Chapter 9
Statistics: Setting the Stage

M. Abdullah Arain, Adil H. Haider, and Zain G. Hashmi

9.1 Introduction

This chapter introduces fundamental statistical concepts in the design of clinical trials that must be considered early on in the planning phases of a project.

A clinical trial is a prospective experimental research study where participants are assigned to one or more intervention arms to assess the impact of those interventions on various predefined health outcomes [1]. Clinical trials are often conducted to definitively answer clinically relevant questions while overcoming many of the limitations of observational studies. In the hierarchy of evidence-based medicine (EBM), results from clinical trials are considered the highest-quality evidence to drive clinical practice [2]. However, shortfalls in study design, methodology, data analysis, and result interpretation can potentially undermine the value of these studies and reduce their applicability to the general practice. This chapter introduces fundamental statistical concepts in the design of clinical trials that must be considered early on in the planning phases of a project. Over the next few pages, we will dive into the concepts of study design in surgical trials, randomization, allocation sequences, allocation concealment, and blinding and will also touch upon non-randomized trials, pragmatic trials, and superiority and inferiority trials. Eventually, statistical errors and power of the study will be discussed along with sample size considerations.

M. A. Arain · A. H. Haider (✉)
Aga Khan University Medical College, Karachi, Pakistan
e-mail: adil.haider@aku.edu

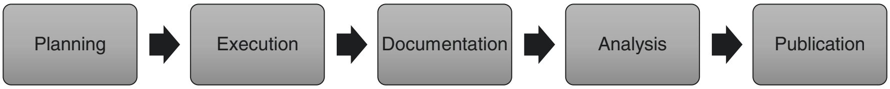
Z. G. Hashmi
Department of Surgery, Sinai Hospital of Baltimore, Baltimore, MD, USA
e-mail: mhashmi@lifebridgehealth.org

T. M. Pawlik, J. A. Sosa (eds.), *Clinical Trials*, Success in Academic Surgery,
https://doi.org/10.1007/978-3-030-35488-6_9

9.2 Setting the Stage

The two important considerations in the development of a clinical trial include (1) asking a clinically important question and (2) having a statistically sound study design to help answer that question. While statistical nuances must be appreciated and understood at each stage of the trial, ignoring them at the design stage can have disastrous downstream consequences. In this regard, only a carefully designed trial that advances clinical knowledge can justify its high costs and resource-intensiveness.

A clinical trial can generally be divided into five phases: planning, execution, documentation, analysis, and publication. The term study design is often used in medical literature to assign an appropriate type to the study while it in fact should refer to the overall planning of all the five phases that together yield a comprehensive publication [3]. Since conducting a trial is a resource-intensive and time-consuming affair, it is extremely disappointing to report less than ideal results just because enough attention was not paid while planning the study.



The Consolidated Standards of Reporting Trials (CONSORT) statement is a set of guidelines that was first introduced in 1996, with the goal of standardizing trial design and conduct. This includes a checklist and a flowchart aimed at standardizing the reporting of clinical trials and guiding the authors on how to be clear, complete, and transparent about their reporting. The statement has undergone several improvements in the last decade to not only aid the reporting of information but also help the readers, reviewers, and scientific journal editors to understand the trial's design, conduct, statistical analysis, and interpretation and critically appraise the publication. The checklist has become a standard of practice when reporting trial data and has been adopted by more than 500 medical journals. The statement also provides a uniformity to the clinical trial literature for later researchers to conveniently select reliable, relevant, and valid studies to include in systematic reviews and meta-analyses [4].

The statistical aspects of a study design include the type of study, sampling, collection of data, and measurement of outcomes/endpoints [5].

9.3 Study Design

Clinical trials fall under the experimental arm of analytical studies investigating the effect of an intervention on the study population. Randomized controlled trials, when designed, carried out, and reported appropriately, represent the most rigorous method of hypothesis testing and are a gold standard in assessing effectiveness of healthcare interventions. However, the results from these trials can suffer from residual biases, especially if adequate methodological diligence is not ensured [6].

Additionally, even though randomization greatly increases the validity of a study, not all clinical trials can be randomized. This is especially true for surgical trials, since the decision to undergo a surgical procedure, when an indication for that procedure exists, is not something that the patients would be willing to leave up to randomization [7]. Outcomes from inadequately designed trials cannot only misguide physicians making treatment decisions for patients at an individual level but also misinform policy makers devising a national public health policy [4]. To overcome these potential pitfalls, a sound understanding of both randomized and nonrandomized trials is necessary.

9.4 Randomization

Randomization is the process by which all the study participants possess an equal chance of being assigned to either the experimental or the control groups. This not only removes any selection biases that can potentially impact the outcomes by randomly distributing the patient characteristics between the groups but also equilibrates all confounding factors yielding a control group that is almost exactly congruent to the treatment group. Any disproportionate assignment to either group would skew the results, by introducing conscious and unconscious prejudices, and make the conclusion invalid. Therefore, any subsequent difference in outcomes between the groups can be demonstrated, after being evaluated by the use of probability theory and the level of significance between the different outcomes, to either be the result of difference in intervention or be merely due to chance alone [2, 7, 8]. Randomization also enables multiple levels of blinding (masking) of the intervention from the stakeholding parties like the researchers, participants, and evaluators [8].

There are many processes that can be used to randomize the participants. The aim of each process is to limit bias and to assemble a similar cohort of individuals between groups, and therefore the process should only be administered to the individuals who agree to participate in the study to ensure the purity of the process.

9.4.1 Fixed Allocation Randomization

Randomization by *fixed allocation* means that the assignments to the separate groups would be made at a predefined probability. Usually this proportion is set at an equal allocation (1:1); however, some situations may allow or even necessitate an unequal allocation (2:1). Some researchers argue that unequal allocation is not consistent with the true equipoise of RCTs and tends to introduce a bias to the results; however, others argue that a 2:1 allocation would have minimal effect on the power of the study but could potentially, with a fixed sample size, decrease the cost of the trial significantly. It can, in fact, increase the power of the study by registering a bigger sample size if the funding of the trial is fixed [7, 9]. For example, the

Scandinavian Simvastatin Study for coronary heart disease prevention, with a fixed sample size and a 2:1 allocation ratio, showed a 3% decrease in power while saving 34% of the cost [10].

9.4.1.1 Simple Randomization

Considered to be the most elementary of the randomization processes, sometimes being as basic as rolling a fair dice or tossing a fair coin, simple randomization conserves the absolute unpredictability and bias prevention of each intervention allotment and thus outclasses all other methods of allocation generation, irrespective of their sophistication and complexity. However, this unpredictability can sometimes become a challenge when the sample size is small because there is then a higher likelihood of disparity in group allocation by chance alone. This disparity diminishes as sample size increases [2, 8].

Fair coin tossing, dice rolling, and shuffled deck of cards dealing are examples of manual methods of drawing lots. These methods, though theoretically ideal for random allocation of intervention, are practically susceptible to non-random contamination. A series of tosses with identical outputs could entice the researchers to intervene with the result of the toss because they perceive the randomness of these results to be non-random. These methods also are difficult to implement and do not leave an audit trail and therefore are not the recommended methods of random sampling. Instead, sequence generation by a table of random numbers or computer generation of random numbers is reliable, unpredictable, reproducible, and easily traceable and should be confidently used in trials [8].

Many investigators have a less than ideal understanding of randomization and frequently assume non-random approaches like alternate assignment and haphazard sampling to be random. Quasi-random is a term commonly used to refer to the assignment of intervention groups based on pre-intervention tests, and while the term may have the word random in it, it serves no more than a misnomer for an approach which completely goes against the ideology of randomization. Assignments based on medical record numbers or date of birth (where odd numbers are placed in one group and even in the other) or alternate assignments (e.g., ABABAB) are non-random methods that are mostly mistaken as random. Systematic sampling should not be considered randomized sampling as well because the outcomes are not, theoretically and practically, based on chance alone. Any study not detailing its randomization process or defining its randomization by a non-random method should be approached with caution [8].

9.4.1.2 Restricted Randomization

Restricted randomization (also called blocked randomization) benefits the researchers who require an equal group size in between groups. It restricts large imbalances in sample size by influencing the acquisition of an allocation sequence that could lead to unequal group sizes. The most commonly used variation of restricted randomization is by random combinations of equal-sized blocks. Participants are examined in blocks of, for example, four individuals at one time. Using this block size will yield six possible combinations of 2 As and 2 Bs in each block (AABB, BBAA, ABAB, BABA, ABBA, BAAB). A random number sequence will then be utilized to select one out of the six blocks, and the sequence of allocation in that particular block is followed for the first four participants. Subsequently, one of these six block combinations will be randomly selected again and its allocation sequence followed for the upcoming four participants and so forth [2, 8]. The downside to this is that some studies could develop an extremely strict exclusion criteria, to keep the population pool comparable, but would end up markedly regulating the participant enrollment and jeopardizing the generalizability of the results [7]. The random allocation rule is another form of restricted randomization where the eventual group sizes would be equal. Often, after the selection of total sample size, a subset is assigned Group A and the remaining end up being Group B. It can be explained by placing in a bowl 100 balls labeled “A” and 100 balls labeled “B” for a total sample size of 200. Then one ball will be drawn at random without replacement and the participant placed in the corresponding group [8].

9.4.1.3 Stratified Randomization

While striving to remove selection bias, randomization tends to establish unwanted chance imbalances. These imbalances can be prevented by dividing the population into strata of prognostic factors (e.g., smokers and nonsmokers). These stratified groups would then undergo blocking randomization to finally yield separate block randomization sequences for the different combinations of prognostic factors. Stratification without blocking would serve no purpose. Even though it is a valid and useful method to curb chance imbalances, the complex stratification procedure provides little benefit in large-scale studies where the effect of chance selection eventually gets balanced on its own. The complicated process can become overwhelming during participant enrollment and sometimes can be a limiting factor for trial collaborations. A possible benefit of stratified randomization is stratification by center in multicenter trials. For small trials, stratification not only provides proper balance between sample groups but also raises the power and precision of the study [2, 7, 8].

9.4.2 Adaptive Allocation Randomization

Adaptive randomization allows for changes in the probability of allocation to the intervention groups with the passage of time. A problem commonly faced with small-sized trials is that simple randomization (with or without stratification of important prognostic variables) results in imbalance of covariates in the intervention groups. This can lead to incorrect interpretation of research outcomes [7, 11, 12].

Baseline adaptive randomization, like the covariate adaptive randomization (minimization), though in itself is of a non-random nature, is an acceptable and valid alternative mode of randomization in such scenarios. Here, all important prognostic factors (covariates) are identified before the initiation of the trial, and every new participant is assigned sequentially to the specific intervention group by relating their covariates to the already specified ones while keeping in mind the previous allocations to each group. This method achieves balanced groups with respect to numbers and covariates [7, 11, 12].

Response adaptive randomization allows for adjustments in group allocations by evaluating intervention responses. Of these, the "play the winner" style assigns the succeeding participant by relying on the previous participant's response to the intervention. A positive response places the successive participant in the same group, and a negative response shifts the next assignment to the other group. In the "two-armed bandit," the probability of positive results keeps on adjusting as the outcome for each participant is added to the count and more and more participants are added to the group with the superior intervention [7].

9.5 Allocation Concealment

After randomization of the study population by allocation sequences, the next crucial step is to implement the allocation sequence in an impartial and unbiased fashion by concealing the sequence until the patient assignments to the intervention group. Allocation concealment refers to the prevention of foreknowledge of the treatment assignment, thus shielding those who enroll participants from being influenced by this knowledge [4]. An unconcealed study defeats the entire purpose of randomization. Allocation concealment is a widely misunderstood concept. Many researchers delve into conversations about randomization techniques when discussing it and some consider it to be related to blinding. Both of these concepts are inaccurate. Allocation concealment, in reality, is the approach used to implement the allocation sequences generated by randomization. Individuals admitting participants to the study should enroll individuals without having any knowledge regarding allocation of intervention groups for the participants. Any such knowledge has a potential to introduce bias and exaggerate treatment effects producing greatly heterogeneous results than should be expected [2, 13]. Popular methods to conceal allocation include use of opaque envelopes to assign groups, a method not widely recommended, or the use of distance randomization, where allocation sequences are handled by a central randomization service and the investigators have to contact the service for each enrolled participant, placing a gap between recruiters and group allocators [2].

9.6 Blinding

After randomization, there would be a group receiving the (new) intervention and another being administered the control, which could be the already existing standard of care or a placebo. Clinical trials always pose a risk that the knowledge of the benefit of the intervention in the stakeholders (e.g., participants, investigators, analysts) can definitely introduce biases into the study and greatly impact the trial outcomes leading to unacceptable results. The participants, if informed of their assigned intervention, with a previous understanding of the advantages of that intervention, could be led to report positive results even if they did not feel a difference. The investigator, if informed of the intervention group, with a previous perception of potential drawbacks of the intervention, would be led to record negative results even if none actually existed. When analysts have knowledge about which groups they are analyzing, a previous inclination to either one could potentially lead them to over-analyze or under-analyze to produce results that agree with their inclination. Knowledge of the groups can potentially change the delivery of care to either group to adjust for a conceived limitation that the group suffers.

To curtail this probable bias, the trial can be blinded; that is, the participants, investigators, and data handlers can be prevented from knowing which participant belongs to which group, thus preventing the stakeholders from projecting their expected outcomes onto the actual results.

A trial may be single-blinded, where the participants in the group do not know details of their assigned intervention; double-blinded, where both the participants and investigators are kept unaware of the assigned intervention; or triple-blinded, where in addition to the participants and investigators, the data analysts are also kept ignorant of the assignments. Rapid un-blinding should be possible in the design of the study to counter any major harmful effects [2, 14, 15].

It is important to distinguish between allocation concealment, which happens before the randomization process is begun to nullify selection bias, and blinding, which happens post-randomization and reduces detection and performance bias. The overall bias reduction is more significant with allocation concealment than with blinding, and the best approach is to employ both when conducting a study [2, 16].

9.7 Non-randomized Studies

In non-randomly assigned controlled studies, two groups are analyzed against each other, one receiving the new intervention while other being the control. This is very similar to an RCT, except the groups are assigned in a non-random manner. Instances where such a study is acceptable could be when the practicality of large-scale administration of the study is a likely impediment or when the logistic requisites of a standard RCT cannot be fulfilled and concerns of costing and patient acceptability become the possible limiting factors. However, the study groups being evaluated still need to be comparable.

Non-randomized studies are discouraged in circumstances where numerous confounding variables are being assessed, the endpoints are multifactorial, or there is a lack of evidence in terms of what outcomes to expect. If the researcher is able to identify the confounders and adjust the analysis accordingly, the evidence produced by these studies would be acceptable given the constraints of performing an RCT in a similar situation.

9.7.1 Advantages of the Non-randomized Trial

The benefit of having a control group even in a non-randomized study can never be underestimated. The control group helps maintain the internal validity of the study by nullifying the impact of temporal trends (other aspects of disease and its care), the regression to the mean (outlying values moderating over time), and the learning curve of the surgeon on the outcomes of the study and facilitates the investigator to deal with these elements during the design of the study. The use of non-randomized control can also increase the generalizability of the study by enrolling a heterogeneous population spread across multiple providers.

9.7.2 Disadvantages of the Non-randomized Trial

The most elementary flaw of non-randomized trials is the confounding bias. The direction of this bias is quite variable and unpredictable. Bias introduced by just selectively registering healthier or sicker patients can turn the results in favor or against the intervention, respectively. Therefore, any “hand-picking” of the subjects must be avoided if possible. These studies also fail to account for social, cultural, economic, and clinical variables which have a potential to affect the outcomes. For a better internal and external validity of the study, it must be replicable and be adjustable to various clinical settings for it to have the desired impact [17].

9.7.3 Examples of Non-randomized Trials

Historical control studies trials are an example of non-randomized trials where the intervention group is compared to a previously assessed historical control group. In this method, everyone receives the intervention. However, they may be limited by changes in diagnostic/therapeutic approaches that accrue over time and thus inherent biases can arise. An example would be the difficulty ascribing a mortality difference to an intervention among patients with coronary artery disease versus historical controls.

Withdrawal studies involve placing participants off treatment to assess the actual benefit of a treatment which has never been proven to be of benefit but is somehow common practice. However, this only allows the most stable patients on the treatment to be selected for the study [7].

Concurrent trials include the crossover design where subjects serve as their own internal control. All subjects are used twice, once in the intervention arm and once in the control arm. Randomization for treatment sequence is also carried out. The major advantage of this method is to account for paired comparisons and mitigate variability secondary to inter-individual differences. Carryover effects are an important consideration for crossover studies. These are effects that "carry over" to the upcoming intervention of the trial from the previous intervention. To counter this effect, studies ensure "wash-in" and "wash out" periods for the succeeding and preceding interventions, respectively. Usually, if more than one intervention is being compared, then a Latin square matrix ($n \times n$) is utilized to ensure that every succeeding intervention is preceded or followed by any other intervention just once. It is believed that this fixed concatenation provides a better control over the carryover effects than by randomization [14].

Factorial study designs commonly involve two interventions to be assessed against the control and a 2 × 2 design is a commonly used factorial design. An important assumption that these studies make is that interventions X and Y independently have no interactions with each other [7].

Control	X + Y
X + Control	Y + Control

9.8 Special Considerations for Surgical Trials

Surgical trials can be classified as those evaluating minor changes in surgical technique, major changes in surgical technique, or surgical versus non-surgical treatments [7]. A common source of error in all these situations can be attributed to the inherent technical variability in the performance of procedures; attempts must be made to mitigate and account for these effects. This especially holds true for novel procedures where the trialist must account for the "learning curve," which entails to the experience of the surgeon on the new procedure which would impact patient selection, operative skills, post-operative care, and additional medical therapy. One way to prevent this is to postpone the initiation of such a trial until adequate expertise has been achieved. Failure to do so may result in an elevated risk of adverse effects and could potentially bias the final results against the new intervention [18].

The underlying principle of equipoise that there must exist a true uncertainty about an intervention's effect in order to justify a clinical trial is even more important for surgical trials. Many patients are concerned about enrolling in surgical trials, especially those involving a novel surgical procedure. A careful explanation of the trial's intents and purpose rooted in equipoise often helps allay these concerns and should always be employed at the time of consent.

Additionally, it may not always be practical to maintain blinding in surgical trials. For example, a comparison of open versus minimally invasive techniques may be hard to blind. However, minor variations in surgical techniques may be amenable to various blinding methods. Moreover, keeping the analytical teams remote from clinical interaction may help maintain blinding in certain situations. Irrespective, careful attention must be paid to enforce and maintain blinding whenever possible.

9.9 Pragmatic Trials and Comparative Effectiveness Research

Randomized controlled trials are conducted to assess if intervention has a biologic impact under strict controlled settings. These trials aim to demonstrate the "potential" of a treatment. Pragmatic controlled trials (PCTs), also called effectiveness trials, are conducted to measure the effectiveness of a treatment in a real-world setting. These trials aim to reach a maximum generalizability of the results while also making sure that the differences in outcomes are a result of the intervention and not due to chance or confounders. In other words, PCTs strive to maintain high external and internal validity [16].

A proposed distinction between explanatory trials (RCTs) and pragmatic trials is that RCTs confirm a physiological or clinical hypothesis while pragmatic trials guide a clinical or policy decision by providing evidence for adoption of the intervention into real-world clinical practice. The need to distinguish between the two became necessary when it was realized that many trials did not adequately inform practice because they were optimized to determine efficacy rather than effectiveness. Since RCTs are performed with relatively small sample sizes, at locations where experienced investigators are conducting these studies with a highly selected population of participants, they could potentially be overestimating benefits and underestimating harm.

The Pragmatic–Explanatory Continuum Indicator Summary (PRECIS) tool attempts to clarify the concept of pragmatism and provides a guide, scoring system, and graphical representation of the pragmatic features of a trial. Included variables are the recruitment of investigators and participants, the intervention and its delivery, follow-up, and the nature, determination, and analysis of outcomes. Most trials could be deemed pragmatic with regard to at least one of these dimensions, but very few end up being pragmatic in all areas [19].

In pragmatic trials, few restrictions are placed on the inclusion criteria to reflect the variation that would exist in the general patient population to ensure generalizability. A larger sample size is needed to cater to the heterogeneity of the population characteristics. These trials usually compare a new treatment to an already existing standard of care rather than a placebo and therefore are preferable as surgical trials where using a placebo or a sham intervention could be considered unethical. Pragmatic trials allow surgeons the flexibility (within set constraints of a real-world setting) to employ their own approaches to the various patients, while also implementing the intervention under trial to the randomly assigned patient. This flexibility in the pragmatic protocols enables academic surgeons to have their conventional practice while also doing research within a defined framework. Instead of measuring surrogate and objective outcomes, like in RCTs, the pragmatic trials focus on patient-centric outcomes like improvement in quality of life (QoL) and follow-up of patients for a longer duration of time. Surgical trials tend to have features of both explanatory trials and pragmatic trials and therefore exist along the continuum of the two designs. Trials designed to eventually aid the clinician to make the best possible decisions for their patients will prove to be most useful [16].

Comparative effectiveness research (CER) differs from clinical trials (especially pragmatic clinical trials) in that it is the conduct and generation of evidence, which incorporates results from observational and experimental researches, including RCTs, to compare the benefits and harms of different interventions to prevent, diagnose, treat, and monitor a clinical condition in the everyday settings and helps improve the delivery of care. A clinical trial is not CER in and of itself; however, CER uses results from clinical trials to inform clinical care. In general, CER aims to assist consumers, clinicians, purchasers, and policy makers to make informed decisions and improve healthcare outcomes for individual patients and patient populations [20].

9.10 Superiority/Inferiority Trials

The type of RCT being conducted depends on the aim of the trial itself. If the aim is demonstrating that the intervention (E) is superior to the control (C), then it is considered to be a superiority trial and the statistical tests executed are the superiority tests. If the results are significant, it could be concluded that intervention produces significantly better outcomes than the control. On the other hand, non-significant results are difficult to categorize as they obviously do not show superiority but also essentially do not show that the intervention was not as equally effective as the control. In fact, there will always exist a small difference in effects when two treatments are non-identical and yet the primary effect could very much be similar. The different interventions could also have the same primary effect but have secondary qualities that could make one more preferable over the other, and it was situations like this that have led to the inception of non-inferiority trials (NITs). Non-inferiority trials could be performed if it is ethically not appropriate to create a placebo group, which is a problem that is commonly encountered in surgical trials. NITs could also be a viable option if the primary outcomes between the groups are expected to be similar but secondary outcomes or safety profiles are anticipated to be better with the new intervention or the new treatment is cheaper and/or easier to administer and is more likely to be applicable in real-life situations [21].

9.11 Outcomes

All RCTs assess response variables or endpoints (outcomes) for which the groups are analyzed against each other. RCTs generally have a diverse set of variables being assessed and the investigator must define and specify the importance of each output variable during the design phase of the study. Each variable could be one of three types of response variables that are measured: dichotomous (measuring event rates), continuous (measuring mean values), and time to event (measuring hazard rates) [4, 7].

The primary endpoint is the predefined response variable that holds the greatest significance for all the involved parties (the patients, investigators, financers, and policy makers) and is mostly the treatment effect variable used when calculating the sample size. It is likely that a trial could be assessing multiple primary variables; however, this comes with its own issues of result interpretation and is discouraged. All other outcomes being evaluated are termed secondary outcomes, and these could be outcomes that were expected and observed and also those which were not expected but were still observed. Adverse effects should always be given importance, irrespective of their status as primary or secondary outcomes. The variable should be defined in way that a third party reading the study should be able to understand and use the same variables. Appropriate use of previously validated guidelines or scales is recommended to enhance the quality of the measurement and make future comparisons possible.

Any unplanned digression from the initially approved protocol must be reported. All changes to the selection criteria, intervention itself, data recording, analytical adjustments, and the reported outcomes ought to be clearly reported [4].

9.12 Types of Errors and Statistical Power

Statistical considerations must be made early during the planning phases of a clinical trial in order to truly understand the results and avoid pitfalls in sample size calculations, power, and statistical significance. Usually the null hypothesis, which states that no observed difference exists between two (or more) groups, is tested using appropriate statistical tests. Several types of errors can occur when interpreting these results which are discussed here.

9.12.1 Type I Error (α)

This is the probability of detecting a statistically significant difference when in fact no difference exists, that is, the chance of a false-positive result.

9.12.2 P-Value (p) and Significance Level (Alpha)

P-value is the probability of a type I error, that is, the probability of detecting a difference as large (or larger) as the actual difference observed given that the null hypothesis is true. The significance level (alpha) refers to the probability that is decided a priori, while *p*-value refers to calculated value obtained after performing a statistical test. Typically, the null hypothesis is rejected if the *p*-value is less than the chosen alpha. Alpha is often chosen arbitrarily but is conventionally set at 0.05 (1 in 20 chance of being incorrect) or 0.01 (1 in 100 chance of being incorrect). In general, the larger the alpha, the larger the required sample size.

9.12.3 Type II Error (β)

This is the probability of not detecting a statistically significant difference when in fact a difference truly exists, that is, the chance of a false-negative result.

9.12.4 Power (1–β)

Power is the probability of detecting a statistically significant difference when in fact a difference truly exists, or alternatively, the probability of rejecting a null hypothesis when it is false. In simpler terms, power quantifies the ability of a study to find true differences. Beta depends on alpha, the sample size, and the measure of true difference between variables (delta). In general, the higher the power, the larger the required sample size. Usually, alpha is set at 0.05 or 0.01 and beta is set at 0.90 or 0.95, while delta and sample size are variable. Delta is typically based on prior research findings and is set at the minimal level at which the differences between groups still remain clinically meaningful.

	When null hypothesis (H_0) is	
Statistical result	True	False
Reject null hypothesis (H_0)	Type I error (α)	Power (1-β), correct result
Fail to reject null hypothesis	Correct result	Type II error (β)

9.13 Sample Size Considerations

A clinical trial should have a sufficient sample size that would ensure the detection of a statistically significant, clinically meaningful effect of an intervention if in fact it truly exists. The end goal is to determine the most conservative sample size in order to avoid overestimates (failure to enroll, high costs) and underestimates (inconclusive results).

The exact details of accurate sample size calculation are beyond the scope of this discussion. However, a few broad concepts need to be understood in order to have a clear view of its mechanics. As alluded to before, in calculating sample sizes, beta and alpha are usually set by convention, while delta is estimated based on prior research. The larger the delta, the smaller the sample size needed to detect a true difference. These differences are typically tested using two-sided tests to detect differences in either direction (since a new treatment may perform better or worse than standard of care/placebo). The significance level used for sample size calculations for two-sided tests is twice that of a one-sided test; therefore, the choice of hypothesis testing has a bearing on sample size calculation. Another design consideration of sample size calculations for clinical trials is the allocation ratio for the probability of assignment to the treatment groups. Most researchers choose a 1:1 allocation ratio for their trials, which means the probability of assignment to the two groups is equal. Although a 1:1 allocation ratio usually maximizes trial power, ratios up to 2:1 minimally reduce the power [4, 22] and require a higher sample size. Lastly, many clinical trials routinely call for interim analyses to serve as an early warning system. If the treatment is overwhelmingly useful or harmful or an expected difference does not result, the trial may need to be stopped earlier due to safety concerns or to preserve resources. When performing these interim analyses, careful consideration of the sample size and initial significance level must be undertaken and adjusted for since the rate of incorrectly rejecting the null hypothesis will be larger.

9.14 Conclusion

Clinical trials, when planned and executed correctly, represent one of the best strategies to determine the clinical benefit or harm as a direct consequence of a new therapeutic intervention. Meticulous, upfront attention during the early design phases of a clinical trial will not only save precious resources, but also will result in the advancement of clinical therapeutics.

References

1. Winter K, Pugh SL, editors. An investigator's introduction to statistical considerations in clinical trials. Urologic oncology: seminars and original investigations. Amsterdam: Elsevier; 2019.
2. Akobeng AK. Understanding randomised controlled trials. Arch Dis Childhood. 2005;90(8):840–4.

3. Rohrig B, du Prel JB, Blettner M. Study design in medical research: part 2 of a series on the evaluation of scientific publications. Dtsch Arztebl Int. 2009;106(11):184–9.
4. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2012;10(1):28–55.
5. Van Belle G, Fisher LB. Biostatistics: a methodology for the health sciences. 2nd ed. Hoboken: Wiley; 2004. p. 15–9.
6. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.
7. Pawlik TM, Sosa JA. Success in academic surgery: clinical trials. 1st Edition. Springer. 2014. Pg 14, section 2.3; ISBN 978-1-4471-4679-7
8. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet. 2002;359(9305):515–9.
9. Peckham E, Brabyn S, Cook L, Devlin T, Dumville J, Torgerson DJ. The use of unequal randomisation in clinical trials—an update. Contemp Clin Trials. 2015;45:113–22.
10. Torgerson DJ, Campbell MK. Use of unequal randomisation to aid the economic efficiency of clinical trials. BMJ. 2000;321(7263):759.
11. Suresh K. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. J Hum Reprod Sci. 2011;4(1):8.
12. Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials: a review. Control Clin Trial. 2002;23(6):662–74.
13. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002;359(9306):614–8.
14. Spieth PM, Kubasch AS, Penzlin AI, Illigens BM-W, Barlinn K, Siepmann T. Randomized controlled trials—a matter of design. Neuropsy Dis Treat. 2016;12:1341.
15. Kendall JM. Designing a research project: randomised controlled trials and their principles. Emerg Med J. 2003;20(2):164–8.
16. Farrokhyar F, Karanicolas PJ, Thoma A, Simunovic M, Bhandari M, Devereaux P, et al. Randomized controlled trials of surgical interventions. Annal Surg. 2010;251(3):409–16.
17. Axelrod DA, Hayward R. Nonrandomized interventional study designs (quasi-experimental designs). Clinical research methods for surgeons. New York: Springer; 2006. p. 63–76.
18. Thoma A, Farrokhyar F, Bhandari M, Tandan V. Users' guide to the surgical literature. How to assess a randomized controlled trial in surgery. Can J Surg. 2004;47(3):200–8.
19. Ford I, Norrie J. Pragmatic trials. N Engl J Med. 2016;375(5):454–63.
20. Chang TI, Winkelmayer WC. Comparative effectiveness research: what is it and why do we need it in nephrology? Nephrol Dia Transplant. 2012;27(6):2156–61.
21. Lesaffre E. Superiority, equivalence, and non-inferiority trials. Bull NYU Hosp Jt Dis. 2008;66(2)
22. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet. 2005;365(9467):1348–53.