocked randomization means that randomization occurs within each center ensuring that about 5 patients in each center will be randomized to the intervention and 5 to the control. If this approach was not used, one center might enroll 10 patients to the intervention and another center, 10 patients to the control group. Recall that the main objective of randomization is to produce between-group comparability. If one knows prior to the study implementation that there might be differences that are not equally distributed between groups (again particularly more likely with small sample sizes) stratified randomization can be used. For example, if age might be an important indicator of drug efficacy, one can randomize within strata of age groups (e.g. 50–59, 60–69 etc.). Within each stratum, randomization can be simple or blocked.

In review, simple randomization is the individual allocation of subjects into the intervention and control groups, block randomization creates small groups (blocks) in which there are equal numbers in each treatment arm so that there are balanced numbers throughout a multi-center trial, and stratified randomization addresses the ability to separate known confounders into strata so that they can no longer confound the study results. Again, randomization is likely the most important key to valid study results because (if the sample size is large enough), it distributes known, and ***more importantly unknown,*** confounders equally to the intervention and control groups.

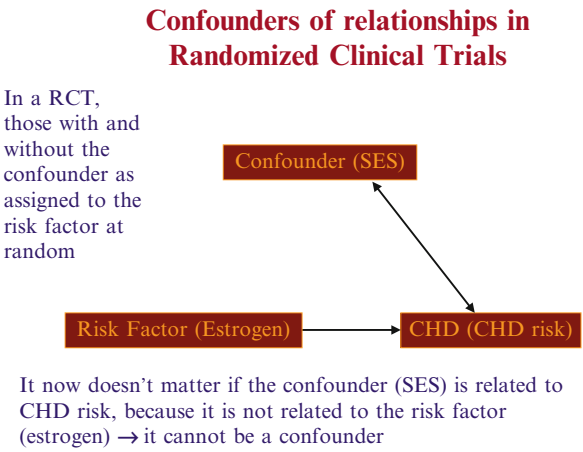

Fig. 3.1 The relationship of confounders to outcome and how they are eliminated in a RCT

Now, as to the problems associated with randomization. As prior discussed, the issue of confounders of relationships is inherent in all clinical research. A confounder is a factor that is associated with both the risk factor and the outcome, and leads to a false apparent association between the risk factor and outcome (see Fig. 3.1). In observational studies, there are several approaches to remove the effect of confounders:

- Most commonly used in case/control studies, one can match the case and control populations on the levels of potential confounders. Through this matching the investigator is assured that both those with a positive outcome (cases) and a negative outcome (controls) have similar levels of the confounder. Since, by definition, a confounder has to be associated with both the risk factor and the outcome; and, since through matching the suspected confounder is not associated with the outcome – then the factor cannot affect the observed differences in the outcome. For example, in a study of stroke, one may match age and race for stroke cases and community controls, with the result that both those with and without strokes will have similar distributions for these variables, and differences in associations with other potential predictors are not likely to be confounded, for example, by higher rates in older or African American populations.
- In all types of observational epidemiological studies, one can statistically/mathematically 'adjust' for the confounders. Such an adjustment allows for the comparison between those with and without the risk factor at a 'fixed level' of the confounding factor. That is, the association between the exposure and the potential confounding factor is removed (those with and without the exposure are assessed at a common level of the confounder), and as such the potential confounder cannot bias the association between the exposure and the outcome. For example, in a longitudinal study assessing the potential impact of hypertension on stroke risk, the analysis can 'adjust' for race and other factors. This adjustment implies that those with and without the exposure (hypertension) are assessed as if race were not associated with both the exposure and outcome.

Table 3.2 Example of the use of propensity scoring

	Before matching			After matching		
Variable (%)	Aspirin	No aspirin	P value	Aspirin	No aspirin	P value
Men	77	56	<.001	70.4	72.1	.33
Diabetes	16.8	11.2	<.001	15	15.3	.83
HTN	53	40.6	<.001	50.3	51.7	.46
CAD Hx	69.7	20.1	<.001	48.3	48.8	.79
CHF	5.5	4.6	.12	5.8	6.6	.43
B-Blocker	35.1	14.2	<.001	26.1	26.5	.79
ACE I	13	11.4	<.001	15.5	15.8	.79

Adapted from: Gum et al. [19]

The Propensity Score has received increased interest. The propensity score was introduced by Rosenbaum and Rubin [18] to provide an alternative method for estimating treatment effects when treatment assignment can be assumed to be unconfounded but is not random. A propensity score is the probability of a unit (e.g., person, classroom, school) being assigned to a particular condition in a study given a set of known covariates (a variable that is possibly predictive of the outcome under study). In an attempt to simulate randomization, propensity scores are used to reduce selection bias by equating groups based upon covariates (this, balances known confounders, but obviously not the unknown confounders). In the analysis of treatment effects, suppose that we have a binary treatment T, an outcome Y, and background variables X. The propensity score is defined as the conditional probability of treatment given background variables. This is operationalized by gathering all the background information that we have on subjects before exposure is known and building a model to predict the probability that they will be in the exposed vs. unexposed group. Groups of subjects with similar propensity scores can then be expected in the aggregate to have similar values of all the background information. Thus, propensity scores can be used in cohort trials, clinical trials without randomization, administrative data base studies, detecting safety signals, secondary questions within RCTs; and, propensity score analyses may be used in either the design or analysis phase. One example of the use of the propensity score is the aspirin and mortality study reported by Gum et al. [19]. In that study, 6,174 subjects underwent stress echocardiography for the evaluation of known or suspected coronary artery disease. Aspirin was being taken by 37 % of the subjects. The main outcome was all cause mortality and the mean follow-up was 3.1 years. In univariate analysis 4.5 % of the subjects receiving aspirin and 4.5 % of those not receiving aspirin died (HR 1.08, 0.85–1.39). Baseline characteristics were dissimilar in 25 of 31 of the covariates. In further analysis using matching by propensity score, 1,351 patients who were taking aspirin were at lower risk for death than 1,351 patients not using aspirin (4 % vs. 8 %, respectively; HR, 0.53; 95 % CI, 0.38–0.74; P = .002). After adjusting for the propensity for using aspirin, as well as other possible confounders and interactions, aspirin use remained associated with a lower risk for death (adjusted HR, 0.56; 95 % CI, 0.40–0.78; P < .001-Table 3.2). The patient characteristics associated with the most aspirin-related reductions in mortality were older age, known coronary artery disease, and impaired exercise capacity.

The major shortcoming with these aforementioned approaches is that one must know what the potential confounders are in order to match or adjust for them; and, it is the ***unknown confounders*** that represent a bigger problem. Another issue is that even if one suspects a confounder, one must be able to appropriately measure it. For example, socio-economic status (usually a combination of education and income) is a commonly addressed confounder; but, the definition of socio-economic status is an issue in which there is disagreement; and, which measures or cut-points to use is/are appropriate is frequently argued. The bottom line is that one can never perfectly measure all known confounders and certainly one cannot measure or perfectly match for unknown confounders. As mentioned, the strength of the RCT is that randomization (performed properly and with a large enough sample size) optimally balances both the known and unknown confounders between the interventional and control groups. But even with an RCT, randomization can be further compromised as will be discussed in some of the following chapters, and by the following example from "Student's" Collected Papers regarding the Lanarkshire Milk Experiment [20].

> *"Student" (i.e., the great William Sealy Gosset) criticized the experiment for it's loss of control over treatment assignment. As quoted: … Student's "contributions to statistics, in spite of a unity of purpose, ranged over a wide field from spurious correlation to Spearman's correlation coefficient. Always kindly and unassuming, he was capable of a generous rage, an instance of which is shown in his criticism of the* Lanarkshire *Milk Experiment. This was a nutritional experiment on a very large scale. For four months 5,000 school children received three-quarters of a pint of raw milk a day, 5,000 children the same quantity of pasteurized milk and 10,000 other children were selected as controls. The experiment, in Gosset's view, was inconclusive in determining whether pasteurized milk was superior in nutritional value to raw milk.*
>
> *This was due to failure to preserve the random selection of controls as originally planned. "In any particular school where there was any group to which these methods (i.e., of random selection) had given an undue proportion of well-fed or ill-nourished children, others were substituted to obtain a more level selection." The teachers were kind-hearted and tended to select ill-nourished as feeders and well-nourished as controls. Student thought that among 20,000 children some 200–300 pairs of twins would be available of which some 50 pairs would be identical-of the same sex and half the remainder nonidentical of the same sex. The 50 pairs of identicals would give more reliable results than the 20,000 dealt with in the experiment, and great expense would be saved. It may be wondered, however, whether Student's suggestion would have proved free from snags. Mothers can be as kind-hearted as teachers, and if one of a pair of identical twins seemed to his mother to be putting on weight…*

Missing Data

In 2008 the FDA requested that the National Research Council (NRC) convene an expert panel and to prepare a report that would be useful. The FDA that would address appropriate methods for analysis of missing data. Recall that the key feature of a RCT is the randomization process; and, this key feature is jeopardized when some of the outcome measures are missing. Missing data can seriously compromise the interpretations of clinical trials. A major source of missing data is the result of

Table 3.3a Eight ideas for limiting missing data in the design of clinical trials

Target a population that is not adequately served by current treatments and hence has an incentive to remain in the study
Include a run-in period (See discussion above regarding run-in periods)
Allow for a flexible treatment regimen
Shorten the follow up time so that participants are less likely to withdraw
Allow the use of rescue medications
For long term efficacy trials consider a withdrawal design
Consider an outcome that is not likely to lead to missing data
Consider add-on designs i.e. where a study treatment is added to existing therapies the patient may be receiving

Table 3.3b Eight ideas for limiting missing data in the conduct of clinical trials

Select investigators with good track records
Set acceptable rates for missing data and monitor during the course of the trial
Provide incentives to investigators and participants to continue the trial
Limit participant burden of data collection
Provide continued access to the trial medication after trial completion
Train investigators and study staff on the importance of trial continuation
Keep up to date contact information on trial participants
Assess the likelihood of participant continuation before enrollment

patients dropping out (discontinuing treatment) for any of a variety of reasons (adverse events, lost to follow up, lack of efficacy or tolerability etc). To the degree possible, these dropouts should be avoided, since there is no foolproof way to analyze data when there is significant (greater than 10 %?) missing data. Continuing to follow the patient after treatment discontinuation is one important step to reduce the degree of lost information. Little et al. summarized eight design ideas and eight ideas for the conduct of clinical trials for limiting missing data (Tables 3.3a and 3.3b [9]).

Since there is no universal method for handling missing data the best strategy is to avoid it. Statistical approaches to missing data will always involve unprovable assumptions, because there is always some uncertainty about the reasons that data is missing. The frequency of missing data is a result of patient dropouts the common reasons for which are: intolerability to the intervention, lack of intervention efficacy, or failure to attend designated appointments. Fleming has pointed out that there are only two reasons a patient can be off study; withdrawal of consent AND refusal to be followed or contacted, or the patient has achieved the required efficacy and safety end points [21]. He suggested six strategies to prevent missing data: first to distinguish nonadherence from nonretention; second to attempt to continue contact with the patient even if they have withdrawn from the study; third, adequately educate the patient during the informed consent process of the scientific relevance of the data they are providing; fourth, protocols should not give a false sense of being able to correct for missing data

with statistical approaches; fifth, protocols should specify targeted levels of data capture; and, sixth, forms and procedures for data collection should be formulated to reduce the likelihood of missing data. The use of run-in periods is an additional strategy, along with the use of flexible doses may be helpful as well.

Losses to follow-up (see below) using the last observation carried forward or baseline observation carried forward analysis is likely to overestimate and/or bias the outcome (since patients lost to follow-up more frequently are not benefitting from the intervention). Imputing the worst possible outcome might underestimate the benefit of the intervention. Missing data can be viewed in several ways. The ideal is if the missing data is "missing completely at random" (MCAR). This is an assumption that is unlikely to hold in most clinical trials because it presumes that the missing data are unrelated to the study variables (an unlikely scenario). A more realistic condition is missing at random (MAR, this might be better stated as missing "mostly" at random), or missing not at random (MNAR).

Because some missing data that does occur in almost every study, and each clinical trial has its own set of challenges, the NRC panel did list four general approaches: complete-case analysis, single imputation methods, estimating-equation methods and methods based on a statistical model. There is no single correct method for handling missing data, as all methods require that untestable assumptions be made. Discussion of these are beyond the scope of this book, but briefly, complete-case analysis simply excludes participants with missing data while with imputation, a single value is filled in for each missing value by using such methods as last observation carried forward or the baseline value carried forward. With estimating-equation methods, cases are weighted based upon the estimate of probability of an outcome being observed. As to the statistical modeling, approaches such as prior probabilities (Bayesian Methods) and multiple imputation where multiple sets of plausible values for missing data are used. Missing data can occur, of course, at random, or there can be differential loss of data, a more important consideration when missing data is assessed. Little et al. outlined six principles for drawing inferences from incomplete data [9].

1. Consider if the missing values are meaningful for analysis
2. Consider a possible causal pathway and how missing data might influence it
3. Consider why data are missing
4. Decide on a set of assumptions about the mechanism for missing data
5. Conduct a statistically valid analysis based on the above
6. Conduct a sensitivity analysis, a statistical technique that attempts to determine how changes in one variable will impact the target variable [7].

Complications of Eligibility Criteria

All generalizations are false, including this one (Mark Twain)

In every study there are substantial gains in statistical power by focusing the intervention in a homogenous patient population likely to respond to treatment, and to exclude patients that could introduce 'noise' by their inconsistent responses to treatment.

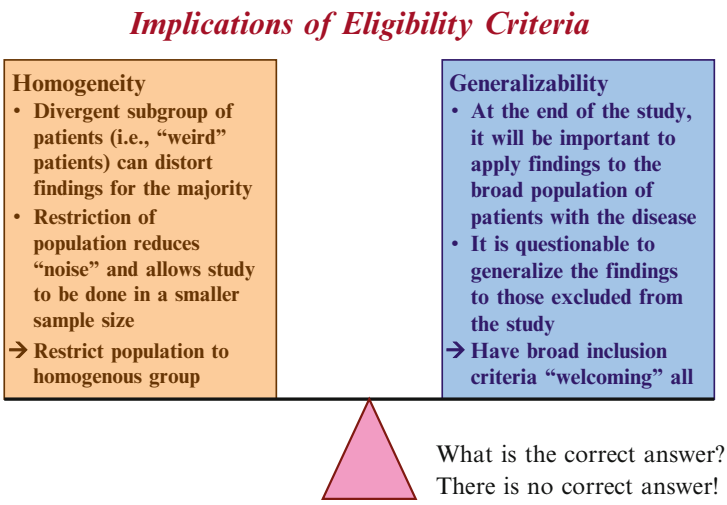

Fig. 3.2 The balance of conflicting issues involved with patient selection

Conversely, at the end of a trial there is a need to generalize the findings to a broad spectrum of patients who could potentially benefit from a superior treatment. These conflicting demands introduce the issue of balancing the inclusion/exclusion (eligibility criteria) such that the enrolled patients are as much alike as possible; but, on the other hand to be as diverse as possible in order to be able to apply the results to the more general population (i.e. generalizability). Figure 3.2 outlines this balance. What is the correct way of achieving this balance? There really is no correct answer, there is always a tradeoff between homogeneity and generalizability; and each study has to address this, given the availability of subjects, along with other considerations. This process of sampling represents one of the reasons that scientific inquiry requires reproducibility of results, that is, one study generally cannot be relied upon to portray 'truth' even if it is a RCT. The process of sampling embraces the concept of generalizability. The issue of generalizability is nicely portrayed in a video entitled 'A Village of 100' [22]. If one wanted to have a representative sample of the world for a study, this video (although predominately focused upon tolerance and understanding), is an excellent way of understanding the issue of generalizability. The central theme of the video asks the question 'if we shrunk the earth's population to a village of precisely 100 people, with all existing ratios remaining the same, what would it look like?' To paraphrase, if we maintained the existing ratios of the earth's population in a study of 100 people, what would our sample look like? The answer – there would be 57 Asians, 21 Europeans, 14 from the Western Hemisphere, 51 females and 49 males, 70 non-white and 30 white, 70 non-Christians and 30 Christians, 89 heterosexuals, 50 % of the worlds wealth would belong to 6 citizens of the USA, 80 would live in sub-standard housing, 70 would be unable to read (a potential problem with IRB approval), 50 would be malnourished, one would have a college education, and 4 would own a computer. When is the last time a study had a population representative of the Village of 100?

Table 3.4 Birmingham v Framingham: comparison of key variables

	Birmingham	Framingham
Population	242,800	62,910
% African American	73.5	5.1
Age		
25–44	30	35
45–64	20	33
65–>	14	13
Median income $	26,700	55,300
Education %		
<High school	25	13
High school	28	23
>High school	48	64
CVD rate	528–582	336–451

For an example of sampling issues, most of the major studies assessing the efficacy of the treatment of extracranial atherosclerosis with endarterectomy had excluded octogenarians on the basis that this patient population may have a response to the challenges of surgery that is different than their younger counterparts [23, 24]. Exclusion of these patients may have contributed to the successful completion of 'positive' trials (finding a benefit for the then new treatment – endarterectomy). However, now that the trials are complete, there is not 'level 5' evidence (data that is a result from RCTs) to guide the management of octogenarians with extracranial atherosclerosis, one of the subpopulations where the need for this information is important. In the absence of this information, thousands of endarterectomies are performed in this older patient population each year under the assumption that the findings from a younger cohort are generalizable to those at older ages. For another example, let's presume that in a multicenter trial that included Framingham Massachusetts, and Birmingham, Alabama, that a representative sample of each was recruited into a study. The makeup of the sample from each is illustrated in Table 3.4. As one can see, there are significant differences in the representative sample populations, and these differences could affect not only the success of the intervention but could also confound its relationship.

Efficacy vs. Effectiveness

Another limitation of RCTs is that they are designed to test safety and efficacy (i.e. does the drug work under optimal circumstances?) and not to answer questions about the effectiveness of a drug, the more relevant question for clinicians and economic analysts (i.e. does the drug work under ordinary circumstances of use?). Thus, the increased use of effectiveness trials has been suggested, to more closely reflect routine clinical practice. Effectiveness trials use a more flexible dosage regimen, and generally a 'usual care' comparator instead of a placebo comparator. Two approaches to this more 'real world trial' is the phase 4 trial (see Chap. 5) or

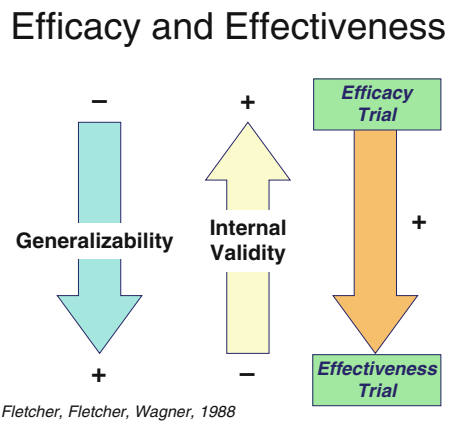

Fig. 3.3 The "Trade-off" between efficacy vs. effectiveness

the prospective, randomized, open-label, blinded end-point – PROBE-Trial. The PROBE Trial is further discussed in the next section entitled "Degree of Masking"). As to phase 4 trials, they are surrounded by some controversy as well. Figure 3.3 compares efficacy and effectiveness trials in terms of some of their more important variables.

Patient Compliance

Run-in Periods

Another issue surrounding RCTs, and one that is almost unique to clinical trials, is the use of run-in periods and their impact on who is eligible to be randomized. Pre-randomization run-in periods are frequently used to select or exclude patients in clinical trials, but the impact of run-in periods on clinical trial interpretation and generalization has not been systematically studied. The controversy regarding run-in periods also addresses the issue of efficacy vs. effectiveness, as the run-in period allows one to exclude patients that are potentially less compliant, or do not tolerate placebo (or whatever other intervention is used in an active comparison group). Although this issue has not been systematically studied, intuitively one can see that the potential for over-estimating the impact of an investigational drug is present when run-in periods are utilized, as the run-in period will likely exclude patients from the study who would not have ideally responded.

A study can achieve high compliance in at least three general ways: designing a simple protocol (complexity makes compliance more difficult); the use of compliance aids such as automatic reminders, telephone calls, calendars, etc; or by selecting subjects based upon pre-study or pre-randomization compliance. Of course, high compliance is a desirable characteristic of any research. High compliance attenuates the argument of whether to use intention to treat vs. compliance only as the primary

analysis. Also, high compliance will optimize the studies power as the "diluting" effect of non-compliers will not be manifest (all other things being equal). While the run-in period increases the proportion of compliers in the trial, it may introduce important differences in the outcomes, particularly if compliers and non-compliers are inherently different in the way they would respond to the intervention of interest. Thus, the effect of run-in periods on generalizability should be considered carefully before implementation. Lang [25] has listed some recommendations for helping to decide whether to use a run-in as part of a clinical trial, including:

1. consider a run-in whenever the contact between study staff and participants is low
2. consider a run-in period for a primary prevention trial because compliance is likely to be more difficult compared to therapeutic trials
3. For any trial, list the key features of the study protocol and see which features compliance could be directly tested prior to randomization
4. before using active agents during a run-in, consider both the expected frequency of occurrence of side effects and the postulated effect of the agent on the outcome of interest
5. all trials can use any available pre-randomization period for the simultaneous purpose of characterizing patients and evaluating compliance, whether of not the compliance information will be used for exclusions

In fairness, as Franciosa points out, clinicians use variants of run-in periods to treat their patients, such as dose titration, or challenge dosing (e.g. using small doses of ACE Inhibitors to rule out excessive responders) [26]. Pablos-Mendez et al. analyzed illustrative examples of reports of clinical trials in which run-in periods were used to exclude non-compliant patients, placebo responders, or patients that could not tolerate or did not respond to active drug [27].

Thus, the use of run-in periods is another reason that the results of RCTs may not accurately portray what the drugs overall effectiveness will be. What can be said is that there does need to be more focus on the details of run-in periods, and as is true of most things the researcher does in designing and implementing a clinical trial, judgments have to be made regarding the best approach to use regarding inclusions and exclusions, as well as judging what the impact of the run-in period is on the ultimate interpretation of a clinical trial. Ultimately, from the perspective of internal validity, it is better to exclude participants before randomization than have participants lost to follow up, cross between study groups, or become non-adherent to intervention protocols after randomization.

Recruitment and Retention

Nothing is more critical to the success of a clinical trial than the recruitment and retention of subjects. As will be discussed in more detail in Chap. 8, there are a number of reasons for failure of the recruitment process including: delayed start-up, and inadequate planning, In terms of patient/subject retention, there are arguably

differences in the handling of clinical patients in contrast to research subjects (although this could and perhaps should be challenged). Losses-to-follow-up need to be kept to a minimum and is discussed later in this chapter.

Degree of Masking (Blinding)

Although the basic concept of clinical trials is to be at equipoise, this does not change the often pre-conceived 'suspicion' that there is a differential benefit of the investigational therapy (e.g. the investigational drug is better than placebo). Thus, if study personnel know the treatment assignment, there may be differential vigilance where the supposed 'inferior group' is more intensively monitored (e.g. 'are you certain you have not had a problem?' they might ask). In this case, unequal evaluations can provide unequal opportunities to differentially 'discover' events. This is why the concept of double-blinding (masking) is an important component of RCTs. There is an argument about which term-blinding or masking-is most appropriate [28], and Fig. 3.4 portray's a humorous example of this argument. But, one cannot always have a double-blind trial, and some would argue that double-blinding distances the trial from a 'real-world' approach. An example where blinding is difficult to achieve might be a surgical vs. medical intervention study where post operative patients may require additional follow-up visits, and each visit imparts an additional opportunity to elicit events. That is, it has been said that 'the patient

Fig. 3.4 A humorous example of blinding (masking) (With permission from Schulz and Grimes [28])

cannot have a fever if the temperature is not taken,' [29] and for RCTs, events cannot be detected without patient contact to assess outcomes.

Of course, masking is not always possible and examples include studies that: might evaluate residual surgical wounds, studies involved with cycling hormone replacement, studies requiring serum (or other) assay or physical measurement, studies that involve participant participation in the treatment (i.e., low fat diet, exercise, etc). Common approaches to these examples are to at least mask the rater (adjudicator), or to move toward a totally objective outcome (e.g. death), or to use an independent observer who does not know treatment to assess outcome. In an effort to study the impact of adjudicator blinding on outcomes, Parmar et al. assessed the effect of blinding race and geography on outcomes ascertainment in an observational study [28]. The primary characteristics of interest were race and geography, and the prespecified acceptable agreement rate between adjudicators was set at >80 %. They selected 116 suspected cardiovascular events that underwent adjudication with usual blinding. At least 3 months later, cases were readjudicated without blinding race and geographic location of the patient, and differences in outcomes ascertainment was assessed using Cohen's κ statistic and agreement rates. Agreement between the blinded and unblinded reviews was good to excellent for all four outcomes. κ statistics were 0.80 (chest pain), 0.85 (heart failure), 0.86 (revascularization) and 0.74 (MI) ($p<0.0001$ for all). Within each outcome, agreement rates were similar for race and geographic groups (agreement 83–100 %). The authors concluded that in observational studies, blinding medical record review for outcomes ascertainment for some types of patient characteristics may be an unwarranted expense.

In order to realize a more 'real-world' approach to clinical trials, the prospective randomized open-label blinded endpoint design (PROBE design) was developed. Randomization is used so that this important component of study design is retained. By using open-label therapy, the drug intervention and its comparator can be clinically titrated as would occur in a doctor's office. Of course, blinding is lost here, but only as to the therapy. In a PROBE design, blinding is maintained as to the ascertainment of the outcome. To test whether the use of open-label vs. double-blind therapy affected outcomes differentially, a meta analysis of PROBE trials and double-blind trials in hypertension was reported by Smith et al. [30]. They found that changes in mean ambulatory blood pressure from double-blind controlled studies and PROBE trials were statistically equivalent.

Selection of Comparison Groups

As the story goes a clinical researcher meets someone on the street who asks "how do you do?" The researcher answers "compared to whom?" When addressing the validity of an outcome difference compared to some control group, it is crucial that the control group be clearly defined. Sometimes studies assess a new (investigational) treatment versus an approved (standard) active treatment (i.e. to assess if the old

'standard' treatment should be replaced with the new treatment), in other cases, studies are assessing if a new treatment should be added (not replacing, but rather supplementing), current treatment. In this latter case, the comparison of interest is the outcome of patients with and without the new treatment. In this instance, masking can only be accomplished by the use of a double-blind technique. Traditionally, placebo treatment has been used as the comparator to investigational treatments, and has been one of the standards of clinical trials.

The use of the placebo comparator has more and more been the subject of ethical concerns. In addition to ethical issues involved with the use of placebos, there are other considerations raised by the use of placebo-controls. For example, an important lesson was learned from the Multiple Risk Factor Intervention Trial (MRFIT) regarding the use and analysis of the placebo control group, which might best be summed up with the question 'why it is important to watch the placebo group?' [31]. MRFIT screened 361,662 patients to randomize high-risk participants (using the Framingham criteria existent at that time) to special intervention (n = 6428) and usual care (n = 6438) with coronary heart disease mortality as the endpoint. The design of this well-conducted study assumed that the risk factor profile of those receiving 'special treatment interventions' would improve, while those patients in the 'usual care' group would continue their current treatments and remain largely unaffected. The special intervention approaches in MRFIT were quite successful, and all risk factor levels were reduced. However, there were also substantial and significant reductions observed in the control group. That both treatment and control groups experienced substantial improvements in their risk factor profile translated to almost identical CHD deaths during the course of the study. Why did the control group fare so well? Several phenomena may have contributed to the improvement in the placebo-control group. First, is the Hawthorne effect, which suggests that just participating in a study is associated with increased health awareness and changes in risk factor profile, irrespective of any intervention [32]. In addition, for the longer-term trials, there are changes in the general population that might alter events. For example, randomization in MRFIT was conducted during the 1980s, a period when health awareness was becoming more widely accepted in the USA, and likely beneficially affected the control group.

Although the ethics of placebo controls is under scrutiny, another principal regarding the placebo-control group is that sometimes being in the placebo group isn't all that bad. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study was launched in 1994 [33]. By the early 1990s there was mounting clinical epidemiologic evidence of reduced cancer risk associated with a higher intake of antioxidants. Treatment with vitamin E and beta carotene were considered unlikely to be harmful, and likely to be helpful; and, the question was asked whether antioxidants could reduce lung cancer-even in smokers. A double-blind, placebo-controlled RCT was launched with a 2 x 2 factorial design (see Chap. 4), and over 7,000 patients in each cell. No benefit was seen with either therapy, but compared to placebo; a disturbing worsening trend was observed in the beta-carotene treated compared with the placebo group.

Table 3.5 Different definitions of "normal"

Property	Term	Consequence of application
Distribution shape	Gaussian	Minus values
Lie within preset %	Percentile	Normal until workup
No additional risk	Risk factor	Assumes altering risk factor improves risk
Societal or political	Culturally desirable	Raises the role of society in medicine
A range before test suggests no disease	Diagnostic	Need to know the predictive value in ones own practice
Therapy beneficial	Therapeutic	New therapies alter this

Frequently, the comparison group or control group is a so called 'normal' population. Inherent to this concept is 'what is normal?'. A wit once opined that 'a normal person is one who is insufficiently tested'. Interestingly, there are a number of scientific definitions of normal (see Table 3.5). One definition of normal might be someone who fits into 97 % of a Gaussian Distribution, another that normal lies within a preset percentile of a laboratory value or values. Other definitions exist, suffice it to say, whatever definition is used it needs to be clearly identified.

Analytic Approach

Intention to Treat and Per-Protocol Analysis

There are three general analytic approaches to clinical trials; intention-to-treat (ITT) analysis (or analysis as randomized), compliers only (or per-protocol) analysis, and analysis by treatment received. Probably the least intuitive and the one that causes most students a problem is ITT. ITT was derived from a principle called the pragmatic attitude [34]. The concept was that one was to compare the effectiveness of the *intention* to administer treatment A vs. the *intention* to administer treatment B, i.e. the comparison of two treatment policies rather than a comparison of two specific treatments. With ITT, everyone assigned to an intervention or control arm is counted in their respective assigned group, whether they ultimately receive none of the treatment, or somewhat less than the trial directed. For example, if in a 1-year trial, a patient is randomized to receive an intervention, but before the intervention is administered, they drop out (for whatever reason) they are analyzed as if they received the treatment for the entire year. The same applies if the patient drops out at any time during the course of the study. Likewise, if it is determined that the patient is not fully compliant with treatment, they are still counted as if they were. In fact, whether there is compliance, administrative, or protocol deviation, patients once randomized are counted as if they completed the trial. Most students initially feel that this is counter-intuitive. Rather the argument would be that one is really interested in what would happen if a patient is randomized to a treatment arm and they take that treatment for the full trial duration and are fully compliant – this, one

would argue, gives one the real information needed about the optimal effect of an intervention (this, by the way, is a description of the compliers only analysis). So why is ITT the scientifically accepted primary analysis for most clinical trials? As mentioned before, randomization is arguably one of the most important aspects of clinical trial design. If patients once randomized to a treatment are not included in the analysis, the process of randomization is compromised. It is not a leap of faith to wonder if patients dropping out of the intervention arm might be different than the patients dropping out of a control arm. Thus, if ITT is not used, one loses the assurance of equal distribution of unknown confounders between the treatment groups, and this thereby tarnishes the basis of randomization. One example of the loss of randomization if ITT is not used might be differential dropouts between the intervention and control arm for adverse events. Also, if patients with more severe disease are more likely to dropout from the placebo arm; or conversely patients who are older, dropout more frequently from the placebo arm thereby removing them from the analysis, this could result in an imbalance between the two comparison groups. Another argument for ITT is that it provides for the most conservative estimate of the intervention effect (if the analysis includes patients that did not get the entire treatment regimen and the regimen is beneficial, clearly the treatment effect will be diluted). Thus, if using ITT analysis reveals a benefit, it adds to the credibility of the effect measure. Of course, one could argue that one could miss a potentially beneficial effect if the intervention effect is diluted. In summary, ITT protects against bias, protects the statistical integrity of the trial, and protects the randomization process.

In the compliers only analysis, the patients that complete the trial and comply fully with that treatment are analyzed. The problem is that if a beneficial effect is seen, one can wonder what the loss of randomization (and thereby equality of confounders between groups) means to that outcome, particularly if an ITT analysis does not demonstrate a difference. The loss of randomization and the loss of balanced confounders between the treatment and control groups is exemplified by an analysis of the Coronary Drug Project, where it was determined that poor compliers to placebo had a worse outcome than good compliers to placebo [35]. This would suggest that there are inherent differences in patients who comply vs. those who do not, and this could differentially be the cause of dropout. The Coronary Drug Project was a trial aimed at comparing clofibrate with placebo in patients with previous myocardial infarction with the outcome of interest being mortality. Initially reported as a favorable intervention (there was a 15 % 5 year mortality in the clofibrate compliers only analysis group, compared to a 19.4 % mortality in the placebo group-$p < .01$); while with ITT analysis there was essentially no difference in outcome (18.2 vs. 19. 4 %−$p < .25$). Given the differences in outcome between placebo compliers and placebo non-compliers, one can only assume the same for the investigational drug group. Likewise, the Anturane Reinfarction Trial was designed to compare anturane with placebo in patients with a prior MI and in whom mortality was the outcome of interest [36]. One thousand six hundred and twenty nine patients were randomized 817 to placebo and 812 to anturane (71 patients were later excluded because it was determined that they did not meet eligibility criteria).

The study initially reported anturane as a favorable intervention (although the $p < .07$), but when the 71 ineligible randomized patients were included in the analysis the $p = 0.20$. Again further analysis demonstrated that in the anturane ineligible patients, overall mortality was 26 % compared to the mortality in the anturane eligible patients that was 9 %.

If one considers the common reasons for subjects not being included in a study, ineligibility is certainly one. In addition, subjects may be dropped from a trial for poor compliance, and/or adverse drug events; and, patients may be excluded from analysis due to protocol deviations or being lost to follow up. Some of the reasons for ineligibility are protocol misinterpretations, clerical error, or wrong diagnosis at the time of randomization. Sometimes the determination of ineligibility is above question (e.g. the patient fell outside of the studies predetermined age limit) but frequently ineligibility requires judgment. The Multicenter Investigation of the Limitation of infarct Size (MILIS) study is an example of this latter concept. MILIS compared propranolol, hyaluronidase, and placebo in patients with early acute MI, in order to observe effects on mortality. Subsequently, some patients were deemed ineligible because the early diagnosis of MI was not substantiated. But, what if the active therapy actually had an effect on preventing or ameliorating the MI? The problem with not including patients in this instance is that more patients could have been withdrawn from the placebo group compared to the active therapy group, and as a result, interpretation of the data would be altered.

Of course, as is true of most things in clinical research there is not just one answer, indeed, one has to carefully assess the trial specifics. For example, Sackett and Gent cite a study comparing heparin to streptokinase in the treatment of acute myocardial infarction [37]. The ITT analysis showed that streptokinase reduced the risk of in-hospital death by 31 % ($p = 0.01$). However, eight patients randomized to the heparin group died after randomization, but before they received the heparin. Analysis restricted to only those who received study drug decreased the benefit of streptokinase (and increased the p value).

In summary, ITT is the most accepted (e.g. by most scientists and the FDA) as the analysis of choice for clinical trials. This is because ITT assures statistical balance (as long as randomization was properly performed), it 'forces' disclosure of all patients randomized in a trial, and most of the arguments against ITT can be rationally addressed.

Analysis-As-Treated is another analytic approach that addresses not the group to which the patient was randomized and not compliers only, but what the patient actually received. This analytic approach is utilized most often when patients cross over from one treatment arm to the other; and, this occurs most often in surgical vs. medical treatment comparisons. For example, patient's randomized to medical treatment (vs. coronary artery bypass surgery) might, at some time during the study, be deemed to need the surgery, and are thus crossed over to the surgical arm and are then assessed as to the treatment they received (i.e. surgery). Like compliers only analysis, this might be an interesting secondary analytic technique, but shares many of the same criticisms discussed earlier for compliers-only analysis. In addition, because such trials cannot easily be double-blinded, even greater criticism can be

leveled against this analytic approach compared to compliers-only analysis. In addition, statistical testing with this analysis by treatment received, is more complicated, not only by the crossovers, but by the inherent nature of the comparison groups. In comparison trials of 1 drug vs. placebo, for example, it is reasonable to assume that if the drug is superior to placebo (or an active control) patients in the drug group will average fewer events in the follow-up period. When this is displayed as survival curves, the survival curves will increasingly separate. In trials comparing surgical to medical therapy, the aforementioned approach may not be reasonable. For example, if patients randomized to surgery have a high early risk (compared to the non-surgical group) and a lower risk later, these risks may cancel and be similar to the number of events under the null hypothesis of no difference between groups. The issue of comparing surgical and non-surgical therapies in clinical trials has been nicely summarized by Howard et al. [38].

Subgroup Analysis

As pointed out by Assmann et al., most clinical trials collect substantial baseline information on each patient in the study [39]. The collection of baseline data has at least four main purposes: (1) to characterize the patients included in the trial, i.e. to determine how successful randomization was (2) to allow assessment of how well the different treatment groups are balanced, (3) to allow for analysis per treatment, (4) to allow for subgroup analysis in order to assess whether treatment differences depend on certain patient characteristics. It is this 4th purpose that is perhaps the most controversial because it can lead to 'data dredging' or has some wits have opined, 'if you interrogate the data enough, you can get it to admit to anything'. For example, Sleight and colleagues, in order to demonstrate the limitations of subgroup analysis, performed subgroup analysis in the ISIS-2 trial by analyzing treatment responses according to the astrological birth sign of the subject [40]. This analysis suggested that the treatment was quite effective and statistically significant for all patients except those born under the sign of Gemini or Libra. The validity of any subgroup observation tends to be inversely proportional to the number of subgroups analyzed. For example, for testing at the 5 % significance level ($p \leq .05$) an erroneous statistically significant difference will be reported (on average) 5 % of the time (i.e. false + rate of 5 %). But, if 20 subgroups are analyzed, the false positive rate would approach 64 % (Table 3.6).

It is true, that meaningful information from subgroup analysis is restricted by multiplicity of testing and low statistical power and that surveys on the adequacy of the reporting of clinical trials consistently find the reporting of subgroup analyses to be wanting. Most studies enroll just enough participants to ensure that the primary efficacy hypothesis can be adequately tested, and this limits the statistical ability to find a difference in subgroup analyses; and, the numbers of subjects available for subgroup analysis is further compounded by loss of compliance, the need for adjustments for multiple testing, etc. Some have taken this to mean that subgroup

Table 3.6 Approximate number of False Positives (FP) occurring with multiple subgroup analyses

No. of tests	Probability of 1 FP	Probability of 2 FPs	Probability of 3 FPs
1	0.05	0.01	0
2	0.10	0.02	0
3	0.14	0.025	0
5	0.23	0.03	0
10	0.40	0.05	0.01
20	0.64	0.10	0.10

analyses are useless. When results from a subgroups analysis are at variance from the overall group outcome, the results are still likely to be true if the subgroup is large, they are pre-specified rather than *post hoc* (i.e. 'after the fact') and they are of limited number (not all post hoc analyses are subgroup analyses, but arguably most are). At the least, whether pre-specified or *post hoc,* subgroup analyses serve to generate questions for subsequent trials, and should not be interpreted as "truth". An exception to this latter principal, is when it comes to safety, here subgroup analyses might "carry more weight". An example of a post-hoc analysis that was "accepted" is the Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) study where LIPITOR 80 mg vs. placebo was administered in 4,731 subjects without CHD who had a stroke or TIA within the preceding 6 months [41]. A higher incidence of hemorrhagic stroke was seen in subgroup analysis in the LIPITOR 80 mg group compared to placebo. Subjects with hemorrhagic stroke on study entry appeared to be at increased risk for hemorrhagic stroke. As a result, Pfizer revised the US Prescribing Information for atorvastatin to include a precaution for its use of 80 mg in patients with a prior history of stroke.

What can be said is that if subgroup analysis is used and interpreted carefully, it can be useful. Even among experts, opinions range from only accepting pre-specified subgroup analyses supported by a very strong *a priori* biological rationale, to a more liberal view in which subgroup analyses, if properly carried out and interpreted, are permitted to play a role in assisting doctors and their patients to choose between treatment options. In reviewing a report that includes subgroup analyses, Cook et al. suggest addressing the following issues (Table 3.7): (1) were the subgroups appropriately defined, (that is, be careful about subgroups that are based upon characteristics measured after randomization e.g. adverse drug events may be more common as reasons for withdrawal from the active treatment arm whereas lack of efficacy may be more common in the placebo arm); (2) were the subgroup analyses planned before the implementation of the study (in contrast to after the study completion or during the conduct of the study); (3) does the study report include enough information to assess the validity of the analysis e.g. the number of subgroup analyses; (4) do the statistical analyses use multiplicity and interaction testing; (5) were the results of subgroup analyses interpreted with caution; (6) is there replication of the subgroup analysis in another independent study; (7) was a dose-response relationship demonstrated in the subgroup; (8) was there reproducibility of the observation within individual sites; and (9) is there a biological explanation.

Table 3.7 Considerations regarding subgroup analyses

Was there potential for patient misclassification
Was the analysis approach Intention-To-Treat
Were subgroups planned *a priori*
Was the subgroup analysis based on trial or biological data
Was there adequate power for subgroup analysis
What are the total number of subgroups analyzed
Are there adjustments for multiple testing
Are there tests for interaction
Are subgroup results emphasized above primary analyses
Are the subgroup analyses placed in proper biological and prior trial data perspective
Are *a priori* analyses distinguished from *a posteriori* analyses

Table 3.8 Goal of RCTs and their relation to hypothesis testing

RCT goal	Superiority	Equivalence
Null hypothesis	New = Old	New < Old + δ
Alternative hypothesis	New $\gtrless$ Old	New = Old + δ

δ is the margin in which the point estimate falls

Traditional Versus Equivalence Testing (Table 3.8)

Most clinical trials have been designed to assess if there is a difference in the efficacy to two (or more) alternative treatment approaches (with placebo usually being the comparator treatment). There are reasons why placebo-controls are preferable to active controls, not the least of which is the ability to distinguish an effective treatment from a less effective treatment. However, if a new treatment is considered to be equally effective but perhaps less expensive and/or invasive, or a placebo-control is considered unethical, then the new treatment needs to be compared to an established therapy and the new treatment would be considered preferable to the established therapy, even if it is just as good (not necessarily better) as the old (Table 3.9). The ethical issues surrounding the use of a placebo-control and the need to show a new treatment to only be as 'good as' (rather than better) has given rise to the recent interest in equivalence or non-inferiority testing. With traditional (superiority) hypothesis testing, the null hypothesis states that 'there is no difference between treatment groups (i.e. New = Old or placebo or standard therapy). Rejecting the null, then allows one to definitively state if one treatment is better (or worse) than another (i.e. New > or < Old). The disadvantage is if at the conclusion of an RCT there is not evidence of a difference, one cannot state that the treatments are the same, or as good as one to the other, only that the data are insufficient to show a difference. That is, when the null hypothesis is not accepted, it is simply the case where it cannot be rejected. The appropriate statement when the null hypothesis is not rejected (accepted) is 'there is not sufficient evidence in these data to establish if a difference exists.'

Table 3.9 Reasons for choosing noninferiority over superiority designs

Comparing new treatment with active control instead of placebo	Unethical to use placebo group in controlled study when there's an established treatment
New treatment not better in primary end point; better in secondary end points	Although no difference between primary efficacy outcomes, difference in secondary end points such as adverse events, quality of life
New treatment not better in primary end point; overall efficiency is better	Non-inferiority in effectiveness and safety; clear superiority in incurred cost produces and overall efficiency
The new treatment can be non-inferior and superior	Non-inferiority testing can be complemented by superiority testing in one study without need for adjustments

Equivalence testing in essence 'flips' the traditional null and alternative hypotheses. Using this approach, the null hypothesis is that the new treatment is worse than the old treatment (i.e. New < Old); that is, rather than assuming that there is no difference, the null hypothesis is that a difference exists and the new treatment is inferior. Just as in traditional testing, the two results available from the statistical test are (1) reject the null hypothesis, or (2) failure to reject the null hypothesis. However, with equivalence/noninferiority testing rejecting the null hypothesis is making the statement that the new treatment is not worse than old treatment, implying the alternative, that is 'that the new treatment is **as good** as the old' (i.e. New = Old). Hence, this approach allows a definitive conclusion that the new treatment is as good as the old.

One caveat is the definition of 'as good as,' which is defined as being in the 'neighborhood' or having a difference that is so small that it is to be considered clinically unimportant (generally, effects within ±2 % – this is known as the equivalence or noninferiority margin usually indicted by the symbol δ). The need for this 'neighborhood' that is considered 'as good as' exposes the first shortcoming of equivalence testing – having to make a statement that 'I reject the null hypothesis that the new treatment is worse than the old, and accept the alternative hypothesis that it is as good – *and by that I mean that it is within at least 2 % of the old*' (the wording in italics are rarely included in the conclusions of a manuscript). A second disadvantage of equivalence/noninferiority testing is that no definitive statement can be made that there is evidence that the new treatment is better or worse. Just as in traditional testing, one never accepts the null hypothesis – one only fails to reject it. Hence if the null is not rejected, all one can really say is that there is ***insufficient evidence in these data*** that the new treatment is as good as the old treatment. Another problem with equivalence/noninferiority testing is that one has to rely on the effectiveness of the active control obtained in previous trials, and on the assumption that the active control would be equally effective under the conditions of the present trial.

An example of an equivalence trial is the Controlled ONset Verapamil INvestigation of Cardiovascular Endpoints study (CONVINCE), a trial that also raised some ethical issues that are different from those usually involved in RCT's [42]. CONVINCE was a large double-blind clinical trial intended to assess the equivalence of verapamil and standard therapy in preventing cardiovascular disease-related events in hypertensive patients. The results of the study indicated that the verapamil

preparation was not equivalent to standard therapy because the upper bound of the 95 % confidence limit (1.18) slightly exceeded the pre-specified boundary of 1.16 for equivalence. However, the study was stopped prematurely for commercial reasons. This not only hobbled the findings in terms of inadequate power, it also meant that participants who had been in the trial for years were subjected to a 'breach in contract'. That is, they had subjected themselves to the risk of an RCT with no ultimate benefit. There was a good deal of criticism borne by the pharmaceutical company involved in the decision to discontinue the study early. Parenthetically, the company involved no longer exists.

In the past, some separated equivalence testing and non-inferiority testing. The question posed by non-inferiority testing being slightly different in that one is asking whether the new intervention is simply not inferior to the comparator (i.e. New $\not<$ Old). One potential advantage of this approach is that statistical significance could be only 'one-tailed' since there is no implication that the analysis is addressing whether the new treatment is better or as good as, only that it is not inferior. There is a good deal of disagreement regarding this latter issue, so that most use the two (equivalence and noninferiority) approaches interchangeably. Weir et al. utilized the non-inferiority approach in evaluating a comparison of valsartin/hydrochlorthiazide (VAL/HCTZ) with amlodipine in the reduction of mean 24-h diastolic BP (DBP) [43]. Noninferiority of the VAL/HCTZ combination to amlodipine was demonstrated, and fewer adverse events were noted with the combination treatment as well. The null hypothesis for this analysis was that the reduction in mean 24-h DBP from baseline to the end of the study with VAL/HCTZ was $\geq$3 mmHg less (the non-inferiority margin) compared with amlodipine. Again, a caveat has been recently raised by LeHenanff et al. and Kaul et al. [44, 45]. LeHananff et al. [45] reviewed studies published between 2003 and 2004 that were listed as equivalence or noninferiority, and noted a number of deficiencies, key among them being the absence of a stated equivalence or non inferiority margin [45].

Equivalence/non-inferiority trials are further discussed in Chap. 4.

Losses to Follow Up (See also Discussion of Missing Data, Above)

Patients who are lost-to-follow-up are critical in clinical trials and are particularly problematic in long-term trials. Patients lost to follow-up might be regarded as having had poor results (that is assumed that they experienced treatment failure); so if there are sufficient numbers of them, trial results can be skewed to less of an effect, even if, in truth, they did not have poor results. If, in the different study arms, there are equal numbers lost to follow-up, and they are lost for the same reasons, lost to follow up would not be as critical, but this is unlikely to occur. Section 4.3.4 of the ICH E-6 Good Clinical Practice: Consolidated Guidance reads, "*Although a subject is not obliged to give his/her reason(s) for withdrawing prematurely from a trial, the investigator should make a reasonable effort to ascertain the reason(s),*

while fully respecting the subject's rights" This excerpt expresses the need for physicians associated with clinical research trials to make a first-hand effort to contact patients who are lost-to-follow-up. In doing so pharmaceutical companies not only look out for the best interest of the patients who enroll in their clinical research trials, but also protect the data outcome of their clinical trials.

Of course, in ITT analysis, patient's lost-to-follow-up is still counted, but the argument is how to count them. Some would argue that it is appropriate to count them as poor outcomes since this will give the most conservative result, while others argue that since their outcome is not known, they should not be counted. In fact, there is little data reported on the actual impact on a study result of patients lost to follow up. In one study, Joshi et al. did address this issue in a long-term follow-up (up to 16 years of follow-up) of patients who had undergone knee arthroplasty. With the concerted effort of full-time personnel and a private detective, all 123 patients initially lost to follow-up were traced. Patients cited a variety of reasons why they did not attend follow-up visits, including: change of residence, inability to travel, displeasure with the physician or staff, financial constraints, satisfaction with the results so that they did not feel follow-up was necessary and poor results. They also found that more women than men were lost to follow-up. A few companies have developed methods of locating and contacting patients that are lost-to-follow-up and processes of handling patient information. These are options that pharmaceutical companies can use to find patients that have become lost-to-follow-up. These lost to follow-up patient locate systems use customized programmed software systems, as well as highly customized research and communication processes.

Surrogate Endpoints

The choice of an outcome is seemingly easy and apparent. For example, mortality is the dominant concern for many situations, and is seldom a difficult outcome to ascertain, unless there is a high loss to follow-up, which should not be a problem if the study is designed properly. However, if all cause death is the outcome this principal holds, if the determination is the specific reason for death, it becomes decidedly more difficult. This difficulty is because many deaths occur either outside the hospital where one has to rely on death certificates as the cause of death, or in hospital, where many patients have multi-organ disease, and trying to parse the specific cause is likely to be difficult. And yet, ascertaining the cause of death is essential for classifying disease-specific mortality in clinical research studies. As mentioned, death certificates often serve as the source of this information with the recognition that the cause of death on the death certificate is often fraught with misclassification (in fact in some states in the US the cause of death is not even entered). The potential for bias from this misclassification, and the fact that obtaining death certificates can often be time consuming and labor intensive is problematic. As a result, many studies also use a proxy–reported statement to determine the cause of death. Halanych et al. [46], assessed the validity of proxy-reported causes of death

Table 3.10 Approximate sample size given the treatment effect and control group "outcome"

	Treatment effect			
Rate in control group (%)	10 %	20 %	30 %	50 %
2	100,000	25,000	10,000	3,000
10	65,000	15,000	6,000	2,000
50	2,100	518	225	80

in 336 participants of the REGARDS Study. Trained experts used study data, medical records, death certificates, and proxy reports to adjudicate deaths. Adjudicated cause of death had a higher rate of agreement with proxy reports (73 %; Cohens kappa = .69) then with death certificates (63 % kappa = .54). Using the adjudicator cause of death as the "gold standard", the sensitivity for proxy reports was 50–89 % (depending on the cause) and specificity; 94–98 %, compared to death certificates, sensitivity 31–81 %. They concluded: "in many settings, proxy reports may represent a better strategy for determining the cause of death than reliance on death certificates".

For many conditions mortality is not a frequent occurrence and only in the largest and longest trials would it be a practical choice. Thus, If the endpoints of interest are rare, RCTs have to be large (and expensive), so the question might arise as to how one can design a study to garner more endpoints? Several considerations for increasing endpoints include: extending the follow-up time, broaden the definition of an event, and, don't use the events of interest rather use surrogate endpoints. An example of this latter point might be a heart disease study in which coronary heart disease events or deaths (direct outcome of interest) and uses the surrogate of incident angina and/or revascularization procedures (this adds events) and even measures of atherosclerosis (moves to continuous measure). In a cancer study, one might be primarily interested in cancer recurrence and/or cancer death (direct), but one can move to the surrogate of tumor size that moves the outcome to a continuous measure.

In 1863, Farr said 'death is a fact, the rest is inference'. In choosing outcomes of interest, death or a disease event is usually the event of interest. However, as previously mentioned, it is frequently necessary to use a surrogate for the endpoint of interest, such as when the disease occurrence is rare and/or far in the future. The main variable that drives sample size and Study Power is the difference in the outcome between the intervention and the control group. Table 3.10 summarizes the sample size necessary based upon these aforementioned differences. One can see from Table 3.10 that most studies would have to be quite large unless the treatment difference is large, and for most outcomes these days, it is common to have treatment differences of no more than 20 %.

A surrogate endpoint is simply a laboratory value, sign, or symptom that is a substitute for the real outcome one is interested in [47]. The assumption is that changes induced in a surrogate endpoint accurately and nearly completely reflect changes in the clinically meaningful endpoint. To realize that assumption, an accurate well-documented model of the outcome of interest is a prerequisite, but it should be understood that the model is only that, and the model may be far from the truth. As

is true of most definitions, there is debate about the best definition for a surrogate endpoint, and it is also important to distinguish surrogate endpoints from intermediate endpoints and statistical correlations. Speaking statistically, Prentice [48] has offered the following definition: '*a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint*.'

Examples of surrogate endpoints include blood pressure reduction in lieu of stroke (this has been termed a 'strong surrogate' by Anand et al.); [49] fasting blood sugar (or hemoglobin HbA1c) in lieu of diabetic complications; and bone mineral density in lieu of fractures. Surrogates are also commonly used early in drug development such as dose ranging or preliminary proof of efficacy ('developmental surrogates'). 'Supportive surrogates' are those outcomes that support and strengthen clinical trial data. The reasons for choosing a surrogate endpoint predominantly revolve around the fact that it might be easier to measure than the clinical endpoint of interest, or that it occurs early in the natural history of the disease of interest (and thus long-term trials are avoided). But as is true of almost any decision one makes in conducting a clinical trial, there are assumptions and compromises one has to make when choosing a surrogate endpoint. For example, many surrogates have been inadequately validated, and many if not most surrogates have several effect pathways (see Fig. 3.5). Other considerations for using a surrogate endpoint are that it should be easier to assess than the corresponding clinical endpoint, and in general, be more frequent; and, that an estimate of the expected clinical benefit should be derivable from the interventions effect upon the surrogate. An example of the controversy regarding surrogate endpoints is highlighted by the discussion of Kelsen [50] regarding the use of tumor regression as an adequate surrogate for new drugs to treat colorectal cancer. On the basis of a meta-analysis, Buyse et al. [51] proposed that surrogate endpoints of efficacy, without direct demonstration of an improvement in survival, could be used to identify effective new agents. The FDA, however, requires that there be a survival advantage before it approves such a drug. That is, a response rate higher than standard therapy (defined as tumor regression >50 %) is by itself an inadequate benefit for drug approval. As stated in the commentary by Kelsen '*the critical question in the debate over the adequacy of response rate as a surrogate endpoint for survival is whether an objective response to treatment is merely associated with a better survival, or whether the tumor regression itself lengthens survival*.'

There are differences in an intermediate endpoint, correlate, and a surrogate endpoint, although an intermediate endpoint may serve as a surrogate. Examples of intermediate endpoints include such things as angina pectoris, or hyperglycemic symptoms i.e. these are not the ultimate outcome of interest (MI, or death etc) but are of value to the patient should they be benefited by an intervention. Another example is from the earlier CHF literature where exercise-walking time was used as an intermediate endpoint as well as a surrogate marker for survival. A number of drugs improved exercise-walking time in the CHF patient; but long-term studies proved that the same agents that improved walking time actually resulted in earlier death. A hypothetical example of a surrogate 'misadventure' is exemplified by a

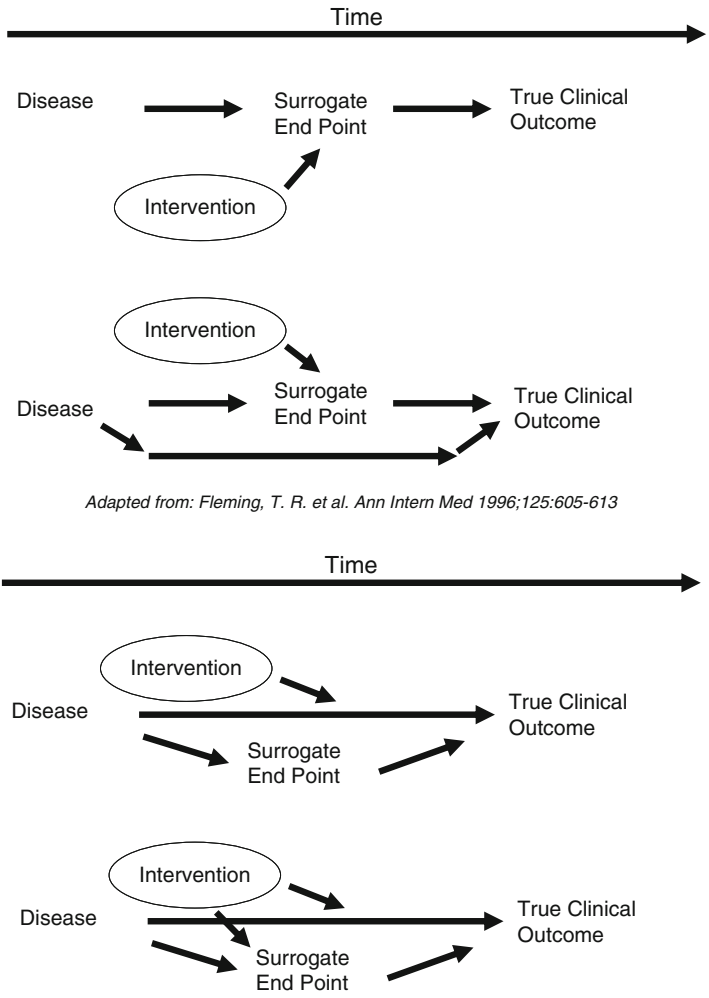

Fig. 3.5 Different surrogate paradigms

scenario where a new drug is used in pneumonia, and it is found to lower the patients white blood count (this used as a surrogate marker for improvement in the patients pneumonia). Subsequently, this hypothetical 'new drug' is found to be cytotoxic to white blood cells but obviously had little effect on the pneumonia. But, perhaps the most glaring example of a surrogate 'misadventure' is represented by a real trial – the Cardiac Arrhythmia Suppression Trial (CAST) [52]. At the time of CAST, premature ventricular contractions (PVC's) were thought to be a good surrogate for ventricular tachycardia or ventricular fibrillation, and thereby for sudden cardiac death (SCD). It was determined that many anti-arrhythmic agents available at the time or being developed reduced PVC's, and it was assumed would benefit the real outcome of interest, SCD. CAST was proposed to test the hypothesis that these

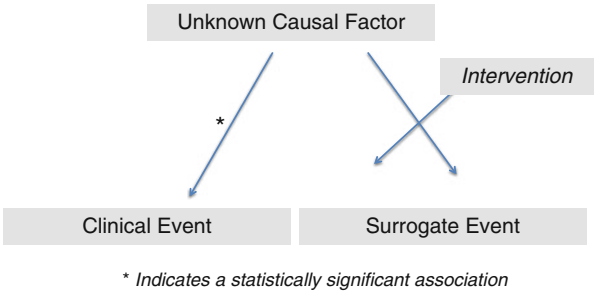

Fig. 3.6 Depicts a correlation (statistically significant) between a causal factor and a clinical event. While treatment impacted the surrogate event, it had no effect on the clinical event since it does not lie in the direct pathway

anti-arrhythmic agents did actually reduce SCD (in a post MI population) and this study was surrounded with some furor about the studies ethics, since a placebo control was part of the study design (it was felt strongly by many that the study was unethical since it was so likely that reduction in PVCs led to a reduction in SCD and how could one therefore justify a placebo arm). In fact, it turned out that the anti-arrhythmic therapy not only failed to reduce SCD, but in some cases it increased its frequency. A final example of surrogate misadventure occurred in 2007, when the Chairman of the FDA Advisory panel that reviewed the safety of rosiglitazone stated that the time has come to abandon surrogate endpoints for the approval of type 2 diabetes drugs. This resulted from the use of glycated hemoglobin as a surrogate for diabetes morbidity and mortality as exemplified in the ADOPT (A Diabetes Outcome Prevention Trial) study where patients taking rosiglitazone had a greater decrease in glycosylated hemoglobin than in patients taking comparator drugs, yet the risks of CHF and cardiovascular ischemia were higher with rosiglitazone [53].

Correlates may or may not be good surrogates. Recall, 'that a surrogate endpoint requires that the effect of the intervention on the surrogate end-point predicts the effect on the clinical outcome-a much stronger condition than correlation.' [47] Another major point of confusion is that between statistical correlation and proof of causality as demonstrated in Fig. 3.6 as discussed by Boissel et al. [54].

In summary, it should be understood that most (many) potential surrogates markers used in clinical research have been inadequately validated and that the surrogate marker must fully (or nearly so) capture the effect of the intervention on the clinical outcome of interest. However, many if not most treatments have several effect pathways and this may not be realized, particularly early in the research of a given intervention. Table 3.11 summarizes some of the issues that support using a surrogate. Surrogate endpoints are most useful in phase 1 and 2 trials where 'proof of concept' or dose-response is being evaluated. One very important additional down-side to the use of surrogate measures is a result of its effect on the safety evaluation of an intervention i.e. the ability to use smaller sample sizes and shorter trials imparted by the use of a surrogate endpoint, in order to gain insight into the benefit of an intervention results in the loss of important safety information.

Table 3.11 Support for and against the use of surrogate outcomes

Support for/against surrogates		
Factor	Favors surrogate	Does not favor surrogate
Biologic plausibility	Epidemiologic evidence extensive; excellent animal models pathogenesis and MOA understood; surrogate is late in causal pathway	Less extensive evidence; no animal model; MOA not understood, surrogate early in causal pathway
Success in clinical trials	Effect on surrogate has predicted outcome with other drugs in class and in disease	Inconsistent results across classes
Risk/benefit	Serious or life-threatening illness and no alternative treatment; large safety database; short term use; difficulty studying clinical endpoint	Less serious disease; little safety data; long term use; easy to study clinical endpoint

MOA mechanism of action

Selection of Endpoints

Table 3.10 makes the point that for most clinical trials, one of the key considerations is the difference in events between the investigational therapy and the control. It is this difference (along with the frequency of events) that drives the sample size and power of the study. From Table 3.10, one can compare the rate in the control group compared to the intervention effect. Thus, if the rate in the control group of the event of interest is high (say 20 %) and the treatment effect is 20 % (i.e. an expected 50 % reduction compared to control), a sample size of 266 patients would be necessary. Compare that to a control rate of 2 % and a treatment effect of 10 % (i.e. a reduction compared to control from 2 to 1.8 %), where a sample size of 97959 would be necessary. Often the question is asked; "What is a meaningful difference in endpoints?"

> A difference to be a difference must make a difference *(Gertrude Stein)*.

Primary and Secondary Endpoints

O'Neil [55] defines an endpoint as "*results, condition or events associated with individual study patients that are used to assess study treatments*". The characteristics of endpoint measures should include those that are easy to diagnose, easy to identify (i.e. no evaluator judgment needed), free of measurement error, reliable with repeated measures, have high internal validity and be directly linked to property of interest, and have good external validity.

Endpoints can be primary, secondary, tertiary, etc. A primary endpoint for a drug in development is a "clinical endpoint that provides evidence sufficient to fully categorize clinically the effect of a treatment that would support a regulatory claim for the

treatment". A secondary endpoint is when there is "additional clinical characterization of a treatment but could not, by itself, be convincing of a clinically significant treatment effect". Tertiary and other endpoints are mostly exploratory. Some questions about secondary endpoints include:

- How does one interpret secondary endpoints when the primary endpoint for which the clinical trial was initially designed does not meet the proposed effect.
- Some argue for caution in making inferences from secondary endpoints, and certainly there are limitations and greater concerns for a secondary endpoint effect that is derived from only one study. The likelihood of replication of the finding in another study of identical size and design as a useful concept to guide this interpretation.
- O'Neill R. (1997) argues that "secondary endpoints ***cannot*** be validly analyzed if the primary endpoint does not demonstrate clear statistical significance" [55], while Davis, C.E. (1997) argues that "secondary endpoints ***can*** be validly analyzed, even if the primary endpoint does not provide clear statistical significance" [55].

In practice, it is rare that trials use a single endpoint, and endpoints frequently cover clinical events, symptoms, physiologic measures, quality of life etc. One example is taken from the "Multiple Sclerosis literature where the result of interest was neurological disability and endpoints included episodes" of focal neurological signs and symptoms, disability rating scales, MRI changes, and CSF changes.

Ultimately the choice of endpoints is a critical and challenging study design decision, based upon considerations such as the phase of development of the clinical question, the specific disease under study, the characteristics of the measure, and the questions the investigator wants answered by the trial. General guidelines in the choice of endpoints include the use of "hard endpoints" whenever possible ("hard" endpoints are clinical landmarks that are well-defined in the study protocol, are definitive with respect to disease process, and not subjective). It is true that some endpoints are useful and reliable even when they require some subjectivity, and the key issue is not the classification of an endpoint as "hard" or "soft", but how prone to measurement error the endpoint is.

Finally other arguments centered on study endpoints are that many advocate having a single primary endpoint, since this is what "drives" sample size calculations; and, multiple endpoints introduces the possibility of Type I error.

Composite Endpoints

It is generally realized that there is an increasing challenge to conduct adequately powered clinical trials. Most trials are designed to assess the time to some first event between two arms of a study. More and more frequently, different clinical events related to the target disease are combined to form a composite endpoint. Composite endpoints (rather than a single endpoint) are being increasingly used as effect sizes for most new interventions are becoming smaller. Effect sizes are becoming smaller

because newer therapies need to be assessed when added to all clinically accepted therapies; and, thus the chance for an incremental change is reduced. For example, when the first therapies for heart failure were introduced, they were basically added to diuretics and digitalis. Now, a new therapy for heart failure would have to show benefit in patients already receiving more powerful diuretics, digitalis, angiotensin converting enzyme inhibitors and/or angiotensin receptor blockers, appropriately used beta adrenergic blocking agents, statins etc. To increase the 'yield' of events, composite endpoints are utilized (a group of individual endpoints that together form a 'single' endpoint for that trial). Thus, the rationale for composite endpoints comes from three basic considerations: statistical issues (sample size considerations due to the need for high event rates in the trial in order to keep the trial relatively small, of shorter duration and with less expense), the pathophysiology of the disease process being studied, and the increasing need to evaluate an overall clinical benefit. There are several downsides associated with the use of composite endpoints, one is that the benefits ascribed to an intervention are assumed to relate to all the components of the composite. Consider the example of a composite endpoint that includes death, MI, and urgent revascularization. In choosing the components of the composite, one should not be driven by the least important variable just because it happens to be the most frequent (e.g. death, MI, urgent revascularization, would be a problem if revascularization turned out to be the main positive finding). Another downside is that the first event within a composite endpoint may not reflect the most clinically important endpoint, and if the study is designed for time to first event, subsequent events within the composite will be missed. Thus incorporating subsequent events is seemingly rational [56]. Montori et al. provided guidelines for interpreting composite endpoints which included asking whether the individual components of composite endpoints were of similar importance, occurred with about the same frequency, had similar relative risk reductions, and had similar biologic mechanisms [57]. Armstrong and Westerhaut added to this by recommending that a strategy for future trials would be to include not just the initial event, but all events and report both per patient and overall rates; and, including a gradation of event severity (e.g. a large MI with heart failure has a very different meaning than a small periprocedural MI or a hemorrhagic stroke vs. a transient left arm weakness).

Freemantle et al. assessed the incidence and quality of reporting of composite endpoints in randomized trials and asked whether composite endpoints provide for greater precision but at the expense of greater uncertainty [58]. Their conclusion was that the reporting of composite outcomes is generally inadequate and as a result, they provided several recommendations regarding the use of composite endpoints such as following the CONSORT guidelines, interpreting the composite endpoint rather than parsing the individual endpoints, and defining the individual components of the composite as secondary outcomes. The reasons for their recommendations stemmed from their observations that in many reports they felt that there was inappropriate attribution of the treatment effects on specific endpoints when only composite endpoints yielded significant results, the effect of dilution when individual endpoints might not all react in the same direction, and the effect of excessively influential endpoints that are not associated with irreversible harm.

Table 3.12 An example of using MACE as a composite endpoint

Acute vs. non acute MI	MACE definition
1.7 (1.2–2.4)	Death; MI; stent thrombosis
1.15 (0.98–1.6)	Death; MI; stent thrombosis; target vessel revascularization
1.13 (0.95–1.4)	Death; MI; stent thrombosis; repeat revascularization
Multi-lesion vs. one lesion attempt	
1.1 (0.75–1.5)	Death; MI; stent thrombosis
1.35 (1.2–1.75)	Death; MI; stent thrombosis; target vessel revascularization
1.25 (0.01–1.52)	Death; MI; stent thrombosis; repeat revascularization

Adapted from: Kip et al. [60]

In an accompanying editorial by Lauer and Topel they list a number of key questions that should be considered when composite endpoints are reported or when an investigator is contemplating their use [59]. First, is whether the end points themselves are of clinical interest to patients and physicians, or are they surrogates; second, how nonfatal endpoints are measured (e.g. is judgment involved in the end point ascertainment, or is it a hard end point); third, how many individual endpoints make up the composite and how are they reported (ideally each component of the composite should be of equal clinical importance – in fact, this is rarely the case); and finally, how are non fatal events analyzed – that is are they subject to competing risks. As they point out, patients who die cannot later experience a non fatal event so a treatment that increases the risk of death may appear to reduce the risk of nonfatal events, and vice versa [59].

Kip et al. [60] reviewed the problems with the use of composite endpoints in cardiovascular studies. The term "major adverse cardiac events:" or MACE is used frequently in cardiovascular studies, a term that was born with the percutaneous coronary intervention studies in the 1990s. Kip et al. noted that MACE encompassed a variety of composite endpoints, the varying definitions of which could lead to different results and conclusions, leading them to the recommendation that MACE as a composite endpoint should be avoided. Table 3.12 from their article demonstrates this latter point rather well.

As mentioned above, composite endpoints are commonly used to increase event rates in an effort to increase statistical power. However, attention towards whether the individual components of the composite are likely to be differentially affected by the intervention is important. Bethel et al. performed a meta-analysis to determine the effect of angiotensin-converting enzyme inhibitors or angiotensin receptor blockers on individual cardiovascular outcomes; and then applied these treatment effects to two different composite cardiovascular endpoints. They found that although composite endpoints did augment event rates, they did not necessarily increase statistical power, and in fact, in some cases reduced it [61]. As they noted, *"occurrence of the composite endpoint must be in keeping with the duration and intensity of follow-up within a clinical trial and should reflect prior knowledge of*

the magnitude of expected treatment benefits. If insufficient data exist to estimate the treatment effect, pooled data based on plausibly similar mechanisms of action may be used instead."

Central to the selection of endpoints is how the endpoints are adjudicated, and for most large clinical trials this is generally accomplished with a centralized system. This is most important when the primary endpoint is a nonfatal event since the definition may be somewhat subjective. The main concern relative to adjudication is to avoid differential misclassification-that is to adjudicate events that are biased by applying the outcome definition variably or by knowing to which treatment assignment the patient was in (as might occur in an open-label study). The idea is that with a central adjudication system in which the adjudicators are blinded as to the treatment assignment and apply the same definitions uniformly, will yield the least biased assessment. However, this aforementioned concept has not been adequately investigated. Granger et al. reviewed the literature concerning the rationale and justification for central adjudication, and came to the conclusion that it has not been shown to improve the ability to determine treatment effects, and may be overly complex and overused. And yet, the FDA and the scientific community derive confidence in the validity of results when central adjudication is performed [62].

Trial Duration

A critical decision in performing or reading about a RCT (or any study for that matter) is the specified duration of follow-up, and how that might influence a meaningful outcome. Many examples and potential problems exist in the literature, but basically in interpreting the results of any study (positive or negative) the question should be asked 'what would have happened had a longer follow-up period been chosen?' An example is the Canadian Implantable Defibrillator Study (CIDS) [63]. CIDS was a RCT comparing the effects of defibrillator implantation to amiodarone in preventing recurrent sudden cardiac death in 659 patients. At the end of study (a mean of 5 months) a 20 % relative risk reduction occurred in all-cause mortality, and a 33 % reduction occurred in arrhythmic mortality, when ICD therapy was compared with amiodarone (this latter reduction did not reach statistical significance). At one center, it was decided to continue the follow-up for an additional mean of 5.6 years in 120 patients who remained on their originally assigned intervention [64]. All-cause mortality was increased in the amiodarone group. The Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering (MIRACL) trial is an example of a potential problem in which study duration could have been problematic (but probably wasn't) [65]. The central hypothesis of MIRACL was that early rapid and profound cholesterol lowering therapy with atorvastatin could reduce early recurrent ischemic events in patients with unstable angina or acute non-Q wave infarction. Often with acute intervention studies, the primary outcome is assessed at 30 days after the sentinel event. From Fig. 3.7 one can see that there was no difference in the primary outcome at 30 days. Fortunately the study specified a 16-week follow-up, and a significant difference was seen at that time point. Had the

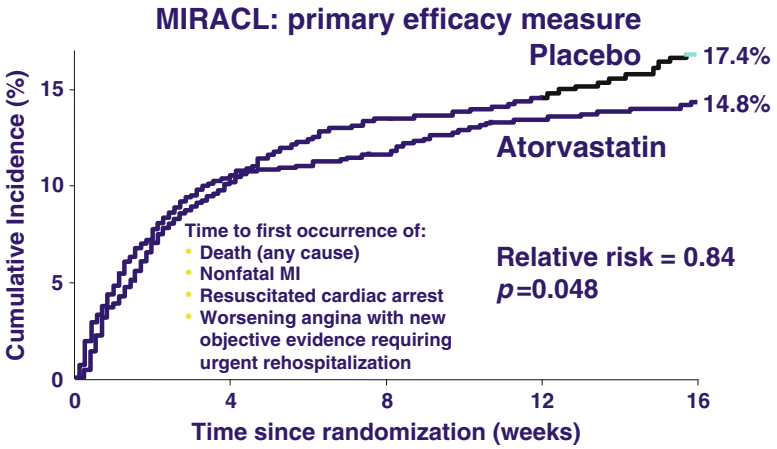

Fig. 3.7 The results of MIRACL for the primary outcome. What would have been the conclusion for the intervention if the pre-specified study endpoint was 1 month? (Adapted from Schwartz et al. [65])

study been stopped at 30 days the ultimate benefit would not have been realized. Finally, an example from the often cited controversial ALLHAT study which demonstrated a greater incidence in new diabetes in the diuretic arm as assessed at the study end of 5 years [66]. The investigators pointed out that this increase in diabetes did not result in a statistically significant difference in adverse outcomes when the diuretic arm was compared to the other treatment arms. Many experts have subsequently opined that the trial duration was too short to assess adverse outcomes from diabetes, and had the study gone on longer that it is likely that a significant difference in adverse complications from diabetes would have occurred.

The Devil Lies in the Interpretation

It is interesting to consider and important to reemphasize, that intelligent people can look at the same data and render differing interpretations. MRFIT is exemplary of this principal, in that it demonstrates how mis-interpretation can have far-reaching effects. One of the conclusions from MRFIT was that reduction in cigarette smoking and cholesterol was effective, but '*possibly an unfavorable response to antihypertensive drug therapy in certain but not all hypertensive subjects*' led to mixed benefits [31]. This 'possibly unfavorable response' (thought to be due to diuretic based hypokalemia) has since been at least questioned if not proven to be false.

Differences in interpretation was also seen in the alpha-tocopherol, beta carotene cancer study [33]. To explain the lack of benefit and potential worsening of cancer risk in the treated patients, the authors opined that perhaps the wrong dose was used, or that the intervention period was to short, since '*no known or described mechanisms and no evidence of serious toxic effects of this substance* (beta carotene)

in humans' had been observed. This points out how ones personal bias can influence ones 'shaping' of the interpretation of a trials results. Finally, there are many examples of trials where an interpretation of the results is initially presented only to find that after publication differing interpretations are rendered. Just consider the recent controversy over the interpretation of the ALLHAT results [66].

Causal Inference, and the role of **the Media** in reporting clinical research will be discussed in chapters 16 and 20.

Conclusions

While randomized clinical trials are the 'gold standard' clinical research design, there remains many aspects of trial design that must be considered before accepting the studies results, even when the study design is a RCT. Starzi et al. in their article entitled 'Randomized Trialomania? The Multicentre Liver Transplant Trials of Tacrolimus' outline many of the roadblocks and pitfalls that can befall even the most conscientious clinical investigator [67]. Ioannidis presents an even more somber view of clinical trials, and has stated 'there is increasing concern that in modern research', false findings may be the majority or even the vast majority of published research claims. He points out that this should not be surprising since it can be proven that most (one can argue many if not most) claimed research findings are false [68]. Also, many feel that misleading interpretations result from an over-reliance on statistical testing, that is, that the strength of evidence is often judged by conventional tests that rely heavily on statistical significance, with less attention paid to the clinical significance or practical importance of treatment effects [69]. Kaul and Diamond cite three particular technical limitations to the interpretation of the results from a clinical trial: the emphasis of statistical significance over clinical importance, the use of composite endpoints, and the use of subgroup analyses (refer to sections on composite endpoints and subgroup analysis above). Relative to the over-reliance on statistical testing is the controversy that surrounds relying on the p value, and as a wit opined *"a p value is no substitute for a brain"* (anonymous source cited in Kaul and Diamond). The significance level that is used most commonly is the P value ≤ 0.05 that represents the maximum probability that is tolerated for rejecting a hypothesis that is in fact true. But in contrast to the $p \leq 0.05$ standard for statistical significance is that there are no guidelines for what difference is clinically significant and some then equate the two. Kaul and Diamond conclude that "while statistical significance tells us whether a difference is likely to be real, it does not place that reality into meaningful clinical context by telling us the difference is small, large, trivial, or important. A formal evaluation of clinical importance (using frequentist confidence intervals, the number needed to treat and the number needed to harm, or Bayesian probabilities), given the overall risk-benefit-cost profile of each therapeutic intervention, should be included in the analysis, interpretation, and presentation of the results of clinical trials." Table 3.13 provides a list of at least 12 misconceptions about P values [70].

Table 3.13 Twelve P-value misconceptions

1. If P = .05, the null hypothesis has only a 5 % chance of being true
2. A nonsignificant difference (e.g., P > .05) means there is no difference between groups
3. A statistically significant finding is clinically important
4. Studies with P values on opposite sides of .05 are conflicting
5. Studies with the same P value provide the same evidence against the null hypothesis
6. P ≤ .05 means that we have observed data that would occur only 5 % of the time under the null hypothesis
7. P ≤ .05 and P < .05 mean the same thing
8. P values are properly written as inequalities (e.g., "P < .02" when P = .015)
9. P ≤ .05 means that if you reject the null hypothesis, the probability of a type I error is only 5 %
10. With a P ≤ .05 threshold for significance, the chance of a type I error will be 5 %
11. You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible
12. A scientific conclusion or treatment policy should be based on whether or not the P value is significant

Adapted from Goodman [70]

One final note of caution revolves around the use of reading or reporting only abstracts in decision-making. As Toma et al. noted, 'not all research presented at scientific meetings is subsequently published, and even when it is, there may be inconsistencies between these results and what is ultimately printed' [71]. They compared RCT abstracts presented at the American College of Cardiology sessions between 1999 and 2002, and subsequent full-length publications. Depending upon the type of presentation (e.g. late breaking trials vs. other trials) 69–79 % were ultimately published; and, discrepancies between meeting abstracts and publication results were common even for the late breaking trials (see Chap. 19 for further discussion of abstracts) [71].

References

1. Glasser SP, Howard G. Clinical trial design issues: at least 10 things you should look for in clinical trials. J Clin Pharmacol. 2006;46:1106–15.
2. Grady D, Herrington D, Bittner V, Blumenthal R, Davidson M, Hlatky M, et al. Cardiovascular disease outcomes during 6.8 years of hormone therapy: Heart and Estrogen/progestin Replacement Study follow-up (HERS II). JAMA. 2002;288:49–57.
3. Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. JAMA. 1998;280:605–13.
4. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. JAMA. 2002;288:321–33.
5. Grady D, Rubin SM, Petitti DB, Fox CS, Black D, Ettinger B, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. Ann Intern Med. 1992;117:1016–37.

6. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. Prev Med. 1991;20:47–63.
7. Sullivan JM, Vander Zwaag R, Hughes JP, Maddock V, Kroetz FW, Ramanathan KB, et al. Estrogen replacement and coronary artery disease. Effect on survival in postmenopausal women. Arch Intern Med. 1990;150:2557–62.
8. Bhatt DL, Cavender MA. Are all clinical trial sites created equal? J Am Coll Cardiol. 2013;61:580–1. doi:10.1016/j.jacc.2012.10.024.
9. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367:1355–60. PMC3771340.
10. Butler J, Subacius H, Vaduganathan M, Fonarow GC, Ambrosy AP, Konstam MA, et al. Relationship between clinical trial site enrollment with participant characteristics, protocol completion, and outcomes: insights from the EVEREST (Efficacy of Vasopressin Antagonism in Heart Failure: Outcome Study with Tolvaptan) trial. J Am Coll Cardiol. 2013;61:571–9. doi:10.1016/j.jacc.2012.10.025.
11. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. 2002;359:57–61.
12. Loscalzo J. Clinical trials in cardiovascular medicine in an era of marginal benefit, bias, and hyperbole. Circulation. 2005;112:3026–9.
13. Bienenfeld L, Frishman W, Glasser SP. The placebo effect in cardiovascular disease. Am Heart J. 1996;132:1207–21.
14. Clark PI, Leaverton PE. Scientific and ethical issues in the use of placebo controls in clinical trials. Annu Rev Public Health. 1994;15:19–38.
15. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. N Engl J Med. 1994;331:394–8.
16. Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. JAMA. 2005;294:2203–9.
17. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ. 1948;ii:769–82.
18. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.
19. Gum PA, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis. JAMA. 2001;286:1187–94.
20. Reviews of statistical and economic books, Student's Collected Papers. J R Stat Soc. 1943;106:278–9.
21. Fleming TR. Addressing missing data in clinical trials. Ann Intern Med. 2010;154:113–7. PMC3319761.
22. A Village of 100 In. 2nd ATS Media ed: A Step Ahead.
23. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Engl J Med. 1991;325:445–53.
24. Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Endarterectomy for asymptomatic carotid artery stenosis. JAMA. 1995;273:1421–8.
25. Lang JM. The use of a run-in to enhance compliance. Stat Med. 1990;9:87–93; discussion –5.
26. Franciosa JA. Commentary on the use of run-in periods in clinical trials. Am J Cardiol. 1999;83:942–4. A9.
27. Pablos-Mendez A, Barr RG, Shea S. Run-in periods in randomized trials: implications for the application of results in clinical practice. JAMA. 1998;279:222–5.
28. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet. 2002;359:696–700.
29. Shem S. The house of god. In: Palgrave Macmillan; 1978:280.
30. Smith DH, Neutel JM, Lacourciere Y, Kempthorne-Rawson J. Prospective, randomized, open-label, blinded-endpoint (PROBE) designed trials yield the same results as double-blind, placebo-controlled trials with respect to ABPM measurements. J Hypertens. 2003;21:1291–8.

31. Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial. Risk factor changes and mortality results. JAMA. 1982;248:1465–77.
32. Mayo E. The human problems of an industrial civilization. New York: Macmillan; 1993.
33. Alpha-Tocopherol T, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. N Engl J Med. 1994;330:1029–35.
34. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999;319:670–4.
35. The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. N Engl J Med. 1980;303:1038–41.
36. The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. N Engl J Med. 1980;302:250–6.
37. Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. N Engl J Med. 1979;301:1410–2.
38. Howard G, Chambless LE, Kronmal RA. Assessing differences in clinical trials comparing surgical vs nonsurgical therapy: using common (statistical) sense. JAMA. 1997;278:1432–6.
39. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000;355:1064–9.
40. Sleight P. Debate: subgroup analyses in clinical trials: fun to look at – but don't believe them! Curr Control Trials Cardiovasc Med. 2000;1:25–7.
41. Amarenco P, Goldstein LB, Szarek M, Sillesen H, Rudolph AE, Callahan 3rd A, et al. Effects of intense low-density lipoprotein cholesterol reduction in patients with stroke or transient ischemic attack: the Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) trial. Stroke. 2007;38:3198–204.
42. Black HR, Elliott WJ, Grandits G, Grambsch P, Lucente T, White WB, et al. Principal results of the Controlled Onset Verapamil Investigation of Cardiovascular End Points (CONVINCE) trial. JAMA. 2003;289:2073–82.
43. Weir MR, Ferdinand KC, Flack JM, Jamerson KA, Daley W, Zelenkofske S. A noninferiority comparison of valsartan/hydrochlorothiazide combination versus amlodipine in black hypertensives. Hypertension. 2005;46:508–13.
44. Kaul S, Diamond GA, Weintraub WS. Trials and tribulations of non-inferiority: the ximelagatran experience. J Am Coll Cardiol. 2005;46:1986–95.
45. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. JAMA. 2006;295:1147–51.
46. Halanych JH, Shuaib F, Parmar G, Tanikella R, Howard VJ, Roth DL, et al. Agreement on cause of death between proxies, death certificates, and clinician adjudicators in the Reasons for Geographic and Racial Differences in Stroke (REGARDS) study. Am J Epidemiol. 2011;173:1319–26. PMC3101067.
47. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996;125:605–13.
48. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med. 1989;8:431–40.
49. Anand IS, Florea VG, Fisher L. Surrogate end points in heart failure. J Am Coll Cardiol. 2002;39:1414–21.
50. Kelsen DP. Surrogate endpoints in assessment of new drugs in colorectal cancer. Lancet. 2000;356:353–4.
51. Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Meta-Analysis Group in Cancer. Lancet. 2000;356:373–8.
52. Greene HL, Roden DM, Katz RJ, Woosley RL, Salerno DM, Henthorn RW. The Cardiac Arrhythmia Suppression Trial: first CAST… then CAST-II. J Am Coll Cardiol. 1992;19:894–8.
53. FDA Adviser Questions Surrogate Endpoints for Diabetes Drug Approvals. In: Medpage Today; 2007.
54. Boissel JP, Collet JP, Moleur P, Haugh M. Surrogate endpoints: a basis for a rational approach. Eur J Clin Pharmacol. 1992;43:235–44.

55. O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. Control Clin Trials. 1997;18:550–6. discussion 61–7.
56. Armstrong PW, Westerhout CM. The power of more than one. Circulation. 2013;127:665–7. doi:10.1161/CIRCULATIONAHA.112.000627.
57. Montori VM, Busse JW, Permanyer-Miralda G, Ferreira I, Guyatt GH. How should clinicians interpret results reflecting the effect of an intervention on composite endpoints: should I dump this lump? ACP J Club. 2005;143:A8.
58. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003;289:2554–9.
59. Lauer MS, Topol EJ. Clinical trials – multiple treatments, multiple end points, and multiple lessons. JAMA. 2003;289:2575–7.
60. Kip KE, Hollabaugh K, Marroquin OC, Williams DO. The problem with composite end points in cardiovascular studies. J Am Coll Cardiol. 2008;51:701–7. doi:10.1016/j.jacc.2007.10.034.
61. Bethel MA, Holman R, Haffner SM, Califf RM, Huntsman-Labed A, Hua TA, et al. Determining the most appropriate components for a composite clinical trial outcome. Am Heart J. 2008;156:633–40. doi:10.1016/j.ahj.2008.05.018.
62. Granger CB, Vogel V, Cummings SR, Held P, Fiedorek F, Lawrence M, et al. Do we need to adjudicate major clinical events? Clin Trials. 2008;5:56–60. doi:10.1177/1740774507087972.
63. Connolly SJ, Gent M, Roberts RS, Dorian P, Roy D, Sheldon RS, et al. Canadian implantable defibrillator study (CIDS): a randomized trial of the implantable cardioverter defibrillator against amiodarone. Circulation. 2000;101:1297–302.
64. Bokhari F, Newman D, Greene M, Korley V, Mangat I, Dorian P. Long-term comparison of the implantable cardioverter defibrillator versus amiodarone: eleven-year follow-up of a subset of patients in the Canadian Implantable Defibrillator Study (CIDS). Circulation. 2004;110:112–6.
65. Schwartz GG, Olsson AG, Ezekowitz MD, Ganz P, Oliver MF, Waters D, et al. Effects of atorvastatin on early recurrent ischemic events in acute coronary syndromes: the MIRACL study: a randomized controlled trial. JAMA. 2001;285:1711–8.
66. The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the antihypertensive and lipid lowering treatment to prevent heart attack trial (ALLHAT). JAMA 2002;288:2981–97.
67. Starzl TE, Donner A, Eliasziw M, Stitt L, Meier P, Fung JJ, et al. Randomised trialomania? The multicentre liver transplant trials of tacrolimus. Lancet. 1995;346:1346–50.
68. Ioannidis JPA. Why most published research findings are false. PLoS. 2005;2:696–701.
69. Kaul S, Diamond GA. Trial and error. How to avoid commonly encountered limitations of published clinical trials. J Am Coll Cardiol. 2010;55:415–27. doi:10.1016/j.jacc.2009.06.065.
70. Goodman SA. A dirty dozen: Twelve P-value misconceptions. Semin Hematol. 2008;45:135–40. doi:10.1053/j.seminhematol.2008.04.003.
71. Toma M, McAlister FA, Bialy L, Adams D, Vandermeer B, Armstrong PW. Transition from meeting abstract to full-length journal article for randomized controlled trials. JAMA. 2006;295:1281–7.