Clinical Trials: II. Randomization and Sample Size

Barbara Tilley
Anthony Schork

Follow this and additional works at: https://scholarlycommons.henryford.com/hfhmedjournal

Part of the Life Sciences Commons, Medical Specialties Commons, and the Public Health Commons

Recommended Citation
Available at: https://scholarlycommons.henryford.com/hfhmedjournal/vol33/iss4/14

This Article is brought to you for free and open access by Henry Ford Health System Scholarly Commons. It has been accepted for inclusion in Henry Ford Hospital Medical Journal by an authorized editor of Henry Ford Health System Scholarly Commons.
Clinical Trials: II. Randomization and Sample Size

Barbara Tilley, PhD,* and Anthony Schork PhD†

This second in a series of articles focuses on clinical trials. Information is provided on determining sample size and on methods of randomization including simple and single- and double-consent randomization, and blocked, stratified, and adaptive procedures to randomize patients to a study group.

A clinical trial is a "prospective study comparing the effect and value of interventions against a control in human subjects" (1). Clinical trials usually include at least two interventions. Ideally, the members of the groups undergoing the interventions should be comparable in every way except for the treatments they receive. Subjects should be followed up from a well-defined starting point to a well-defined ending point. Subjects are usually randomly assigned to the different treatment groups, and a sample size is chosen to ensure that enough patients are followed up to reach a definitive conclusion about the efficacy of each intervention. Additionally, clinical trials are usually performed in a blinded fashion, ie, neither the clinician providing nor the patient receiving the treatment knows which treatment is being used. The focus of this article is on those aspects of clinical trials relating to randomization and sample size.

Randomization

The following appeared in the New England Journal of Medicine (2): "Somewhere between 1910 and 1912 in this country, a random patient, with a random disease, consulting a doctor chosen at random, had, for the first time in the history of mankind, a better than fifty-fifty chance of profiting from this encounter." This statement could be considered a tender jibe at the process of randomization in which statisticians place so much faith. However, randomization fulfills an essential role in the design and consequent validity of a clinical trial. Randomization means that the choice of treatments for a study patient is determined by a chance or random process.

Randomized designs have several advantages (1). Randomization removes the potential bias in allocating subjects to a treatment regimen. For example, when the benefit of exercise after myocardial infarction was being studied, patients had several clinic visits prior to random placement in either the control or the treatment group. The staff involved in the study strongly believed in the benefit of exercise. Some had "favorite" patients and were disappointed when some of these patients were not assigned to the exercise program. If the staff had influenced assignments, it is possible that all favorite patients would have been assigned to the exercise program.

Another advantage of randomization is in assuring that variables (such as age, sex, and health status) are on the average evenly balanced among the treatment regimens. For example, if instead of using randomization all patients who are most ill are given treatment A, and all remaining patients are given treatment B, then treatment A could appear to be less effective than treatment B, because the patients receiving treatment A were more ill at the beginning of the study.

In addition, random assignment to a treatment group guarantees that statistical tests will have valid levels of significance.

Randomization also has disadvantages. Some physicians are concerned that randomization undermines the doctor-patient relationship. Randomization is less acceptable to the patient than assignment based on individual therapeutic consideration, and patients sometimes refuse to understand that they are being assigned in a random manner even when they are informed. Figure 1 shows the traditional method of randomization. To minimize patient concerns, Zelen (3) proposed single- and double-consent randomizations. When the method of single-consent randomization is used, consent is not sought if the patient is assigned to the control group and the control group is receiving standard care. Consent is required if the patient is randomly placed in the treatment group (Fig 2). In analyzing the trial, patients randomized to treatment are members of the treatment group whether they have consented and received treatment.
because parents are likely to give consent for the new treatment. In the double-consent randomization, patients are randomly assigned treatment group A or B and then asked whether they wish to receive treatment A or B (Fig 3). Again, when doing the analysis, patients are maintained as a part of the group to which they are randomized, regardless of treatment assignment. The Table gives the efficiency of these designs. If the probability of accepting the treatment is high, as in the newborn study, then only a small number of additional patients are required in order to do this type of study. However, as the probability of accepting the assigned treatment decreases, the sample-size requirement increases. If patients have only a 60% probability of accepting the treatment to which they are assigned, 25 times more patients are required to perform the trial in the double-consent design, and 2.8 times more patients are required in the single-consent design. Also these designs do not allow blinding.

Randomization has another disadvantage, especially if the total number of patients in the study is small. Randomization provides only a balance of prognostic variables on the average. Sometimes, by chance, unequal numbers of patients with the same prognostic variables are assigned to the treatment groups. Suppose, for example, that most male patients were randomized to treatment A by chance. If men do not respond as quickly to treatment as women, it may be incorrectly concluded that treatment A is inferior to treatment B.

Finally, implementation of the actual randomization process can be cumbersome because the allocation of treatments should be unpredictable. The clinician should have no knowledge of which treatment the next patient will be randomized to receive. Once a randomization list is developed, it should not be made available to the person performing the study. Instead, the list should be kept by someone outside the study who will give out the treatment assignment after receiving some patient-identifying information. In the exercise treatment example, it is easy to see that if treatment assignments were known and two patients became eligible on the same day, some bias could occur in the treatment assignment.

Other alternatives include numbered envelopes containing the treatment assignments that are given to the investigators. These envelopes must be truly opaque and well sealed. Curious staff members have been known to hold envelopes up to strong light to see the contents. It is also important to record the envelope number as a part of the patient data. Later these numbers should be checked against date of study entry to assure that the order of treatment assignment was maintained. Another approach used by the Beta Blocker Heart Attack Trial (BHAT) was an automated system. Data that defined eligibility (age, other medical conditions, blood

---

**Table**

**Efficiencies of Single- and Double-Consent Randomized Design**

<table>
<thead>
<tr>
<th>Probability of Acceptance</th>
<th>Single Consent Break Even Accrual Factor¹</th>
<th>Double Consent Break Even Accrual Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>4</td>
<td>—</td>
</tr>
<tr>
<td>0.60</td>
<td>2.8</td>
<td>25.0</td>
</tr>
<tr>
<td>0.70</td>
<td>2.0</td>
<td>6.2</td>
</tr>
<tr>
<td>0.80</td>
<td>1.6</td>
<td>2.8</td>
</tr>
<tr>
<td>0.90</td>
<td>1.2</td>
<td>1.6</td>
</tr>
<tr>
<td>0.95</td>
<td>1.1</td>
<td>1.2</td>
</tr>
</tbody>
</table>

¹ The number given multiplied by the proposed sample size for the trial is the number of patients to be studied. For example, if the probability of acceptance is 0.5 and the original sample size was 50 per group, 200 per group would be needed for the single consent, and the study could not be done with a double-consent design.

---

*Zelen (3)*

---

1. The number given multiplied by the proposed sample size for the trial is the number of patients to be studied. For example, if the probability of acceptance is 0.5 and the original sample size was 50 per group, 200 per group would be needed for the single consent, and the study could not be done with a double-consent design.
pressure, etc) would be provided from clinical centers to the coordinating center and entered interactively into a computer. After checking criteria and determining that the patient was eligible, the computer generated the treatment assignment and recorded the assignment in the data base for that patient.

**Performing randomization**

Several different methods are used to randomize patient treatments. The easiest to apply is *simple randomization*, which proceeds for all subjects without restriction. Randomly generated numbers from tables (5) or from automated random-number generators can be used to assign patients to treatment groups. Random numbers could also be generated by rolling a die or dice. For the two-treatment case, treatment assignment could also be generated by tossing a coin. If simple randomization is to be used and two treatments compared, the odd numbers (or heads on a coin) could represent treatment A and the even numbers (or tails on a coin) treatment B. Simple randomization is easy to implement, but the groups could be unbalanced in small studies.

*Blocked randomization* is used to ensure that after a specific number of patients are entered, the same number of patients is assigned to each treatment group. If the block size is eight, and there are two treatments, then four of every eight patients entered would be on treatment A and four on treatment B. Blocked randomization has two important advantages. If the type of subject recruited for study changes during the entry period, blocking will produce more comparable groups, and if the trial is terminated at any time, balance will still be achieved. There are two disadvantages. The investigator might be able to break the code if the length of the block is known. For example, if seven patients had been entered in a block known to contain eight patients, and three were assigned to treatment A and four to treatment B, the investigator would know that the next patient would receive treatment A. Thus, it is advisable to vary the length of the block randomly. Also, analysis of data can be more cumbersome if blocked randomization is used. However, most investigators ignore the blocking and proceed with standard analyses. By taking this approach, the true significance level is probably lower than computed, which means there will be a greater chance of failing to find a difference between treatment groups when a true difference exists (6).

*Stratified randomization* is used to ensure that possible prognostic factors that might influence the outcome of treatment are balanced among treatments (7). In this procedure, patients are divided (stratified) into homogeneous groups (eg, by sex, age, race, or some other factor) before randomization; then, usual randomization procedures are applied to subjects within each stratum. Note that only important variables should be used to stratify to avoid groups with very sparse data. It is possible to adjust for some imbalances in prognostic variables during data analysis (8). (Analysis will be discussed in the next article of this series.) Stratification on area of infarct (anterior, inferior) and time of infarct (less than three hours, three hours or more) was used by the western Washington randomized trial for the use of intracoronary streptokinase (9). In a multicenter clinical trial, the clinical center is often an important factor for stratification.

BHAT (10) and the Lipid Research Clinic study (11) used both blocking and stratified random sampling. In BHAT, patients were stratified only by clinical center, and block size was randomly varied with four, six, or eight patients per block.

**Treatment allocation**

Usually patients are allocated evenly among treatments. This is called *fixed allocation*. For example, if there are two treatments, A and B, 50% of patients would be allocated to treatment A, and 50% to treatment B. The allocation probabilities are not altered as the study progresses. However, for some studies, investigators may want a fixed allocation that assigns more than 50% of patients to a particular treatment or treatments. For example, if there is sufficient information about toxicity in the control group but little information about toxicity on the new therapy, a two-to-one allocation (ie, two patients randomized to treatment for every one patient randomized as a control subject) may be chosen (12). This allows more information to be gathered about the new intervention.

*Adaptive randomization* procedures are used to change allocation probabilities as the study progresses. Adaptive procedures are applied when the changing probabilities result from information about either prognostic variables measured at entry to the study (baseline) or the response variables.

Baseline adaptive randomization procedures include biased coin randomization (13), which balances the number of subjects on each regimen based on previous assignments, and minimization randomization (14), which changes assignment probabilities as a function of the distribution of the prognostic factors for subjects already randomized. When biased coin randomization is used, the proportion of patients on each treatment is calculated after a new patient is randomized. The goal is to have equal numbers of patients in each group. If either proportion exceeds a prespecified amount, for example 0.6, the allocation probability is changed from 0.5 to 0.67 to increase allocation to the group with fewer patients.

Minimization is an extension of biased randomization. As an example, let the prognostic variables of interest be age
group, gender, and previous history of myocardial infarction. Suppose 20 patients have already been entered in a trial, and the distribution between treatment and control groups is as follows:

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Previous MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;50</td>
<td>5</td>
<td>Never</td>
</tr>
<tr>
<td>50-69</td>
<td>3</td>
<td>&gt;5 years ago</td>
</tr>
<tr>
<td>≥70</td>
<td>4</td>
<td>≤5 years ago</td>
</tr>
</tbody>
</table>

The effect of adding a new patient to each group is assessed by computing the sum of the absolute values of the differences (differences without the sign) and choosing the assignment that gives the smallest total difference. If a new patient (age 70, male, and never had an MI) is assigned to treatment, the absolute value of the difference for age 70 would be 2 (5-3); for gender, 0 (10-10); and for previous MI, 0 (7-7).

The total difference would be 2 (2 + 0 + 0). If the patient was assigned as a control, the absolute value of the difference for age would be 0 (4-4); for gender, 2 (9-11); and for previous MI, 2 (6-8). The total would be 4 (0 + 2 + 2). Thus, to minimize differences between groups, the new patient would be assigned to treatment. As the number of variables to be considered and number of study groups increase, calculation becomes more complex. Also, some variables may be considered more important than others, so weights could be added. Because of this complexity, minimization randomization is usually done by computer.

These techniques protect against a severe imbalance in either the number of patients per treatment or in the prognostic variables. As noted, these methods are cumbersome to implement, and minimization requires special methods of data analysis to take this type of randomization into account (15). If baseline adaptive procedures are used, it is generally unnecessary to block. If prognostic variables are used to alter patient assignment, stratification is usually unnecessary.

Response adaptive randomization procedures use information on subject response to intervention during the course of the trial to determine allocation of the next subject. These include “play-the-winner” randomization (16), a process that calls for staying with the winning intervention until a failure occurs and then switching to the opposite intervention. An extreme example of this type of randomization was the previously mentioned study in newborns (4). Because the treatment was so successful, only one patient was randomized to the control group. “Two-armed bandit” randomization allows updating the probability of success as soon as the outcome for each subject is known. This process is used to adjust the probabilities of allocation so that a higher proportion of future subjects receive the currently better intervention or treatment. To use this approach response by treatment must be known early in the trial.

Sample Size

In addition to deciding on a randomization scheme, it is important to consider sample size when designing a clinical trial. According to Friedman et al (1), “clinical trials should have sufficient statistical power to detect differences between groups considered to be of clinical interest. Calculation of sample size with provision for adequate levels of significance and power is an essential part of planning.” Several key features in determining the size of sample needed for a clinical trial should be considered in detail, using as much clinical data as possible to assist in computation of number of subjects. First, investigators should determine the most important outcome or response variable. The distinction as to whether this variable is measured on a discrete (usually dichotomous, eg, success/failure, lived/died, or yes/no) or continuous scale (eg, change in lung capacity) must be understood. Second, the investigators should estimate the smallest difference between the interventions that they wish to detect based on the trial. In determining this difference, the distinction between a statistically and clinically significant difference should be understood. It is possible to design a trial with a large sample size to detect a small treatment difference. If only a large treatment difference is of clinical importance, then the study could be done at less expense with a smaller sample size. Third, the investigators should state the hypothesis to be tested, usually a null hypothesis implying that the response will be the same for all interventions. The alternative hypothesis should also be clearly stated. Does the alternative state that the treatments are different, implying two-sided tests, or that one treatment is better than the others, implying a one-sided test? Fourth, the allowable alpha (probability of rejecting the null hypothesis when it is true, or probability of a Type I error) and beta (the probability of failing to reject the null hypothesis when it is false, or probability of a Type II error) should be determined. Power (the probability of rejecting the null hypothesis when it is truly false) is one minus beta. In clinical trials a Type I error is made if the investigator concludes two treatments are different when, in truth, the treatments have the same effect. A Type II error is made when the investigator fails to find differences between two treatments, but in truth, the treatments are different. Frequently alpha is set at 5% (0.05 level) and the power at 80% (Beta = 20%). Because it is unlikely that drugs from a negative trial will be retested, beta is set to 1% or 5% in Phase II chemotherapy trials to avoid missing an effective drug. Fifth, the variability of the primary outcome
Randomization and Sample Size

or response should be determined from the literature or from a pilot study. If the response is a continuous measure, this variability is usually expressed as a standard deviation or standard error. The variability of the response can greatly affect the sample size. In a study of chronic obstructive pulmonary disease (COPD) (17), the variability in measurement of pulmonary function was unexpectedly extremely large, and as a consequence, the results of the trial were equivocal. Sample size for a new trial of the effect of smoking cessation on pulmonary function for subjects with COPD was estimated to be as high as 11,000 patients, based on the variance from this previous study.

Special formulas are available to assist the investigator in estimating requisite sample sizes per intervention group. The choice of formula depends on the outcome to be measured. Lachin's review article (18) gives an excellent overview of sample size calculations for clinical trials including calculations when the outcome is a proportion or a continuous measure. Methods have been developed (19) to take into account the time to the outcome event (eg, time to death) when this is important. All of the above sample size calculations can be done using an interactive software package (STPLAN) (20) that is available in many institutions across the country.

Schork and Remington (21) have developed formulas for computing sample size when drop-outs (patients who refuse to participate in the study) are expected to be a problem, as often happens in studies with long periods of follow-up. In BHAT (10), both drop-outs and drop-ins were a concern. Before the beginning of the trial, propranolol was becoming popular as a treatment after a new myocardial infarction, even though it had not been approved for this use in the United States. Thus, study investigators expected some patients who were randomized to the control group to become drop-ins when their private physicians put them on propranolol. In calculating sample size, the methods of Halperin et al (22) were adapted to take into account both drops-outs and drop-ins.

Recently, Schoenfeld (23) developed sample size calculations for a study where time to event for two treatments is being studied but where it is also necessary to adjust the groups for possible differences in prognostic factors. Methods (24) are also available for calculating sample size when the comparison is a historical control group. However, before using historical controls, the disadvantages of this type of comparison group should be considered (25).

Often trials are planned with no knowledge of the variability in the outcome measures or of patient acceptance of the new treatment (ie, the possible drop-out rate). In this situation, a carefully planned pilot study on a small number of patients may allow the investigator to plan a clinical trial more effectively.

Calculation of sample size and power is also essential to an ethical trial. If the study is intended to be a clinical trial but too few patients are studied to allow investigators to draw conclusions, then patients may have been needlessly subjected to an experiment. A dramatic example is the smoking cessation study. An investigator who planned to study the effect of smoking cessation on COPD with a sample size of 200 or even 300 per group could not come to any conclusion because of the variability in the outcome measure. An investigator who determines that the available sample size for a planned trial is too small has several alternatives including abandoning the study, increasing the recruitment period, or trying to interest other institutions in a cooperative trial, which was done with the COPD study.

At the other extreme, if the sample size is too large, more patients than necessary will receive the inferior treatment. Because of the concern that patients or animals may be subjected to unnecessary treatment in clinical trials when the data base is too large or too small, most human-subject and animal-care review committees consider sample size and power to be an ethical as well as a scientific issue. Giammona and Glantz (26) provide an excellent discussion of the role of such committees in evaluating the statistical design of research.

References


