A Primer For Evaluating Clinical Trials

Gary H. Lyman, MD, MPH, and Nicole M. Kuderer

An understanding of the limitations of design and conduct of clinical trials helps clinicians to assess the results

Background: Evidence-based medicine demands the use of information from clinical trials to direct medical care. Knowledge of the principles of trial design and conduct is important to assess the validity of results.

Methods: The authors review the key principles behind clinical study design and conduct, and they summarize important biases and confounding issues.

Results: Clear hypotheses, a well-described study population, precise measurements, freedom from bias, and consideration of any interactions are attributes of good clinical trials.

Conclusions: The greatest level of evidence in support of a difference in outcome is associated with randomized, controlled clinical trials, particularly when combined with other randomized trials in a systematic fashion (meta-analysis).

Introduction

Clinicians are faced daily with the challenge of understanding and evaluating the results of clinical investigations published in the medical literature. Clinical researchers also face the task of translating their ideas into reasonable hypotheses and designing appropriate studies to test these hypotheses. Editors and reviewers experience similar challenges in evaluating clinical reports and selecting those of sufficient interest and quality to justify publication. It is essential that clinician readers of the medical literature and investigators conducting clinical research, as well as reviewers and editors of medical journals, become familiar with the fundamentals of clinical research methods.

This primer for designing, analyzing, and evaluating the results of clinical investigations may serve as an introduction to these issues for students and young clinicians as well as a refresher or aid to readers and established investigators of the medical literature. This primer is not a comprehensive treatise on this vast subject; however, it provides a framework for designing studies, analyzing data, and reading the medical literature for quick and ready reference.

Table 1 presents five basic questions to be addressed when designing and analyzing a clinical investigation or when reviewing or reading a published report of a clinical trial. Researchers, reviewers, and readers of the medical literature should be able to address and successfully answer each of these fundamental questions in the evaluation of any clinical trial. The optimal evaluation of a clinical investigation requires both an understanding of the clinical issues involved and a basic awareness of statistical methods.

**Question 1: Are the Study Hypotheses Clearly Stated and Relevant?**

Developing a clear statement of the clinical question or hypothesis to be studied presents a difficult but critical challenge to the inexperienced investigator or reader of the medical literature. In many cases, the researcher has a good understanding of the issues involved but has difficulty in explicitly stating or translating these into testable hypotheses. A clear statement of the hypothesis of interest requires a definition of both the dependent (outcome) variables and the independent (treatment, prognostic factors) variables of interest. The primary hypothesis generally addresses the effect of an intervention variable, such as treatment, on the outcome variables of interest, such as response or survival (Fig 1). Secondary hypotheses generally address the effect of the intervention on outcome among specified subgroups (subgroup analysis). In randomized clinical trials, the primary hypothesis is usually evident, while in nonrandomized trials and especially in uncontrolled trials, it may be much less apparent. Both the primary and any secondary hypotheses should be stated in advance of the study and should include any planned subgroup analyses.
Testable hypotheses must be formulated in terms of measurable outcomes. The measures of interest must be clearly stated, and the method of measurement must be described. Clinical measurement scales include continuous, categorical, and time-to-event observations. Time-to-event measures should specify the beginning and ending points and should clearly define the criteria for an event. Methods utilized to assure the completeness and quality of the measured endpoints should be discussed. The measurement scale used will determine the appropriate summary measures and applicable statistical methods. Finally, the importance and clinical relevance of the study question should be clearly stated. Even a well-designed and carefully analyzed trial cannot justify a study of limited clinical importance.

**Question 2: Is the Study Population Adequately Described?**

No investigation can be properly evaluated without a detailed understanding of the study population. Investigators often fail to adequately define and account for the subjects evaluated in a study. The investigator must find the proper balance between restricting eligibility in order to obtain a relatively uniform group of subjects and minimizing eligibility restrictions in order to provide greater ability to generalize the study results. Every study evaluation should specify the eligibility criteria for entry, including any reasons for exclusion. Obviously, specifying such criteria may not fully account for differences in referral patterns to a particular institution or in selection bias imposed by clinicians or patients. While this is of great concern in uncontrolled or historically controlled studies, it is also important to the interpretation and extrapolation of the results of randomized, controlled trials. Fig 2 provides a flow chart illustrating subject evaluation, entry, and loss during the process of a clinical trial. The number of subjects, reasons for ineligibility or study refusal, failure to randomize, alteration in treatment, or withdrawal from study should be presented, as well as a description of the specific population on which subsequent analysis is based. An inequality rate of 10% or less for major outcome measures should be achievable in most studies. In addition, incomplete data can seriously restrict the sample population considered in subgroup and multivariate analyses and can potentially bias the results.

![Flow chart](image)

**Question 3: Are the Observed Differences Due to Random Error (Chance)?**

The outcomes of a clinical study are seldom exactly equal among the study groups. The differences observed may represent true outcome differences. However, it is imperative that the investigator and the reader consider the possibility that the differences are due to either random error or systematic error. Random error is the result of either biologic or measurement variation, whereas systematic error is the result of a variety of biases that can affect the results of a trial. The process of evaluating the outcomes of a study for random error includes both estimation and statistical testing. Estimates summarizing the distribution of measured variables may include point estimates (such as means or proportions) and measures of precision (such as confidence intervals). Confidence limits represent the upper and lower bounds likely to contain the true value of the variable. In an uncontrolled study, the number of subjects should be sufficient to achieve the desired level of precision in estimating an outcome (such as the response rate).

Statistical testing involves an assessment of the probability of obtaining an observed difference in outcome when there is actually no true difference between the groups (false-positive rate; \( P \) value). When the \( P \) value is greater than a specified critical value (eg, .05), the observed difference is considered to be not statistically significant. Random error is the result of either biologic or measurement variation, whereas systematic error is the result of a variety of biases that can affect the results of a trial. The process of evaluating the outcomes of a study for random error includes both estimation and statistical testing. Estimates summarizing the distribution of measured variables may include point estimates (such as means or proportions) and measures of precision (such as confidence intervals). Confidence limits represent the upper and lower bounds likely to contain the true value of the variable. In an uncontrolled study, the number of subjects should be sufficient to achieve the desired level of precision in estimating an outcome (such as the response rate).

Repeated or interim analyses will increase the chances of observing a statistically significant difference purely due to chance. The number of interim analyses planned and any measures taken to appropriately adjust the final analysis should be discussed. Similarly, the risk of a false-positive result will increase with multiple subgroup comparisons. Subgroup analysis should be considered in the design of the study and should be limited to those with sufficient biologic or clinical rationale.

Random error should be considered and controlled for in the design, conduct, and analysis of a clinical investigation. Evidence supporting a true effect on outcome is summarized in Table 2. It is imperative that sufficient numbers of subjects are included to ensure that a nonsignificant difference will not be due to a false-negative result. The probability of obtaining a significant result when a real difference exists is termed the *power*. A sample size large enough to achieve a power of 80% (and
The most effective approach to controlling bias is in the design of a clinical investigation. As illustrated in Fig 3, the level of evidence in support of a true difference in outcome measures will be greater for controlled clinical trials, especially when treatment is randomly assigned and when there is blinding of both the observer and the subject to the treatment administered (double-blinding). In the process of randomization, both known and unknown confounding factors will be evenly distributed among the treatment groups on the average. The balance of important factors within treatment groups can be ensured by randomizing separately within subgroups (stratification). Support for an outcome difference may be provided by nonrandomized studies with one or more concurrent control groups that are preferably matched on important prognostic factors. However, historically controlled, uncontrolled, and purely descriptive studies provide relatively limited evidence in support of an observed outcome difference. Perhaps the greatest level of evidence comes from a systematic and quantitative overview of properly conducted clinical trials in the form of a meta-analysis.

Bias can also be introduced in the conduct and measurement phase of a study. The quality of observation, measurement, and data recording represent additional information that affects the validity of study results. Great care must be taken to obtain complete information and to account for any missing data.

**Question 4: Are the Observed Differences Due to Systematic Error (Bias)?**

Apparent differences in measured outcomes may appear to be clinically and statistically significant and yet may be the result of systematic error or bias within the study. Even the most careful measurement and elegant statistical analysis will not salvage a biased clinical trial. The most common types of bias in clinical investigations are those related to subject selection, outcome measurement, and confounding. Confounding represents the modification of the true relationship between treatment and outcome by another factor, eg, prognostic factor (Fig 1). This occurs when the factor is associated with both the outcome of interest and treatment group assignment. Confounding can obscure a true outcome difference when it exists or can create an apparent difference that does not exist.

Despite the randomization process and other safeguards in the design and conduct of a study, the distribution of known prognostic factors within treatment groups should be compared in the analysis. In randomized trials, outcomes should be compared among the groups based on the original treatment assignment rather than based on the treatment received (intention-to-treat analysis). The investigator should present the distribution of known prognostic factors among the treatment groups. Any prognostic factor associated with both outcome and treatment group assignment must be considered as a potential confounding factor and should be properly adjusted in the analysis. If actual confounding has occurred, the relationship between treatment and outcome will be either strengthened or weakened in subsequent analysis.

Adjustment for confounding can be achieved through either stratified or multivariate analysis. In stratified analysis, the relationship between treatment and outcome is evaluated separately among the subgroups of a prognostic factor. The major strengths of such analysis are the ease and clarity of presentation, while the major weakness is that of small numbers in the subgroups considered. The Mantel-Haenszel method permits the combining of subgroups, thus providing an adjusted overall test.

Multivariate analysis permits evaluation and adjustment for confounding for any variable contained in the model. The coefficients for variables within the model represent the rate of change in outcome with change in the variable. The effect of each variable on outcome is adjusted for the effects of the other variables retained in the model. While multivariate methods avoid the problem of small subgroups, subjects missing one or more clinical measures will be deleted from the analysis, which may bias the results. The investigator should present the actual number of subjects evaluated in each model. Multivariate analysis is a sophisticated process requiring knowledge of the assumptions on which the model is based. In addition, the results are often difficult to display and explain to an inexperienced reader. Finally, the reader must recognize that the observed outcome differences may still be the result of an imbalance of unmeasured or unrecognized confounding factors that cannot be addressed in any analysis.

**Question 5: Are the Observed Differences Modified by Other Factors?**
When confounding is present, the relationship between treatment and outcome is altered by the confounding factor. Interaction occurs when the relationship between treatment and outcome is different within subgroups of the factor. Synergy occurs when the relationship between treatment and outcome is greater than expected in a subgroup, while a decrease in the expected relationship is termed antagonism. When interaction is present, the relationship between treatment and outcome should be presented separately for each subgroup. Combining results across subgroups of an interaction term will produce an inaccurate measure that does not apply to any subgroup.

Inclusion of a variable in a multivariate model will adjust for confounding but not for interaction. If interaction is present, the investigator must either present separate models for each subgroup or include an interaction term (product term) in the model along with each variable. Such models are often difficult to present and interpret; therefore, the presence of interaction is often not properly considered or addressed in the presentation of clinical trial results. The optimal consideration of interaction is aided by an understanding of the underlying biological mechanisms and relationships of the clinical measures involved.

Conclusions

The appropriate design, presentation, and evaluation of a clinical investigation require an explicit definition of the study population, a clear statement of the primary and secondary hypotheses in terms of measurable outcomes, and careful consideration of any observed differences for precision, bias, and possible interaction. The presence of interaction is often not properly considered or addressed in the presentation of clinical trial results. The optimal consideration of interaction is aided by an understanding of the underlying biological mechanisms and relationships of the clinical measures involved.

Table 3: Structure of an Article

<table>
<thead>
<tr>
<th>Abstract</th>
<th>Present structured overview of article including introduction, methods, results, and conclusions</th>
</tr>
</thead>
</table>
| Introduction | State basis and clinical background
Define primary and secondary outcomes
State primary and secondary hypotheses |
| Methods | Define patient population with inclusion/exclusion criteria
Describe study design (randomization method, blinding, controls, etc.)
Define primary and secondary outcome measures
Describe variable data based on clinically important differences
Describe statistical analyses including:
Summary of sample size based on any type of measure
Correlation coefficients and power
Subgroup analyses planned prior to study initiation |
| Results | Describe flow of patients considered, registered, randomized, treated, and available for outcome in each group
Summarize descriptive statistics including mean estimates and confidence limits
Display distribution of prognostic factors in each treatment group
Estimate difference in outcome for treatment effects for primary and secondary outcomes based on intention-to-treat
Describe evaluation of prognostic factors including adjustment for confounding and interaction |
| Conclusions | Provide specific interpretation of results including precision (confidence limits and bias) of treatment difference
Discuss limitations and strengths of study
Discuss results in context of previous studies
Discuss need for future studies and propose specific testable hypotheses based on results |
| References | Support introduction and methods
Support interpretation and discussion in light of previous work |

Resource Bibliography


