J Exp Stroke Transl Med (2010) 3(1): 19-27
Society for Experimental Stroke (www.s4es.org)

Fundamentals of clinical trial design

Scott R. Evans, Ph.D.

Department of Statistics, Harvard University, Boston, MA

Abstract

Most errors in clinical trials are a result of poor planning. Fancy statistical methods cannot rescue design flaws. Thus careful planning with clear foresight is crucial. Issues in trial conduct and analyses should be anticipated during trial design and thoughtfully addressed. Fundamental clinical trial design issues are discussed.

Keywords: p-value; confidence intervals; intent-to-treat; missing data; multiplicity; subgroup analyses; causation

1. Introduction

The objective of clinical trials is to establish the effect of an intervention. Treatment effects are efficiently isolated by controlling for bias and confounding and by minimizing variation. Key features of clinical trials that are used to meet this objective are randomization (possibly with stratification), adherence to intent-to-treat (ITT) principles, blinding, prospective evaluation, and use of a control group. Compared to other types of study designs (e.g., case-control studies, cohort studies, case reports), randomized trials have high validity but are more difficult and expensive to conduct.

2. Design Issues

There are many issues that must be considered when designing clinical trials. Fundamental issues including clearly defining the research question, minimizing variation, randomization and stratification, blinding, placebos/shams, selection of a control group, selection of the target population, the selection of endpoints, sample size, and planning for interim analyses will be discussed and common terms are defined (Table 1).

2.1 What is the question?

The design of every clinical trial starts with a primary clinical research question. Clarity and understanding of the research question can require much deliberation often entailing a transition from a vague concept (e.g., "to see if the drug works" or "to look at the neuro-biology of the drug") to a particular hypothesis that can be tested or a quantity that can be estimated using specific data collection instruments with a particular duration of therapy. Secondary research questions may also be of interest but the trial design usually is constructed to address the primary research question.

There are two strategies for framing the research question. The most common is hypothesis testing where researchers construct a null hypothesis (often "no effect" or "no difference") that is assumed to be true and evidence is sought to disprove it. An alternative hypothesis (the statement that is desired to be claimed) is also constructed (often the presence of an effect or difference between groups). Evidence is sought to support the alternative hypothesis. The second strategy is estimation. For example a trial might be designed to estimate the difference in response rates between two therapies with appropriate precision. Appropriate precision might be measured by the width of a confidence interval of the difference between the two response rates.

Clinical trials are classified into phases based on the objectives of the trial. Phase I trials are the first studies of an intervention conducted in humans. Phase I trials have small sample sizes (e.g., <20), may enroll healthy human participants, and are used to investigate pharmacokinetics, pharmacodynamics, and toxicity. Phase II trials are typically conducted to investigate a dose response relationship, identify an optimal dose, and to investigate safety issues. Phase III trials are generally large trials (i.e., many study participants) designed to "confirm" efficacy of an intervention. They are sometimes called "confirmatory trials" or "registration trials" in the context of pharmaceutical development. Phase IV trials are conducted after registration of an intervention. They are generally very large and are typically conducted by pharmaceutical companies for marketing purposes and to gain broader experience with the intervention.

Although clinical trials are conducted prospectively, one can think of them as being designed retrospectively. That is, there is a vision of the scientific claim (i.e., answer to the research question) that a project team would like to make at the end of the trial. In

---

* **Correspondence should be sent to:**

Scott R. Evans, Ph.D., Department of Statistics, Harvard University, 651 Huntington Avenue, FBX 513, Boston, MA, 02115. Phone: 614.432.2998; Fax: 617.432.3163; Email: evans@sdac.harvard.edu

order to make that claim, appropriate analyses must be conducted in order to justify the claim. In order to conduct the appropriate analyses, specific data must be collected in a manner suitable to conduct the analyses. In order to collect these necessary data, a thorough plan for data collection must be developed. This sequential retrospective strategy continues until a trial design has been constructed to address the research question.

Table 1. Terms in clinical trial design

Alternative Hypothesis	Claim that would like to be made at the end of the trial. The scientific method states that to prove something, assume the compliment is true and then look for contradictory evidence. If sufficient contradictory evidence is observed, then the desired claim has been proven. Typically the alternative hypothesis is something that the investigator desires to prove (e.g., that a new treatment is superior to placebo). The investigator thus assumes that the compliment (called the null hypothesis) is true and then looks for evidence to disprove the null hypothesis and hence claim the alternative hypothesis to be true.
Intent-to-treat (ITT)	Strategy for conducting a trial and analyzing data. The strategy implies "analyze as randomized" regardless of adherence or treatment received.
Null Hypothesis	Claim that an investigator desires to disprove. See "Alternative Hypothesis".
Phase I	The first studies conducted in humans using an experimental intervention. These trials often have small sample sizes (e.g., <20), may enroll healthy human participants, and are used to investigate pharmacokinetics, pharmacodynamics, and toxicity.
Phase II	Trials typically conducted to investigate a dose response relationship, identify an optimal dose, and to investigate safety issues.
Phase III	Generally large trials (i.e., many study participants) designed to "confirm" efficacy of an intervention. They are sometimes called "confirmatory trials" or "registration trials" in the context of pharmaceutical development.
Phase IV	Trials carried out after registration of an intervention. They are generally very large and are typically conducted by pharmaceutical companies for marketing purposes and to gain broader experience with the intervention.
Power	The probability of rejecting a null hypothesis when it should be rejected. In superiority trials (e.g., trials designed to show that a new treatment is superior to placebo) this means the probability of identifying a treatment effect when indeed a true treatment effect exists.
Type I Error	The probability if rejecting the null hypothesis when it should not be rejected (i.e., a false positive). In superiority trials this means the probability of (incorrectly) identifying a treatment effect when indeed a true treatment effect does not exist.
Type II Error	The probability of failing to reject the null hypothesis when it should be rejected (i.e., a false negative). Type II error is the compliment of "power". In superiority trials this means the probability of failing to identifying a treatment effect when indeed a true treatment effect exists.

Once the research question is well understood and associated hypotheses have been constructed then the project team must evaluate the characteristics of the disease, the therapies, the target population, and the measurement instruments. Each disease and therapy will have its own challenges. Neurologic data has many challenging characteristics. First, some neurologic outcomes can be subject to lots of variation (e.g., cognitive outcomes). Second, some neurologic outcomes are subjective in nature (e.g., pain, fatigue, anxiety, depression). Thirdly, some neurologic outcomes lack a gold standard definition or diagnosis (e.g., neuropathy, dementia). Forth, neurologic outcomes can be high dimensional (e.g., neuroimaging outcomes or genomic information, that cannot be captured using a single numeric score). Fifth,

composite outcomes are common (e.g., cognitive measures, instruments assessing depression or quality of life). Consider a trial to evaluate treatments for pain. Researchers should consider the subjective and transient nature of pain, the heterogeneity of pain expression, the placebo effect often encountered in pain trials, and the likely use of concomitant and rescue medications. Design must be customized to address these challenges. The goal of design is to construct the most efficient design within research constraints that will address the research question while considering these characteristics.

2.2 Minimizing variation

The larger the variation, the more difficult it is to identify treatment effects. Thus minimizing variation is a fundamental element of clinical trial design. Minimizing variation can be accomplished in several ways. One important method for reducing variation is to construct consistent and uniform endpoint definitions. Ideally endpoints could be measured objectively (e.g., via a laboratory test) however many endpoints are based on subjective evaluation. For example, the diagnosis of neuropathy or dementia may be an endpoint. However these diagnoses are partly subjective. Variation in these diagnoses can be minimized with clear definitions and consistent evaluations.

A common design feature is the use of central labs for quantitating laboratory parameters to eliminate between-lab variation or the use of central evaluators to eliminate between-evaluator variation. For example, the AIDS Clinical Trials Group (ACTG) uses a central laboratory to quantitate HIV-1 RNA viral load on all of its studies while trials using imaging modalities for diagnose stroke might consider using a central imaging laboratory to quantitate all images.

Variation can also be reduced with standardization of the manner in which study participants are treated and evaluated via training. For example, in studies that involve imaging, it is very important to have an imaging protocol that standardizes the manner in which images are collected to reduce added variation due to inconsistent patient positioning. Training modules can be developed to instruct site personnel on the appropriate administration of evaluations. For example extensive training on the administration of neuropsychological exams was conducted in the International Neurological HIV Study (ACTG A5199) and a training module was developed to instruct sites on the proper administration of the NeuroScreen that is employed in the Adult Longitudinally Linked Randomized Treatment (ALLRT) trials (ACTG A5001).

2.3 Randomization and stratification

Randomization is a powerful tool that helps control for bias in clinical trials. It essentially eliminates the bias associated with treatment selection. Although randomization cannot ensure between-treatment balance with respect to all participant characteristics, it does ensure the expectation of balance. Importantly randomization ensures this expectation of balance for all factors even if the factors are unknown or unmeasured. This expectation of balance that randomization provides combined with the ITT principle, provides the foundation for statistical inference.

Trials commonly employ *stratified* randomization to ensure that treatment groups are balanced with respect to confounding variables. In stratified randomization, separate randomization schedules are prepared for each stratum. For example, gender is a potential confounder for estimating the effects of interventions to treat or prevent stroke (e.g., a between-group imbalance with respect to gender could distort the estimate of the intervention effect). Thus trials investigating the effects of such interventions might employ stratified randomization based on gender. For example, two randomization schedules may be utilized; one for males and another for females. Stratified randomization ensures that the number of male participants in each treatment group is similar and that the number of female participants in each treatment group is similar. Stratification has a few limitations. First, stratification can only be utilized for known and measurable confounders. Secondly, although one can stratify on multiple variables, one has to be wary of over-stratification (i.e., too many strata for the given sample size). The sample size must be large enough to enroll several participants for each treatment from each stratum.

2.4 Blinding

Blinding is a fundamental tool in clinical trial design and a powerful method for preventing and reducing bias. Blinding refers to keeping study participants, investigators, or assessors unaware of the assigned intervention so that this knowledge will not affect their behavior, noting that a change in behavior can be subtle, unnoticeable, and unintentional. When study participants are blinded, they may be less likely to have biased psychological or physical responses to intervention, less likely to use adjunct intervention, less likely to drop out of the study, and more likely to adhere to the intervention. Blinding of study participants is particularly important for patient reported outcomes (e.g., pain) since knowledge of treatment assignment could affect their response. When trial investigators are blinded, they may be less likely to transfer inclinations to study participants, less likely to differentially apply adjunctive therapy, adjust a dose, withdraw study participants, or encourage participants to continue participation. When assessors are blinded, they may be less likely to have biases affect their outcome assessments. In a placebo controlled

trial for an intervention for multiple sclerosis, an evaluation was performed by both blinded and unblinded neurologists. A benefit of the intervention was suggested when using the assessments from neurologists that were not blinded, but not when using the assessments from the blinded neurologists. In this case, the blinded assessment is thought to be more objective.

Clinical trialists often use the terms “single-blind” to indicate blinding of study participants, “double-blind” to indicate blinding of study participants and investigators, and “triple-blind” to indicate blinding of participants, investigators, and the sponsor and assessors. Trials without blinding are often referred to as “open-label”.

Successful blinding is not trivial. In a placebo-controlled trial, a placebo must be created to look, smell, and taste just like the intervention. For example a concern for a trial evaluating the effects of minocycline on cognitive function may be that minocycline can cause a change in skin pigmentation, thus unblinding the intervention. Blinding can be challenging or impractical in many trials. For example surgical trials often cannot be double-blind for ethical reasons. The effects of the intervention may also be a threat to the blind. For example, an injection site reaction of swelling or itching may indicate an active intervention rather than a sham injection. Researchers could then consider using a sham injection that induces a similar reaction.

In late phase clinical trials, it is common to compare two active interventions. These interventions may have different treatment schedules (e.g., dosing frequencies), may be administered via different routes (e.g., oral vs. intravenously), or may look, taste, or smell different. A typical way to blind such a study is the “double-dummy” approach that utilizes two placebos, one for each intervention. This is often easier than trying to make the two interventions look like each other. Participants are then randomized to receive one active treatment and one placebo (but are blinded). The downside of this approach is that the treatment schedules become more complicated (i.e., each participant must adhere to two regimens).

When blinding is implemented in a clinical trial, a plan for assessing the effectiveness of the blinding may be arranged. This usually requires two blinding questionnaires, one completed by the trial participant and the other completed by the local investigator or person that conducts the evaluation of the trial participant. Having “double-blind” in the title of a trial does not imply that blinding was successful. Reviews of blinded trials suggest that many trials experience issues that jeopardize the blind. For example in a study assessing zinc for the treatment of the common cold(Prasad *et al* 2000) the blinding failed because the taste and aftertaste of zinc was distinctive. Creative designs can be utilized to help maintain the blind. For example, OHARA and the ACTG are developing a study to evaluate the use of gentian violet (GV) for the treatment of oral candidiasis. GV has staining potential which could jeopardize the blind when the assessors conduct oral examinations after treatment. A staining cough drop could be given to study participants prior to evaluation to help maintain the blind.

Unplanned unblinding should only be undertaken to protect participant safety (i.e., if the treatment assignment is critical for making immediate therapeutic decisions).

Blinding has been poorly reported in the literature. Researchers should explicitly state whether a study was blinded, who was blinded, how blinding was achieved, the reasons for any unplanned unblinding, and state the results of an evaluation of the success of the blinding.

2.5 Placebos/Shams

A placebo can be defined as an inert pill, injection, or other sham intervention that masks as an active intervention in an effort to maintain blinding of treatment assignment. It is termed the “sugar pill” and does not contain an active ingredient for treating the underlying disease or syndrome but is used in clinical trials as a control to account for the natural history of disease and for psychological effects. One disadvantage to the use of placebos is that sometimes they can be costly to obtain.

Although the placebo pill or injection has no activity for the disease being treated, it can provide impressive treatment effects. This is especially true when the endpoint is subjective (e.g., pain, depression, anxiety, or other patient reported outcomes). Evans et.al. (Evans *et al* 2007) reported a significant improvement in pain in the placebo arm of a trial investigating an intervention for the treatment of painful HIV-associated peripheral neuropathy.

There can be many logistic and ethical concerns in clinical trials where neither a placebo, nor a sham control can be applied. The inability to use placebos is common in the development of devices. For example, surgical trials rarely have a sham/placebo.

2.6 Selection of a control group

The selection of a control group is a critical decision in clinical trial design. The control group provides data about what would have happened to participants if they were not treated or had received a different intervention. Without a control group, researchers would be unable to discriminate the effects caused by the investigational intervention from effects due to the

natural history of the disease, patient or clinician expectations, or the effects of other interventions.

There are three primary types of control groups: 1) historical controls, 2) placebo/sham controls and 3) active controls. The selection of a control group depends on the research question of interest. If it is desirable to show any effect, then placebo-controls are the most credible and should be considered as a first option. However placebo controls may not be ethical in some cases and thus active controls may be utilized. If it is desirable to show noninferiority or superiority to other active interventions then active controls may be utilized.

Historical controls are obtained from studies that have already been conducted and are often published in the medical literature. The data for such controls is external to the trial being designed and will be compared with data collected in the trial being designed. The advantage of using historical controls is that the current trial will require fewer participants and thus use of historical controls provides an attractive option from a cost and efficiency perspective. The drawback of trials that utilize historical controls is that they are non-randomized studies (i.e., the comparison of newly enrolled trial participants to the historical controls is a non-randomized comparison) and thus subject to considerable bias, requiring additional assumptions when making group comparison (although note that the historical controls themselves may have been drawn from randomized trials). Historical controls are rarely used in clinical trials for drug development due to the concerns for bias. However, when historical data are very reliable, well documented and other disease and treatment conditions have not changed since the historical trial was conducted, then they can be considered. Historical controls have become common in device trials when placebo-controls are not a viable option. Historical controls can be helpful in interpreting the results from trials for which placebo controls are not ethical (e.g., oncology trials).

An active control is an active intervention that has often shown effectiveness to treat the disease under study. Often an active control is selected because it is the standard of care (SOC) treatment for the disease under study. Active controls are selected for use in noninferiority trials. Active controls and placebo controls can be used simultaneously and provide useful data. For example, if the new intervention was unable to show superiority to placebo, but an active control group was able to demonstrate superiority to placebo, then this may be evidence that the new intervention is not effective. However, if the active control with established efficacy did not demonstrate superiority to placebo, then it is possible the trial was flawed or may have been underpowered because of the placebo response or variability being unexpected high.

2.7 Selection of a population and entry criteria

In selecting a population to enroll into a trial, researchers must consider the target use of the intervention since it will be desirable to generalize the results of the trial to the target population. However researchers also select entry criteria to help ensure a high quality trial and to address the specific objectives of the trial.

The selection of a population can depend on the trial phase since different phases have different objectives. Early phase trials tend to select populations that are more homogenous since it is easier to reduce response variation and thus isolate effects. Later phase trials tend to target more heterogeneous populations since it is desirable to have the results of such trials to be generalizable to the population in which the intervention will be utilized in practice. It is often desirable for this targeted patient population to be as large as possible to maximize the impact of the intervention. Thus phase III trials tend to have more relaxed entry criteria that are representative (both in demographics and underlying disease status) to the patient population for which the intervention is targeted to treat.

When constructing entry criteria, the safety of the study participant is paramount. Researcher should consider the appropriateness of recruiting participants with various conditions into the trial. The ability to accrue study participants can also affect the selection of entry criteria. Although strict entry criteria may be scientifically desirable in some cases, studies with strict entry criteria may be difficult to accrue particularly when the disease is rare or alternative interventions or trials are available. Entry criteria may need to be relaxed so that enrollment can be completed within a reasonable time frame.

Researchers should also consider restricting entry criteria to reduce variation and potential for bias. Participants that enroll with confounding indications that could influence treatment outcome could be excluded to reduce potential bias. For example, in a trial evaluating interventions for HIV-associated painful neuropathy, conditions that may confound an evaluation of neuropathy such as diabetes or a B12 deficiency may be considered exclusionary.

2.8 Selection of endpoints

The selection of endpoints in a clinical trial is extremely important and requires a marriage of clinical relevance with statistical reasoning. The motivation for every clinical trial begins with a scientific question. The primary objective of the trial is to address the scientific question by collecting appropriate data. The

selection of the primary endpoint is made to address the primary objective of the trial. The primary endpoint should be clinically relevant, interpretable, sensitive to the effects of intervention, practical and affordable to measure, and ideally can be measured in an unbiased manner.

Endpoints can generally be categorized by their scale of measurement. The three most common types of endpoints in clinical trials are continuous endpoints (e.g., pain on a visual analogue scale), categorical (including binary, e.g., response vs. no response) endpoints, and event-time endpoints (e.g., time to death). The scale of the primary endpoint impacts the analyses, trial power, and thus costs.

In many situations, more than one efficacy endpoints are used to address the primary objective. This creates a multiplicity issue since multiple tests will be conducted. Decisions regarding how the statistical error rates (e.g., Type I error) will be controlled should be described in the protocol and in the statistical analysis plan.

Endpoints can be classified as being objective or subjective. Objective endpoints are those that can be measured without prejudice or favor. Death is an objective endpoint in trials of stroke. Subjective endpoints are more susceptible to individual interpretation. For example, neuropathy trials employ pain as a subjective endpoint. Other examples of subjective endpoints include depression, anxiety, or sleep quality. Objective endpoints are generally preferred to subjective endpoints since they are less subject to bias.

1). Composite endpoints

An intervention can have effects on several important endpoints. Composite endpoints combine a number of endpoints into a single measure. The CHARISMA (Bhatt *et al* 2006), MATCH (Diener *et al* 2004), and CAPRIE (Committee 1996) studies of clopidogrel for the prevention of vascular ischemic events use combinations of MI, stroke, death, and re-hospitalization as components of composite endpoints. The advantages of composite endpoints are that they may result in a more completed characterization of intervention effects as there may be interest in a variety of outcomes. Composite endpoints may also result in higher power and resulting smaller sample sizes in event-driven trials since more events will be observed (assuming that the effect size is unchanged). Composite endpoints may also reduce the bias due to competing risks and informative censoring. This is because one event can censor other events and if data were only analyzed on a single component then informative censoring can occur. Composite endpoints may also help avoid the multiplicity issue of evaluating many endpoints individually.

Composite endpoints have several limitations. Firstly, significance of the composite does not necessarily imply significance of the components nor does significance of the components necessarily imply significance of the composite. For example one intervention could be better on one component but worse on another and thus result in a non-significant composite. Another concern with composite endpoints is that the interpretation can be challenging particularly when the relative importance of the components differs and the intervention effects on the components also differ. For example, how do we interpret a study in which the overall event rate in one arm is lower but the types of events occurring in that arm are more serious? Higher event rates and larger effects for less important components could lead to a misinterpretation of intervention impact. It is also possible that intervention effects for different components can go in different directions. Power can be reduced if there is little effect on some of the components (i.e., the intervention effect is diluted with the inclusion of these components).

When designing trials with composite endpoints, it is advisable to consider including events that are more severe (e.g., death) than the events of interest as part of the definition of the composite to avoid the bias induced by informative censoring. It is also advisable to collect data and evaluate each of the components as secondary analyses. This means that study participants should continue to be followed for other components after experiencing a component event. When utilizing a composite endpoint, there are several considerations including: (i) whether the components are of similar importance, (ii) whether the components occur with similar frequency, and (iii) whether the treatment effect is similar across the components.

2). Surrogate Endpoints.

In the treatment of some diseases, it may take a very long time to observe the definitive endpoint (e.g., death). A surrogate endpoint is a measure that is predictive of the clinical event but takes a shorter time to observe. The definitive endpoint often measures clinical benefit whereas the surrogate endpoint tracks the progress or extent of disease. Surrogate endpoints could also be used when the clinical endpoint is too expensive or difficult to measure, or not ethical to measure.

An example of a surrogate endpoint is blood pressure for hemorrhagic stroke.

Surrogate markers must be validated. Ideally evaluation of the surrogate endpoint would result in the same conclusions if the definitive endpoint had been used. The criteria for a surrogate marker are: (1) the marker is predictive of the clinical event, and (2) the

intervention effect on the clinical outcome manifests itself entirely through its effect on the marker. It is important to note that significant correlation does not necessarily imply that a marker will be an acceptable surrogate.

2.9 Preventing missing data and encouraging adherence to protocol

Missing data is one of the biggest threats to the integrity of a clinical trial. Missing data can create biased estimates of treatment effects. Thus it is important when designing a trial to consider methods that can prevent missing data. Researchers can prevent missing data by designing simple clinical trials (e.g., designing protocols that are easy to adhere to; having easy instructions; having patient visits and evaluations that are not too burdensome; having short, clear case report forms that are easy to complete, etc.) and adhering to the ITT principle (i.e., following all patients after randomization for the scheduled duration of follow-up regardless of treatment status, etc.).

Similarly it is important to consider adherence to protocol (e.g., treatment adherence) in order address the biological aspect of treatment comparisons. Envision a trial comparing two treatments in which the trial participants in both groups do not adhere to the assigned intervention. Then when evaluating the trial endpoints, the two interventions will appear to have similar effects regardless of any differences in the biological effects of the two interventions. Note however that the fact that trial participants in neither intervention arm adhere to therapy may indicate that the two interventions do not differ with respect to the strategy of applying the intervention (i.e., making a decision to treat a patient). Researchers need to be careful about influencing participant adherence since the goal of the trial may be to evaluate the strategy of how the interventions will work in practice (which may not include incentives to motivate patients similar to that used in the trial).

2.10 Sample size

Sample size is an important element of trial design because too large of a sample size is wasteful of resources but too small of a sample size could result in inconclusive results. Calculation of the sample size requires a clearly defined objective. The analyses to address the objective must then be envisioned via a hypothesis to be tested or a quantity to be estimated. The sample size is then based on the planned analyses. A typical conceptual strategy based on hypothesis testing is as follows:

1. Formulate null and alternative hypotheses. For example, the null hypothesis might be that the response rate in the intervention and placebo arms of a trial are the same and the alternative hypothesis is that the response rate in the intervention arm is greater than the placebo arm by a certain amount (typically selected as the “minimum clinically relevant difference”).
2. Select the Type I error rate. Type I error is the probability of incorrectly rejecting the null hypothesis when the null hypothesis is true. In the example above, a Type I error often implies that you incorrectly conclude that an intervention is effective (since the alternative hypothesis is that the response rate in the intervention is greater than in the placebo arm). In regulatory settings for Phase III trials, the Type I error is set at 5%. In other instances the investigator can evaluate the “cost” of a Type I error and decide upon an acceptable level of Type I error given other design constraints. For example, when evaluating a new intervention, an investigator may consider using a smaller Type I error (e.g., 1%) when a safe and effective intervention already exists or when the new intervention appears to be “risky”. Alternatively a larger Type I error (e.g., 10%) might be considered when a safe and effective intervention does not exist and when the new intervention appears to have low risk.
3. Select the Type II error rate. Type II error is the probability of incorrectly failing to reject the null hypothesis when the null hypothesis should be rejected. The implication of a Type II error in the example above is that an effective intervention is not identified as effective. The compliment of Type II error is “power”, i.e., the probability of rejecting the null hypothesis when it should be rejected. Type II error and power are not generally regulated and thus investigators can evaluate the Type II error that is acceptable. For example, when evaluating a new intervention for a serious disease that has no effective treatment, the investigator may opt for a lower Type II error (e.g., 10%) and thus higher power (90%), but may allow Type II error to be higher (e.g., 20%) when effective alternative interventions are available. Typically Type II error is set at 10-20%.
4. Obtain estimates of quantities that may be needed (e.g., estimates of variation or a control group response rate). This may require searching the literature for prior data or running pilot studies.
5. Select the minimum sample size such that two conditions hold: (1) if the hull hypothesis

is true then the probability of incorrectly rejecting is no more than the selected Type I error rate, and (2) if the alternative hypothesis is true then the probability of incorrectly failing to reject is no more than the selected Type II error (or equivalently that the probability of correctly rejecting the null hypothesis is the selected power).

The selection of quantities such as the "minimum clinically relevant difference", Type I error, and Type II error, reflects the assumptions, limitations, and compromises of the study design, and thus require diligent consideration. Since assumptions are made when sizing the trial (e.g., via an estimate of variation), evaluation of the sensitivity of the required sample size to variation in these assumptions is prudent as the assumptions may turn-out to be incorrect. Interim analyses can be used to evaluate the accuracy of these assumptions and potentially make sample size adjustments should the assumptions not hold. Sample size calculations may also need to be adjusted for the possibility of a lack of adherence or participant drop-out. In general, the following increases the required sample size: lower Type I error, lower Type II error, larger variation, and the desire to detect a smaller effect size or have greater precision.

An alternative method for calculating the sample size is to identify a primary quantity to be estimated and then estimate it with acceptable precision. For example, the quantity to be estimated may be the between-group difference in the mean response. A sample size is then calculated to ensure that there is a high probability that this quantity is estimated with acceptable precision as measured by say the width of the confidence interval for the between-group difference in means.

2.11 Planning for interim analyses

Interim analysis should be considered during trial design since it can affect the sample size and planning of the trial. When trials are very large or long in duration, when the interventions have associated serious safety concerns, or when the disease being studied is very serious, then interim data monitoring should be considered. Typically a group of independent experts (i.e., people not associated with the trial but with relevant expertise in the disease or treatments being studied, e.g., clinicians and statisticians) are recruited to form a Data Safety Monitoring Board (DSMB). The DSMB meets regularly to review data from the trial to ensure participant safety and efficacy, that trial objectives can be met, to assess trial design assumptions, and assess the overall risk-benefit of the intervention. The project team typically remains blinded to these data if applicable. The DSMB then makes recommendations to the trial sponsor regarding whether the trial should continue as planned or whether modifications to the trial design are needed.

Careful planning of interim analyses is prudent in trial design. Care must be taken to avoid inflation of statistical error rates associated with multiple testing to avoid other biases that can arise by examining data prior to trial completion, and to maintain the trial blind.

3. Common Structural Designs

Many structural designs can be considered when planning a clinical trial. Common clinical trial designs include single-arm trials, placebo-controlled trials, crossover trials, factorial trials, noninferiority trials, and designs for validating a diagnostic device. The choice of the structural design depends on the specific research questions of interest, characteristics of the disease and therapy, the endpoints, the availability of a control group, and on the availability of funding. Structural designs are discussed in an accompanying article in this special issue.

4. Summary

This manuscript summarizes and discusses fundamental issues in clinical trial design. A clear understanding of the research question is a most important first step in designing a clinical trial. Minimizing variation in trial design will help to elucidate treatment effects. Randomization helps to eliminate bias associated with treatment selection. Stratified randomization can be used to help ensure that treatment groups are balanced with respect to potentially confounding variables. Blinding participants and trial investigators helps to prevent and reduce bias. Placebos are utilized so that blinding can be accomplished. Control groups help to discriminate between intervention effects and natural history. There are three primary types of control groups: (1) historical controls, (2) placebo/sham controls, and (3) active controls. The selection of a control group depends on the research question, ethical constraints, the feasibility of blinding, the availability of quality data, and the ability to recruit participants. The selection of entry criteria is guided by the desire to generalize the results, concerns for participant safety, and minimizing bias associated with confounding conditions. Endpoints are selected to address the objectives of the trial and should be clinically relevant, interpretable, sensitive to the effects of an intervention, practical and affordable to obtain, and measured in an unbiased manner. Composite endpoints combine a number of component endpoints into a single measure. Surrogate endpoints are measures that are predictive of a clinical event but take a shorter time to observe than the clinical endpoint of interest. Interim analyses should be considered for larger trials of long duration or trials of serious disease or trials that evaluate potentially harmful interventions. Sample size should be consi-

dered carefully so as not to be wasteful of resources and to ensure that a trial reaches conclusive results.

There are many issues to consider during the design of a clinical trial. Researchers should understand these issues when designing clinical trials.

Acknowledgement

The author would like to thank Dr. Justin McArthur and Dr. John Griffin for their invitation to participate as part of the ANAs Summer Course for Clinical and Translational Research in the Neurosciences. The author thanks the students and faculty in the course for their helpful feedback. This work was supported in part by Neurologic AIDS Research Consortium (NS32228) and the Statistical and Data Management Center for the AIDS Clinical Trials Group (U01 068634).

References

Bhatt DL, Fox KAA, Hacke W, Berger PB, Black HR, Boden WE, Cacoub P, Cohen EA, Creager MA, Easton JD, Flather MD, Haffner SM, Hamm CW, Hankey GJ, Johnston SC, Mak K-H, Mas J-L, Montalescot G, Pearson TA, Steg PG, Steinhubl SR, Weber MA, Brennan DM, Fabry-Ribaudo L, Booth J, Topol EJ (2006) Clopidogrel and Aspirin Versus Aspirin Alone for the Prevention of Atherothrombotic Events. *N Engl J Med* 16:1706-17.

CAPRIE Committee (1996) A Randomised, Blinded, Trial of Clopidogrel Versus Aspirin in Patients at Risk of Ischaemic Events (CAPRIE) . *Lancet* 348:1329–39.

Diener H, Bogousslavsky J, Brass L, Cimminiello C, Csiba L, Kaste M, Leys D, Matias-Guiu J, Rupprecht H, investigators obotM (2004) Aspirin and Clopidogrel Compared with Clopidogrel Alone after Recent Ischaemic Stroke or Rransient Ischaemic Attack in High-Risk Patients (MATCH): Randomized, Double-Blind, Placebo-Controlled Trial. *Lancet* 364.

Evans S, Simpson D, Kitch D, King A, Clifford D, Cohen B, MacArthur J (2007) A Randomized Trial Evaluating Prosaptide™ for HIV-Associated Sensory Neuropathies: Use of an Electronic Diary to Record Neuropathic Pain. *PLoS ONE* 2:e551.doi:10.1371/journal.pone.0000551.

Prasad A, Fitzgerald J, Bao B, Beck F, Chandrasekar P (2000) Duration of Symptoms and Plasma Cytokine Levels in Patients with the Common Cold Treated with Zinc Acetate. *Ann Intern Med* 133:245-52.