

A Review of Methods for Ensuring the Comparability of Comparison Groups in Randomized Clinical Trials

Vance W. Berger*

National Cancer Institute, University of Maryland Baltimore County, Biometry Research Group, National Cancer Institute, Executive Plaza North, Suite 3131, 6130 Executive Boulevard, MSC 7354, Bethesda, MD 20892-7354, USA

Abstract: While different design features of medical studies ostensibly serve different functions, many fall under the umbrella of methods aimed at ensuring the comparability of the comparison groups. Randomization rightly occupies the top spot in the hierarchy of design types, as it eliminates some biases (that is, systematic differences in comparison groups) that no other design can claim to eliminate. It is often assumed, and sometimes even asserted explicitly, that randomization by itself suffices to ensure that the comparison groups are sufficiently comparable that they would differ only randomly, but two points need to be made in this context. First, the assertion is not true. Second, even if it were true, it would still not be a cause for complacency, because even random baseline imbalances can wreak havoc on the valid interpretation of randomized clinical trials. Additional methods, beyond randomization, are therefore seen to be essential to the design of a good randomized clinical trial. Such methods include masking, allocation concealment, restrictions on the randomization, adjustment for prognostic variables, and the intent-to-treat approach to data analysis. Masking aims to ensure that those individuals in any one group formed by randomization are treated as similarly as possible subsequent to randomization as those in any other group formed by randomization. In contrast, allocation concealment and restricted randomization aim to create groups that start off as comparable. Adjustment for prognostic variables aims to change the comparison groups themselves to make them comparable. For example, one might find gender to be both predictive of outcome and unbalanced across treatment groups, and so one would compare the treatment groups not overall but rather first only among females and second only among males. The intent-to-treat approach aims to keep similar groups similar by not allowing for patient selection based on post-randomization outcomes (including failure to comply with the protocol). The key to understanding masking, allocation concealment, and randomization is to recognize that none of them are binary phenomena, even though they are often incorrectly understood to be. So one must question how these methods are actually carried out, rather than contenting oneself with the vague statement that these methods were performed. This review will shed light on the distinction between the process and the outcome of each of these methods (masking, allocation concealment, and randomization), and will also consider issues related to adjustment for prognostic covariates.

1. INTRODUCTION

While many sophisticated design features are used in the evaluation of medical interventions, and justified by appeal to complicated arguments, in fact the key to an unbiased comparison is readily understood by appeal to a simple principle. Specifically, there should be no unfair advantages awarded to any of the comparison groups. It is precisely this simple principle which underlies many of the complex features of modern medical studies, and randomized clinical trials. The easiest way to see this is to ask what would happen without a particular design feature. This question will be explored for the specific design features of randomization, masking, allocation concealment, restricted randomization, and adjustment for prognostic covariates, and intent-to-treat. An excellent discussion of many facets of these design features can be found in [1].

2. Randomization

Randomization is often seen to be a cure for whatever ails any study. The term “randomized evidence” has gained

popularity, and suggests a level of rigor that is beyond reproach. After all, if a study was randomized, then that is the end of any discussion of biases. But how does randomization exert its influence so as to eliminate biases? And does it actually eliminate all biases, or only some of them? These questions were considered in the context of selection bias, in Section 1.6 of [2], where it was pointed out that a lack of selection bias can be phrased as neither the set of subjects nor the set of treatments affecting the identity of the other. That is, self-selection would create a bias, because those subjects who choose one treatment, for example an elective surgery, would likely differ systematically (possibly having greater disease severity) from those who choose another less invasive treatment.

While randomization is actually a collection of techniques, and not just one technique [3], most forms of randomization eliminate the bias caused by self-selection, by virtue of eliminating the self-selection itself. That is, most forms of randomization used in practice do not allow either the subject or the investigator to select the treatment. Rather, the treatment is assigned according to an allocation sequence prepared in advance. In this way, randomization eliminates one bias that could create imbalances in both measured and unmeasured prognostic covariates. See Section 1.7 of [2] for more details, and for the obvious next question – where does this leave us?

*Address correspondence to this author at the National Cancer Institute, Biometry Research Group, Executive Plaza North, Suite 3131, 6130 Executive Boulevard, MSC 7354, Bethesda, MD 20892-7354; Tel: (301) 435-5303; Fax: (301) 402-0816; E-mail: bergerv@mail.nih.gov

Are there other biases to consider? The existence of an allocation sequence created in advance opens up the possibility of subversion through the prediction of future allocations by either observing the allocation sequence directly or detecting patterns among the allocations already made. Prediction by pattern detection will be discussed in Section 5, and here only the direct observation of the allocation sequence is being discussed. In theory, the allocation sequence is to be kept hidden from all parties involved in the conduct of the trial, especially those investigators who enroll patients. This is because if investigators know the upcoming treatment assignment, then they can deny enrollment to prospective patients whom they (the investigators) do not want to receive the particular treatment that is to be allocated next [4]. In fact, there is evidence to suggest that such subversions, in which healthier patients are enrolled when one treatment is due up next and sicker patients are enrolled when the other treatment is due up next, occur in actual randomized clinical trials. See [5] and Chapter 3 of [2].

The design feature that would eliminate this type of selection bias is allocation concealment, which will be discussed in greater detail in Section 4. For now, though, it is enough to understand that randomization by itself does not eliminate all biases.

3. Masking

While randomization aims to create treatment groups that are comparable to begin with, masking aims instead to keep them comparable subsequent to treatment [1]. As in Section 2, we again seek to understand the benefits of a design feature by considering what happens in its absence. Given the complexities of ancillary treatments and care, it is not hard to see that without masking, even groups that start out identical to each other can diverge subsequent to treatment due to factors having nothing to do with the treatments themselves. This is because knowledge of the treatments administered can create expectations in both the patient and the investigator, and these expectations may drive the decisions to render additional treatments or not. Consider the difference between “You seem to be in pain, so I will administer some pain medication” (knowing that a placebo was given as the study medication) and “The discomfort you feel at the moment is temporary, so you probably do not need pain medication, right?” (knowing that the active treatment was given as the study medication).

This is not the only way in which knowledge of the identity of treatments already administered can create differences between the treatment groups. It is also possible for subjective measures to be scored differently, for patients to differentially seek out information and/or treatments depending on their confidence in their treatment, or for investigators to spend more time with patients in one group or the other. Friedman, Furberg, and DeMets [6; Chapter 6] discussed a trial of coronary bypass surgery *vs.* medical treatment [7], and the differential and time-varying nature of tobacco use across the two groups. Specifically, the two groups were balanced with respect to smoking at baseline, but during follow-up, smoking was more prevalent in the medical group. Any difference in outcomes between the

groups is then confounded, as it could be due to the treatments themselves, but it could also be due, in whole or in part, to tobacco use.

For these reasons, as well as others, masking is a standard design feature when it is feasible to keep the identities of the treatments from the patients and/or the investigators. When masking is not possible, and both the patients and the investigators know what treatment was administered to which patients, it is still best to use masked assessors, so that those investigators who evaluate patients and record data in the case report forms do so without knowing the identities of the treatments each patient received. See, for example, [8], in which masked assessors were used because there were exercise and vitamin D groups. It is also best, in an unmasked trial, to at least mask which of the treatments is the active one. For example, a placebo group could be supplemented with some informational brochure, and called the “information group”, to conceal its role as the control group. This step could minimize, but probably would not completely eliminate, the effect of unmasking, because at least none of the patients would be made to feel like they were unlucky and did not get the treatment they wanted.

Senn [9; Section 3.8] notes that 1) masking offers no protection against bias when the goal is to establish equivalence, instead of superiority; and 2) masking may be only partial, or “veiled”, in complicated trials with multiple control groups. That is, each of two treatments may have its own specific placebo group designed to match it, as well as possible. In this case, it may still be clear if a patient is in one pair (consisting of an active treatment group and its matched placebo) or another, and so any direct comparison of the two active treatment groups would most certainly not be a masked one.

4. Allocation Concealment

As discussed in Section 2, randomization is not enough, especially if upcoming allocations can be observed, because in this case, it is possible to enroll healthier patients when one treatment is due up next, and sicker patients when the other treatment is due up next. Allocation concealment is generally defined as the negation of this ability to observe upcoming allocations. That is, allocation concealment is generally defined as the inability to observe upcoming treatment allocations, thereby denying investigators the opportunity to selectively channel some patients into one treatment group and other patients into another treatment group. But this definition is not unique. Hewitt *et al.* [10], for example, define allocation concealment as adequate if the investigator(s) who recruit(s) patients differs from the one(s) who execute(s) the allocation sequence. Any definition of allocation concealment will bear some resemblance to that of masking, but the key difference is in the timing. While masking requires ignorance of the treatment identities until all evaluations are recorded and the database is locked, allocation concealment requires ignorance of each treatment identity only until it is assigned.

Perfect masking implies allocation concealment, but the reverse implication does not hold. In fact, some would even say that while masking is possible only some of the time,

allocation concealment is always possible. See, e.g., [11]. However, as pointed out in Section 2.5 of [2], this is not the case unless one uses distorted definitions of masking and allocation concealment. Both masking and allocation concealment, as well as any other process, may be defined as either the action or the result. Attempting to mask a study is a good idea, but it does not guarantee true masking (ignorance of the treatment identities). Likewise, the process of allocation concealment, whatever process is used, does not guarantee that the investigators have no knowledge of upcoming treatments. So the correct statement is that the process of masking is always possible (simply do not tell patients or investigators who received which treatments, even if they can figure it out for themselves anyway), and the process of allocation concealment (in fact, more than one such process) is always possible. But neither result can always be guaranteed.

It is possible to guess which treatment was received based on distinguishing tastes, side effects, smells, or other characteristics of the treatments; so true masking cannot ever be ensured. The mechanisms for subverting allocation concealment are more subtle, and will be described in terms of an unmasked trial, although the same phenomenon may occur also in partially masked trials (even those that are labeled as masked, with no mention of any unmasking that may have occurred during the trial). This discussion is deferred until Section 5, so that restricted randomization can be discussed first.

5. Restricted Randomization

Unrestricted randomization, in which each allocation is independent of the other allocations (such as would be obtained by flipping coins repeatedly), is not used in practice very often. One reason for this may be the possibility of chronological bias, which occurs when the distribution of some patient characteristic varies over time, and then gets confounded with treatments if the treatments are unbalanced over time. For example, if a trial is conducted in a college town between July and October, then the age distribution would be expected to change during the course of the study. This is because there would be more older patients early in the trial (when many of the college students are away for the Summer), and more younger patients later in the trial (when school is back in session). Even if the treatment allocations are balanced at the end of the trial, there is still confounding if there is an unbalance in early September, because whichever treatment group was allocated more during the Summer months will tend to have older patients, and the other treatment group will tend to have younger patients. See Section 5.3.1 of [2], or [12], for more information on chronological bias.

To prevent chronological bias, and also to ensure balance at the end of the trial, most trials use restricted randomization, in which balance is forced both at the end and at intermediate points during the trial. The most common form of restricted randomization is the permuted blocks procedure, which is described in detail in [13]. See also [1; Section 10.8], [2; Section 5.3.1], [6; Chapter 5], and [12]. Briefly, the idea is to require perfect balance in the numbers allocated to each treatment group within each block, and,

therefore, overall in the trial after the completion of each block. But there is a price to pay for this avoidance of chronological bias. Specifically, allocation concealment should now be reconsidered in light of a second threat to it. Not only can future allocations be observed directly by those with access to the allocation sequence, but also future allocations can be predicted by those without access to the allocation sequence.

Consider, for example, permuted blocks of size two. Then each pair of consecutive patients with the even accession number exceeding the odd accession number by one (so, for example, Patients #1 and #2 taken as a pair, or Patients #5 and #6 taken as a pair, but not Patients #6 and #7) must be assigned to different treatment groups. Even without having ever directly observed the allocation sequence, one can still infer that the second allocation will be to the control group in a two-arm trial if the first allocation was to the active group. With blocks of size two, prediction of future allocations is tantamount to direct observation, because any knowledge so gained would be perfect knowledge. This is not the case, however, with larger block sizes.

For example, with blocks of size four, and two treatment groups (say A and B), the six possible block configurations are AABB, ABAB, ABBA, BAAB, BABA, and BBAA. The fourth allocation in any of these configurations would be known with certainty, given knowledge of the previous three allocations, as would the third allocation of the AABB and BBAA configurations, given knowledge of the first two allocations. However, the second allocation would not be known with certainty for any configuration, and yet (imperfect) prediction is still possible by virtue of the fact that the second allocation is more likely to disagree with the first allocation than it is to agree with it. This being the case, allocation concealment must be viewed as somewhat of a continuous phenomenon, as opposed to the binary phenomenon it is usually taken to be. That is, “the degree of blinding [masking] guaranteed cannot exceed the degree of randomization employed” [9; Section 3.8]. Once it is recognized that there can be an extent of allocation concealment, it becomes easier to evaluate some restrictions on a randomization procedure as being better than other restrictions.

For one thing, smaller block sizes will better control chronological bias, but at the expense of allowing greater prediction. A solution was found that appears to address both prediction and chronological bias, specifically varying the block sizes. However, this solution does not generally achieve what it purports to. Specifically, in some cases, the randomized blocks procedure with varied block sizes can result in greater prediction of future allocations than the corresponding procedure with fixed block sizes [2], [12]. This is because sometimes a pattern will emerge that will allow one to deduce the block size, and also because one does not need to know the block size to be able to predict the upcoming allocations with reasonably good odds [9; Section 3.8].

The maximal procedure, which is less restricted than the randomized block procedure (with either fixed or varied block sizes), tends to minimize prediction better than the

randomized block procedure while matching the level of control over chronological bias [2; Section 5.3.4], [12]. The idea is to still restrict the randomization, but not nearly as much as randomized blocks would. Given that the maximal imbalance is controlled, there seems to be little point in further requiring perfect balance after every second or fourth allocation. The maximal procedure drops these excessive restrictions.

6. Covariate Adjustment

Generally, covariates are attributes of patients that are predictive of response. For example, severity of disease at baseline, or some marker thereof, would be both a patient characteristic and predictive of subsequent response. Even in a randomized trial, disease severity may be unbalanced across treatment groups, with one treatment group having predominantly sicker patients, and another treatment group having predominantly healthier patients. In such a case, a comparison of the treatment groups that does not account for the difference in severity of disease would not be a valid one, because even in the absence of any treatment effect, one treatment group would be expected to demonstrate better responses than the other. One solution to this problem is to adjust the analysis (as opposed to the randomization; this is referred to as post-stratification, as opposed to pre-stratification [14; Section 6.3]) for disease severity. The mechanism for doing so depends on the nature of the covariate, the simplest method applying to binary and nominally scaled covariates, such as gender or race. For such covariates, adjustment (post-stratification) consists of comparing the treatment groups separately within each level of the covariate, and combining the resulting measures of treatment effect.

For example, one could compare treatments just for men, and then just for women. The theory here is that while the treatment groups may not be comparable, they may be comparable among only men, and likewise among only women. So instead of trying to force the overall groups to be comparable, as the aforementioned methods would do, adjustment tries to restrict consideration to those subgroups that already are comparable. The overall analysis would consider all subgroups, or, in our example, with gender as the only covariate, both men and women, but it would do so by taking some sort of weighted average of gender-specific measures of treatment effect. It is also possible, of course, to simply present only the gender-specific analyses, and never combine them. Because of the inherent variability in the data, however, more robust estimates generally result from making use of as much data as possible. When permutation tests are used to avoid the need to make unreasonable assumptions, there are actually several ways to adjust for a binary covariate. These include restricting the permutations to fix the level of covariate imbalance at its observed value, using an adjusted test statistic, or both [15].

More complicated adjustment methods are required when the covariate is ordered. If the response variable is taken to be normally distributed (in the real world, no variables have a normal distribution [16], although many variables are taken to have normal distributions), then it is common to adjust for continuous covariates with the analysis of covariance

(ANCOVA), which specifies a common variance for the response variable regardless of the treatment group or the covariate values. However, the mean of the response variable is taken to be linearly related to the covariate, or some transformation of it. Because of this presumed linear relationship, the responses are compared even across the covariate values, whereas this was not the case with the adjustment methods for binary and nominal covariates. Nonparametric alternatives to the ANCOVA are also available [17]. See Chapter 7 of [9] for more information on ANCOVA.

If the covariate is ordered, but not continuous, then more specialized methods are required. One possibility is to consider the pairs of covariate and response values, and then determine which of these pairs are more indicative of a treatment effect than others [18]. It will generally turn out that only a partial ordering is possible, because some pairs of values will not be comparable to others. This obstacle is not insurmountable, and the analysis can proceed anyway, for example, by comparing each patient in one treatment group to each patient in the other treatment group, and ignoring both the ties and the cases in which the respective outcomes are not comparable. Each comparison that remains would favor one treatment group or the other, and one can base the analysis on how often each of the treatment groups is favored. See [18] for further details.

It was mentioned earlier in this section that covariates tend to be patient characteristics, or attributes, but they need not be. Recall that selection bias is possible even in randomized trials, especially if the trial is unmasked and the randomization is restricted. This is because in such a case, it is possible to predict the upcoming allocations. Consider, for example, blocks of size two, so that each block is either AB or BA. Assuming that the block type cannot be foreseen, the investigator would have no way to know the treatment to be assigned to any patient who is first in his or her block. If the treatments are equally likely, then the likelihood of any such patient receiving treatment A is 50%. In an unmasked trial, the investigator knows the treatment assigned to the first patient in a block before the second patient in that same block is identified.

As such, if the first patient in the block received A, then the second has a 0% chance of receiving A. Conversely, if the first patient in the block received B, then the second has a 100% chance of receiving A. This chance of receiving A (conditional on the previous allocations), which was discussed in [2, 4], and Section 3.8 of [9], has been called the reverse propensity score (RPS) [19], and can itself be used as a covariate. Of course, the RPS is somewhat of an unusual covariate, in that it is more a credential than an attribute. That is, if the RPS is in fact predictive of future responses, then this is more similar to credentials (MVP awards in sports or high grades in school) predicting future success but not causing it. The analogy to the more usual covariates (attributes or characteristics) would be intelligence for school outcomes and athleticism for sports outcomes. These may actually cause success, whereas clearly credentials (occurring after the success, and in fact based upon that success) cannot. Still, the RPS is a useful predictor, because if there is selection bias, then patients with better values of the usual

covariates are selected when the RPS value is high, and patients with worse values of the usual covariates are selected when the RPS value is low. In this sense, the RPS summarizes the other covariates, and may even be redundant given these other covariates, except that some other covariates may not be measured. If there is no selection bias, then the RPS is a useless predictor, and there is no need to adjust for it.

In fact, the extent to which the RPS predicts the outcomes within each treatment group forms the basis of the most direct test for selection bias, the Berger-Exner test [2; Section 6.5], [4]. If selection bias is found, then the treatment groups may be unbalanced even after adjusting for the observed covariates, and so it is a good idea to salvage the trial by adjusting for the RPS, either in addition to or instead of other covariates.

7. Intent-To-Treat

Suppose that in a given trial the combination of randomization, masking, and allocation concealment worked together to produce perfectly balanced groups at baseline, and that subsequent to baseline, these groups were treated identically. Suppose further that the active treatment group causes a particular adverse reaction in 40% of the patient population (so these patients drop out before having the opportunity to respond), but is effective for 30% of those patients not experiencing this adverse reaction. Finally, suppose that the control treatment is effective in 20% of the patients. Now, overall the active response rate is $(0\%)(40\%)+(30\%)(60\%)=18\%$, so the control group is actually more effective than the active group. But considering only the compliers in the active group would give the misleading impression that its response rate was actually 30%. The selection based on post-randomization events (in this case, compliance) created an imbalance, in which the healthiest patients in the active group were compared to the set of all patients in the control group. This is not a valid procedure [9; Section 7.2.10].

The intent-to-treat approach [9; Chapter 11] to data analysis involves including in the analysis all subjects who were randomized, regardless of whether or not they complied or even provided any data. If no data were provided for a particular patient, then some sort of data imputation would be required. In some cases, this is not problematic, as, for example, if a blood pressure measurement is missing because the patient died prior to the measurement being taken. Clearly, death is worse than any blood pressure outcome, and so if a rank-based analysis is to be used, then the worst rank would be assigned to the death [20]. The intent-to-treat approach does not concern itself with contamination (a control patient obtaining and taking the active medication against the protocol regulations). Treatment groups are defined by randomization only. In this way, the baseline balance is protected against differential drop-out rates, among other things.

SUMMARY

One cornerstone of good research is that comparison groups be comparable. Much effort goes in to ensuring that this is the case, and in fact, some of the processes aimed at

ensuring comparability have become so engrained in modern research that many researchers use them by rote, without understanding what these tools actually can and cannot provide. Randomization is one good example, as many have some sense that randomization ensures validity, but cannot go much further than that in terms of explaining how or why. This ignorance is not benign, as it gets in the way of a true understanding of the limitations of randomization. There are also limitations to masking and allocation concealment, which are often used to help randomization bring about comparable groups, in that the effort to mask treatment identities or conceal allocations is rarely completely successful. Covariate adjustment is used along with randomization, but does not produce groups that are comparable. Rather, the goal here is to find subgroups that are comparable, and restrict treatment comparisons to these groups. Covariate adjustment is also limited, because it is not possible to adjust for a covariate that is not measured. The limitations of randomization, masking, allocation concealment, and covariate adjustment are not reasons not to use these measures. However, candid consideration needs to be given to the extent to which these methods, alone or in combination, have produced the desired result of comparable comparison groups.

REFERENCES

- [1] Meinert CL. *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York, 1986.
- [2] Berger VW. *Selection Bias and Covariate Imbalances in Randomized Clinical Trials*, John Wiley and Sons, Chichester, 2005.
- [3] Berger VW, Bears JD. When can a clinical trial be called "randomized"? *Vaccine* 2003; 21: 468-472.
- [4] Berger VW, Exner DV. Detecting selection bias in randomized clinical trials. *Control Clin Trials* 1999; 20:319-327.
- [5] Berger VW, Weinstein S. Ensuring the Comparability of Comparison Groups: Is Randomization Enough? *Control Clin Trials* 2004; 25: 515-524.
- [6] Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*, Third Edition. Mosby, St. Louis, 1996
- [7] European Coronary Surgery Study Group. Coronary-Artery Bypass Surgery in Stable Angina Pectoris: Survival at Two Years. *Lancet* 1979; 313: 889-893.
- [8] Latham NK, Anderson CS, Lee A, Bennett DA, Moseley A, Cameron ID. A randomized, controlled trial of quadriceps resistance exercise and vitamin D in frail older people: The Frailty Interventions Trial in Elderly Subjects (FITNESS). *J Am Geriatr Soc* 2003; 51: 291-299.
- [9] Senn S. *Statistical Issues in Drug Development*. John Wiley and Sons, Chichester, 1997.
- [10] Hewitt C, Hahn S, Torgerson DJ, Watson J, Bland JM. Adequacy and Reporting of Allocation Concealment: Review of Recent Trials Published in Four General Medical Journals. *BMJ* 2005; 330: 1057-1058.
- [11] Forder PM, Gebski VJ, Keech AC. Allocation concealment and blinding: When ignorance is bliss. *Med J Aust* 2005; 182: 87-89.
- [12] Berger VW, Ivanova A, Deloria Knoll M. Minimizing predictability while retaining balance through the use of less restrictive randomization procedures. *Stat Med* 2003; 22: 3017-3028.
- [13] Berger VW. Quantifying the Magnitude of Baseline Covariate Imbalances Resulting from Selection Bias in Randomized Clinical Trials (with discussion). *Biomet J* 2005; 47: 119-127.
- [14] Fleiss JL. *The Design and Analysis of Clinical Experiments*, John Wiley and Sons, New York, 1986.
- [15] Berger VW. Nonparametric Adjustment Techniques for Binary Covariates. *Biomet J* 2005; 47: 199-205.
- [16] Geary RC. Testing for Normality. *Biometrika* 1947; 34: 209-242.

- [17] Brunner E, Dette H, Munk A. Box-Type Approximations in Nonparametric Factorial Designs. *JASA* 1997; 92: 1494-1502.
- [18] Berger VW, Zhou YY, Ivanova A, Tremmel L. Adjusting for Ordinal Covariates by Inducing a Partial Ordering. *Biomet J* 2004; 46: 48-55.
- [19] Berger VW. The Reverse Propensity Score To Manage Baseline Imbalances in Randomized Trials. *Stat Med* 2005; 24, in press.
- [20] Lachin JM. Worst-rank score analysis with informatively missing observations in clinical trials. *Control Clin Trials* 1999; 20: 408-422.

Received: June 09, 2005

Revised: June 27, 2005

Accepted: June 30, 2005