

***Part I***  
***Is There a Problem  
with Reliability in  
Medical Studies?***



# 1

## ***An Evolution of Comparative Methodology***

Our starting point is the desire to evaluate a new medical intervention to determine if it might be useful in clinical practice. We mean this to be quite inclusive, so the medical intervention might be a drug, a vaccine, a diet, a screening test, acupuncture, cognitive therapy, or anything else that might be used to promote health and/or treat or prevent a disease. Over the years, many different methods have been employed in the name of evaluating a new medical intervention (or evaluating a new use for an existing medical intervention). These methods include the design, conduct, and analysis of studies. In this chapter we critically evaluate some of the designs that have been used to evaluate medical interventions, filling in many of the details behind the outline provided by Smith (2003). We pay special attention to the existence and nature of the control group, and the method for determining who gets which treatments.

### **1.1 SINGLE-SUBJECT STUDIES**

One might imagine a time when the standard evaluation of a medical intervention consisted in applying it to a single subject, and noting if this subject appeared to improve or deteriorate. Today we are well aware of the limitations of such single-subject studies, but it still may be useful to clarify what precisely these limitations are, to better understand the need for other methods. Consider a situation in which the natural history of a disease is known with absolute certainty, and

#### 4 *An evolution of comparative methodology*

there is literally no variation across patients. In such a case, if one patient were demonstrated to deviate from this known path following treatment with a given agent, then the response would have to be attributed to the agent, and the study would be convincing. For example, if some new potion could be applied to a corpse and bring the corpse back to life, then one corpse would be sufficient to provide strong and convincing evidence.

It is common to refer to randomization as the basis for inference; see, for example, Berger (2000), Berger *et al.* (2002), and Berger and Bears (2003). In most cases it is true that randomization serves as the basis for inference. However, in the case just described, there is an alternative basis for inference. This is because we know the potential outcomes both with and without the potion. The former was observed and the latter follows from the certainty of the natural history in the absence of the potion. This makes the prediction causal, as opposed to probabilistic (Runde, 1996). With knowledge of both potential outcomes, we are in a position to perform causal inference, and compute a legitimate and valid  $p$ -value. The probability of observing so extreme a result under the null hypothesis that the potion does not work is exactly zero, so this is the  $p$ -value. Likewise, if at some point in the future we find ourselves in a position to conduct trials on clones, so that the assumption of exchangeability or identical potential outcomes (Greenland and Robins, 1986) becomes tenable, then randomization would be unnecessary as we would still be able to perform causal inference, and compute a legitimate and valid  $p$ -value of zero if any between-group difference is found.

### 1.2 CASE SERIES AND COHORT STUDIES

The conditions described above are not likely to be met in actual clinical practice. How well, then, does a single-subject study perform when the natural history is not known with certainty? The answer is not very well, because the experiences of a single subject represent the outcome of a single Bernoulli trial with unknown success probability. There is little hope, based on this single Bernoulli trial, of estimating the success probability, let alone of establishing its difference from the success probability in the absence of treatment. Intuitively, it is clear that a larger sample size will offer some benefits, because the

**Historical controls 5**

sampling variability is reduced with increases in the sample size. As such, a case series or cohort of consecutive patients with the same or a similar diagnosis all treated the same way might be preferred to a single-subject study. These designs are often used for Phase II studies, to ascertain preliminary indications of efficacy.

While the case-series approach is certainly preferable (at least from a scientific perspective, but from an ethical perspective this point could be debated) to a single subject study, it also retains some of the same drawbacks of the single-subject study. In particular, the case-series design does not address the fact that the evaluation of a medical intervention is necessarily comparative. An individual may not care, for example, if a vaccine is 'good' or 'bad' in a vacuum, but this individual may have come to understand that these descriptors refer to the vaccine being better or worse, respectively, than the absence of the vaccine. Such an evaluation can be made only with a comparative study.

**1.3 HISTORICAL CONTROLS**

One of the simplest comparative designs is the historical control design, in which the experiences of a current case series are compared to the historical experiences of a prior cohort. This design allows for an assessment of 'better' or 'worse', which is certainly a strength. However, the comparison is confounded with both time trends and selection processes. That is, unintentionally or otherwise, especially good responders may be selected for the current case series. Clearly, this would bias the results in favor of the intervention used for the current case series. Conversely, especially bad responders may be selected for the current case series; this would bias the results in favor of the intervention used for the historical controls. Even if the current cohort is quite similar to the historical cohort prior to either being treated, they still may differ once treated for reasons having nothing to do with the treatments being compared. For example, ancillary care may be better now than it was in the past; this would bias the results in favor of the intervention used for the current case series. Conversely, managed care may deny the current cohort some health benefits that were available to the historical cohort; this would bias the results in favor of the intervention used for the historical controls. As we see, time itself

## **6 An evolution of comparative methodology**

is an important covariate that should be balanced across groups. This can be accomplished with parallel control groups, or groups that are treated at the same time.

### **1.4 PARALLEL CONTROL GROUPS**

Many studies fall in the category of parallel control. For example, one could assess the effects of smoking by comparing those who do smoke today to those who do not. Each group might be followed up for some period of time, and any occurrences of cancer or heart disease would be noted, and compared across groups. This design would balance the effects of time across groups, and would therefore eliminate this source of bias. However, there are other sources of bias that remain, most notably self-selection bias. Consider that those subjects who choose to smoke may differ in important ways from those who do not. For example, they may engage in riskier behavior, or may drink more alcohol, or may eat fewer fruits and vegetables. This means that even if a clear difference between the experiences of the smokers and the experiences of the non-smokers is found, this difference may not be attributable to smoking itself. To attribute the differences in outcomes to the agents studied requires that the comparison groups be as comparable as possible in every way other than the difference in their treatments.

### **1.5 MATCHED STUDIES**

The next improvement in study design is matching, including case-control studies (Breslow and Day, 1980). In our development, this can include both prospective and retrospective designs. As an example of the latter, the Los Angeles Retirement Study of Endometrial Cancer (Mack *et al.*, 1976; Breslow and Day, 1980, Section 5.1) was a case-control study designed to study the effect of exogenous estrogens on the risk of endometrial cancer. There were 63 cases of endometrial cancer identified from 1971 to 1975 in a retirement community near Los Angeles. Each case was matched to four controls, all of whom were alive and living in the community at the time the case (of endometrial cancer) was diagnosed, were born within one year of the case to whom they were matched, had the same marital status, and

**Matched studies 7**

had entered the community at approximately the same time as the case. In addition, controls were chosen from among women who had not had a hysterectomy prior to the time the case was diagnosed and who were therefore still at risk for the disease. One purpose of the study was to determine whether gall bladder disease was associated with endometrial cancer.

Here, the search would have been retrospective, to find out which cases and which controls had experienced gall bladder disease that would have occurred prior to the present time. It is also possible to study the effects of gender, height, genetic profiles, exposure to particular carcinogens, or exposure to particular viruses with a prospective variation of the same design. One would find cases, defined, for example, as those having gall bladder disease, and then match each of these cases to some number of controls. Now all cases and controls could be followed prospectively to determine if they develop endometrial cancer. The basis for inference in this design is the exchangeability of the cases and the controls. That is, it is hoped that the cases and controls are identical to each other in every way other than the 'caseness', or that which makes the cases become cases and that which makes the controls remain controls (not cases).

It is possible, at least in theory, to match on any number of prognostic variables, so any number of variables can be balanced across cases in controls within each matched set. If all potential confounding variables are known and measured then, again at least in theory, randomization may be considered unnecessary (Villar and Carroli, 1996). The problem is that there are often prognostic variables that are not measured. For example, subjective health perceived by a patient can predict clinical outcomes and even mortality, even after adjusting for other observed predictors (Fayers and Sprangers, 2002). In fact, as Madersbacher *et al.* (2004) pointed out, 'The comparison of new treatment modalities with so-called matched controls and particularly historic controlled series . . . confuse the results by introducing errors resulting from case selection bias, stage migration, differences in follow-up, and the evolution of supportive care. These confounding, recall, and detection biases are particularly problematic for the results of oncologic trials because the respective surgical or medical therapies can be associated with considerable treatment-related morbidity.'

It is entirely possible that, in a case-control study, the matching does not balance the unknown and/or unmeasured covariates, such

## 8 *An evolution of comparative methodology*

as subjective health perceptions. The same criticism applies to deterministic designs mistakenly referred to as ‘randomized’, such as minimization. For this reason, it is considered ideal to randomize, at least when doing so is feasible and ethical. Clearly, it is not always feasible or ethical to randomize. For example, it is not possible to randomize subjects to different genders (surgical interventions to modify the gender may be possible, but this is not the same as being born to a given gender). It is certainly conceivable to randomize subjects to exposure to carcinogens, but this is not ethical. As such, there is still a place for matched designs that are not randomized. From this point on, however, we restrict attention to randomized trials.

We note that alternating designs, and other deterministic designs including those in which allocations are based on the social security number, are often called ‘randomized’ (Berger and Bears, 2003), yet these are poor substitutes for true randomization. Nature is not in the business of randomizing the order in which patients show up to clinics, so there is no sense in which alternating designs represent truly randomized designs. As with all non-randomized studies, the observed data represent the only outcome that could have occurred. This means that if randomization is the basis for inference, and there was no true randomization, then the only valid  $p$ -value would have to assume the uninteresting value of 1.00. Only if the non-randomized study is performed in clones, or in a patient population whose natural history is known with certainty, would another basis for inference be available, to allow for the valid calculation of a more interesting  $p$ -value. Of course, assumptions can also serve as the basis for inference, as when a population is assumed to follow the normal distribution and sampling is assumed to be random.

While these two ‘assumptions’ are more often better described as violations of known facts to the contrary, a less objectionable assumption was discussed recently by Gallin *et al.* (2003) with regard to a study to evaluate the efficacy of itraconazole. Specifically, the treatments (itraconazole and placebo) were alternated over time periods within each patient, and this continued until the occurrence of an event or the end of the study. With the assumption that the probabilities of events in the two groups did not depend on time or exposure to prior treatments, it is possible to compute valid  $p$ -values, even in the absence of randomization. Now the study by Gallin *et al.* (2003) actually did use randomization, but it appears that the assumption of

time homogeneity, and not randomization itself, served as the basis for inference and the calculation of  $p$ -values.

## 1.6 RANDOMIZATION

Randomization is often said to balance all covariates, at least in distribution, across the treatment groups. For example, Beller, Gebski, and Keech (2002) state that 'Allocation of participants to specific treatment groups in a random fashion ensures that each group is, on average, as alike as possible to the other group(s). The process of randomization aims to ensure similar levels of all risk factors in each group; not only known, but also unknown, characteristics are rendered comparable, resulting in similar numbers or levels of outcomes in each group, except for either the play of chance or a real effect of the intervention(s).' While it is certainly true that randomization is used for the purpose of ensuring comparability between or among comparison groups, we will see in Chapter 2 that it is categorically not true that this goal is achieved. However, it is worth reviewing the logic behind this statement to see where it can break down.

One basic tenet of most forms of randomization is that there is no opportunity for the subject to select a treatment, and no opportunity for the investigator to assign a treatment based on subject characteristics. Exceptions exist, at least to some extent; for example, the consumer principle of randomization would allow subjects to select not the treatment *per se* but rather the probability with which they are to receive each treatment (Bird, 2001). In most cases, however, there is no consumer choice, and allocation probabilities are determined in advance. Often, but not always, these probabilities are the same for all treatment groups, to achieve balance in sample sizes across the treatment groups. For simplicity, unless otherwise noted, we will consider only two-arm randomized trials with equal allocation probabilities to the two groups (1:1 randomization). Our development will apply more broadly, however, allowing for more treatment groups and unequal randomization.

The idea of randomization is to overlay a sequence of units (subjects, or patients) onto a sequence of treatment conditions. If neither sequence can influence the other, then there should be no bias in the assignment of the treatments, and the comparison groups should be

## 10 *An evolution of comparative methodology*

comparable. We note that the bias from self-selection designs can be viewed as the influence of the former sequence on the latter sequence. Specifically, the identity of the units (or, in this case, patients) in the first sequence, and their ability to select treatments, not only influences but also determines the sequence of treatment assignments (Berger and Christophi, 2003). This is why it is not valid to compare the group of patients who were treated with (by virtue of having selected) one treatment to the group of patients who were treated with (by virtue of having selected) another treatment. In fact, sometimes there are contraindications that allow some patients to use one treatment but not another. In fact, eligibility for a chemotherapy protocol was recently found to be a good prognostic factor for invasive bladder cancer after radical cystectomy (Madersbacher *et al.*, 2004).

With randomization (understood to exclude the consumer variety), there should be no such influence of the subjects on the treatment assignments. The veracity of this statement depends on the nature of the randomization procedure. Consider, for example, randomization by tossing coins. If the coin tossing takes place only after the subject to be randomized has been identified, then it would be possible to take into consideration the preferences of this subject by rejecting the outcome of the coin toss until the preferred outcome is observed.

Schulz (1995a) defines randomization as follows: 'First, an unpredictable allocation sequence must be generated based on a random procedure. Second, strict implementation of that schedule must be secured through an assignment mechanism (allocation concealment process) that prevents foreknowledge of the treatment assignment'. He goes on to call it a 'mistake' that many medical researchers regard only the sequence generation process as the randomization itself. We disagree, and find good reason to follow instead Berger and Bears (2003) in defining randomization strictly on how the allocation sequence is generated, randomly or not. That is, a trial is randomized if, and only if, the accession numbers from any one treatment group constitute a random sample from the set of all accession numbers used. In taking this as the definition of randomization, we are not denying the importance of allocation concealment, but, for reasons that will become clear in the remainder of this book and especially in Chapter 2, there are good reasons to regard the two processes, randomization and allocation concealment, as markedly distinct entities, to allow for consideration of one without the other.

## 1.7 ADVANCE RANDOMIZATION

The type of subversion considered in the previous paragraph would represent a breakdown in the integrity of the randomization itself (later we will discuss other subversions of the allocation process that have nothing to do with any breakdown in the randomization itself), and can occur in any trial for which randomization takes place only after the (human) subjects were already selected. In practice, not only are the allocation proportions determined prior to the initiation of patient recruitment, but in fact the allocation sequence itself is also determined in advance, prior to the initiation of patient recruitment. This design feature ensures that the sequence of subjects to be randomized cannot influence the sequence of treatment assignments (Berger and Christophi, 2003). In this sense, randomization has served its purpose. We can say that randomization has contributed to the balancing of both measured and unmeasured covariates. Certainly, it has done a better job of this than any deterministic design could (Moses, 1995). Is it possible, however, for the direction of the influence to be reversed? That is, can the sequence of treatment assignments influence the sequence of subjects to be randomized?

At first, this notion seems preposterous. How can a treatment assignment alter the baseline characteristics (pre-randomization) of a patient? However, we need to take a broader view of this potential influence than simply the influence of a given treatment assignment on the corresponding patient. In fact, it is clear that once the patient is selected to be randomized, there can be no influence of the treatment allocation on that patient (at least not on any patient characteristics prior to randomization). However, if it is known that the next allocation will be to a given treatment group, then this advance knowledge may lead to selective patient recruitment. This concern can be addressed, of course, by determining each allocation only after the patient to be enrolled is identified, as was suggested by Clarke (2002). But either the allocation to be made or the patient to be enrolled has to be selected first; whichever it is may influence the other.

The biases possible with randomization only after patient accrual are at least as serious as the biases possible with advance randomization, so the best approach seems to be to randomize first, and then recruit the patients, but to try to do so in such a way that the influence of the treatment assignments on the patients enrolled is minimized,

## 12 *An evolution of comparative methodology*

or preferably eliminated. Berger and Christophi (2003) enumerated some conditions under which this reverse influence can be completely eliminated. For example, if the trial is performed in clones, all of whom are identical in every way to each other, then there can be no preferential patient selection to any treatment group. If all eligible patients must be enrolled, and can neither refuse consent (possibly after being discouraged by an investigator who was aware of the upcoming treatment assignment) or denied enrollment, then there would be no way for the treatment assignments to influence the patient selection (although it could, of course, influence the patient evaluation after enrollment).

In practice, however, both investigators and patients enjoy enrollment discretion, and studies are not done in clones. Still, if the patients to be randomized can be all assembled at once, prior to randomization, and then randomized all at once, then there is no opportunity to act on any advance knowledge. However, most trials are sequential, in the sense of using staggered patient entry, and patients are randomized as they are enrolled, often due to the need for immediate treatment of the disease that qualified them for the trial in the first place. This leaves one other hope for eliminating the influence of the treatment assignments on the patient selection. If there is absolutely no advance knowledge of upcoming allocations, then there is no opportunity to preferentially select better responders into one treatment group or the other. This is the idea behind allocation concealment (Schulz, 1995a, 1995b, 1996), which is essentially the masking (or concealing) of each allocation just until it is executed. That is, if the allocation itself reveals the nature of the treatment assigned, this would not constitute a violation of allocation concealment, because it occurs only after the patient to be allocated has already been selected.

### 1.8 ALLOCATION CONCEALMENT

Discussions of the imperfections of masking are quite relevant to the evaluation of the success of allocation concealment. For example, in a discussion of the distinction between a claim of masking and true masking, Oxtoby *et al.* (1989) pointed out that 'the presumption that a plan to which one has aspired has come to fruition by virtue of aspiration alone is not science, and is particularly inapposite for a

**Allocation concealment 13**

profession which should have a reputation for making clear distinctions between fantasy and reality'. Masking may be defined as either the process (researchers not revealing treatment codes until the database is locked) or the result (complete ignorance of all trial participants as to which patients received which treatments until the database is locked). Analogously, then, allocation concealment may be defined as either the process (researchers not revealing treatment codes until the patient is randomized) or the result (complete ignorance of all trial participants as to which patients received which treatments until the patient is randomized). It is often said that masking is possible only some of the time, while allocation concealment is always possible. The reason for this sentiment is clear enough. It is hard to imagine how to mask a trial comparing a surgery to a medical treatment, for example. Yet allocation concealment would still be possible even in this case, because the unmasking of each patient would occur only after the allocation, and after the selection of the patient.

Of course, there are cases in which sham surgery is ethical, and might cause the study to be as well masked as trials comparing a medicine to a placebo (Jones *et al.*, 2003). Yet even in cases in which sham surgery is deemed unethical, there is still something troubling about stating that masking is possible only some of the time, while allocation concealment is always possible. Specifically, Berger and Christophi (2003) noted that the process of masking is always possible, and pointed out that

this confusion of the two definitions is a double-standard. If masking is possible only some of the time, then clearly reference is being made to the result, and not the process. To be fair, then, one would have to ask if the *result* of allocation concealment is always possible. Sealed envelopes have been held to lights, phantom patients have been enrolled, and locked files have been raided to determine upcoming treatment allocations in successful subversions of allocation concealment (Schulz, 1995a) . . . so only the *process* of allocation concealment, but not its result, can be ensured.

In future chapters, we will have more to say about the specific mechanisms by which allocation concealment can be subverted. For now, we highlight the two key points, which are as follows. First, even in trials labeled as 'randomized', randomization can be conducted with or without error, or not at all. Second, even in trials that are properly randomized, without error or subversion, a lack of the result of

## 14 *An evolution of comparative methodology*

allocation concealment can occur even when the trial claims allocation concealment (the process). Hence, baseline covariate imbalances across treatment groups can be systematic even in such trials. We call such systematic baseline covariate imbalances across treatment groups selection bias, although the term 'selection bias' has come to have many different meanings in different contexts (Mark, 1997).

Our interest in selection bias is confined to the type of selection bias that interferes with internal validity, or a fair and unbiased comparison of the treatment groups. The mechanism for this type of selection bias is most easily understood in the context of non-randomized designs, and especially self-selection designs. It is commonly believed that randomization by itself will eliminate this type of selection bias, but in fact, as we have seen and will explore further in later chapters, it does not. Moreover, such selection bias can occur even when the randomization was performed successfully, and not subverted. Of course, the randomization itself may be subverted too. For example, as we have seen, randomization may occur only after patient selection, and this can degenerate into what essentially becomes a *de facto* non-randomized trial.

### 1.9 RESIDUAL SELECTION BIAS

We refer to the selection bias that interferes with internal validity in randomized trials with patient selection preceding randomization as first-order residual selection bias, to distinguish it from its related form that occurs in non-randomized studies. It may be tempting to believe that simply performing the randomization in advance would eliminate all such selection bias, but this is not true either, as future allocations may be predictable. We refer to the selection bias that interferes with internal validity in *advance* randomized trials as second-order residual selection bias. It may be tempting to believe that allocation concealment would eliminate all such selection bias. Indeed, the objective of allocation concealment would eliminate all such selection bias, but again, the claim of allocation concealment refers to the process, and this is not sufficient to ensure that allocation concealment has achieved its objective. We refer to the selection bias that interferes with internal validity in advance randomized trials with imperfect or unsuccessful (subverted) allocation concealment as third-order

***Residual selection bias*** 15

residual selection bias. This will be the type of selection bias with which we are most concerned, as it is the only one that does not have a simple countermeasure that is generally known and utilized.

In Chapter 2 we will discuss, in greater detail, the mechanisms by which selection bias may occur even in properly randomized trials, at least as we define the term 'properly randomized'. In Chapter 3 we will provide some evidence that this type of selection bias actually occurs, and is not merely a hypothetical concern. In Chapter 4 we will discuss the impact one can expect selection bias to have on the results of trials. In Chapter 5 we will discuss measures (beyond randomization) that can be taken to prevent selection bias. In Chapter 6 we will discuss methods by which selection bias can be detected, or hopefully ruled out, from any given randomized trial. In Chapter 7 we will discuss methods that can be used to adjust for selection bias, in case it is found but useful between-group comparisons need to be salvaged anyway. In Chapter 8 we will summarize the overall recommendations for managing selection bias in randomized trials.

