Chapter 1
CHAPTER - 1

INTRODUCTION TO CLINICAL TRIALS AND RECEIVER OPERATING CHARACTERISTIC CURVES

Clinical trials are research studies in which people help doctors find ways to improve health and disease care. Each study tries to answer scientific questions and to find better ways to prevent, diagnose, treat. Most clinical research that involves the testing of a new drug progresses in an orderly series of steps, called phases - US National Institute of Cancer.

1.1 Introduction

A Clinical trial is prospective study designed to determine the effectiveness of a treatment or a surgical procedure administrated to patients with a specific disease. The main purpose of a clinical trial is to find a better way to prevent, diagnose or treat a disease. These clinical trials are performed to understand the effect and safety of a new drug which is to be used for testing its potential efficacy. Usually, the new drug or a surgical procedure is first experimented on laboratory animals.
The evolution of clinical trials and their developments is worth learning. From the available information it is known that from the ancient days to till date, significant changes have been brought out in the process of clinical trials. The contributions in subsequent years in this area were made by Avicenna (980-1037), Leonardo da Vinci (1451-1519), Ambrose Pare (1510-1590), Maitland (1668-1748), Lady Wartley Mantague, James Lind (1716-1794), William Withering (1741-1799), a few to mention.

Later in nineteenth century P. C. Alexander Louis (1878-1872) brought out the evolution of clinical trials towards the application of Statistics in Medicine. In the succeeding years Abraham made a trial on Penicillin and another well-known trial was made by the British Medical Research Council (BMRC) under the leadership of Sir Austin Bradford Hill in 1946, to evaluate Streptomycin in the treatment of Pulmonary Tuberculosis.

When a new treatment for a particular disease is proposed, the key question is whether the treatment is safe and efficient to be adopted by physicians as part of their routine clinical practice. Except for diseases that are completely intractable, addressing this question requires comparison of the new treatment to existing standard therapies. Inevitably, the evaluation process also involves ethical, logistical, economic and regularity issues. To deal with this complexity, in recent decades the medical and scientific communities have developed an elaborate infrastructure for conducting clinical trials, which are experiments to evaluate the effects of medical treatments on human subjects. In order to account for the uncountable variation in the experimental
environment, statistical methods are required for designing a clinical trial and to perform data analysis.

1.2 The Three-Phase Paradigm of Clinical Trials

In a clinical trial, a new treatment E is to be evaluated at each stage of process. It is conducted in different phases, each phase addressing queries in a distinct experimental environment. This is called the three-phase paradigm.

The characteristics of the three phases are outlined below.

**Phase I:**

1. The main goal is to find the safest dose and to observe the side effects of a new drug.

2. Limited number of patients are involved and patients who have undergone standard treatment

3. The maximum tolerance dose is established.

4. Small cohorts are formed and the knowledge of investigator is essential

5. If the new drug is found safe than it is considered worthy for further study

**Phase II:**

1. This phase contains two components, viz., Phase II A and Phase II B.

2. In phase II A, the effectiveness of the drug/treatment is determined. If the drug/treat is effective, we proceed to II B; else stop.

3. Phase II B, is used to estimate the therapeutic effectiveness of the drug is estimated

4. Less number of patients are involved
5. The two-way procedure is also known as Double-Sampling (Cox, 1958)

**Phase III:**

1. This phase is a comparative one, where the new treatment is compared with standard treatment, to identify the better one.

2. Involvement of patients is high.

3. Patients are divided into two groups, one receiving the standard treatment (control group) and the other receiving new treatment (experimental group).

4. This phase needs some careful planning and design such as time, number of treatments, duration, treatment allocation ratio and so on.

5. In this phase, major importance is given to statistical techniques.

In the simplest form the *three-phase clinical trial paradigm* begins with a phase I trial to determine the safe dose of new treatment.

In phase II, each patient's outcome is characterized by a binary variable Y, indicating whether a desirable response to therapy with E has been achieved or not.

If the response rate in phase II is high then it is rationale to conduct a confirmatory phase III trial. A phase II trial should be smaller and completed much more quickly, in order to provide a feasible means to screen E.

In phase III, the new treatment is compared with a standard treatment, S. In this phase patients are randomized between E and S in order to obtain unbiased comparisons.

Like any experiment, a clinical trial also requires a plan or a *design* for collection of data. The different activities of a clinical trial can be shown by a flow chart given in figure 1.1.
Figure: 1.1: The logical process of clinical trial

* MTD: Minimum Tolerance Dose
The normal duration of a clinical trial goes for a period of 10-12 years. After obtaining FDA approval another phase identified as Phase – IV is adopted. Every clinical trial invariably needs the approval from Food and Drug Authority (FDA) before commencement of phase IV.

Phase IV trial is about the survivalance studies or post-marketing research about a new drug which has been approved for marketing. This phase is processed either in hospital or general practitioners following the guidelines of Good Clinical Practice (GCP). The outcomes of the phase-IV trial are monitored to collect data to address issues that support product success in a real world clinical practice. The main reasons for processing phase IV trial is to find out the side effects and safety of the drug, risks and benefits in long term use of the drug and also to observe how well the drug works when it is used more widely in clinical practice.

The following are few statistical techniques which are useful in phase III of a clinical trial.
1.3 Statistical Designs

Two types of statistical designs are used for conducting a clinical trial. There are briefly outlined below.

1.3.1 Fixed-Sample trials: In these trials the sample size is fixed. The assignment of treatments is made by following randomization methods.

a) *Simple Randomized Design*: This is most commonly used in which the allocation of treatments to patients is done with the help of random numbers. If A and B are two treatments to be allocated to 10 patients the assignment is done as given below.

- Read one random number from tables or compute one
- If the number is odd, assign treatment A otherwise assign B
- Continue until 10 patients are assigned.

This is easy to implement but leads to unbalanced allocation.

b) *Stratified Randomized Design*: In this design, patients are grouped into 'strata' using a categorical factor like gender. Within each stratum patients are randomly assigned to treatments. Suppose, a study involves n patients and with two treatments A and B: these n patients are divided into two strata of say \( n_1 \) and \( n_2 \). Now within each stratum, patients are randomly assigned to treatments A and B respectively.

c) *Cross Over Design*: It is a combination of simple randomized design and paired comparison design. A common way of using this design is to give the sequence of treatments A followed by B to the first half of the patients and sequence B
followed by A to the other half. A patient is assigned to one of the two sequences by random allocation.

d) Factorial Design: Factorial designs are used to study the effect of different treatment combinations simultaneously. In a \((r \times s)\) factorial design, one of the two treatments is administered at \(r\) levels, the other at \(s\) levels. The goal is estimate the main effects and possible interactions of treatments like drug and placebo.

1.3.2 Sequential trials: In these trials sampling is continued until there is enough evidence to stop further sampling. So, the sample size is not fixed in these trials. The following are two such designs.

a) Open Sequential design: In this procedure, patients are taken into the study in pairs; one is randomly assigned to treatment A and the other to treatment B. The order of administration is random and the treatments are made indistinguishable to the patients. Basing on the distribution type and critical values \((a, b)\), the trial is terminated if \(x \leq a\) or \(x > b\). This helps the investigator to observe the difference in the response due to treatments and also to take a decision on whether the trial is to be terminated or not.

b) Closed Sequential design: This design is based on boundary points which are from the tables of Binomial distribution and Normal distribution. In this design preferences are drawn between the treatments. Here, pairs of patients will be assigned to one or the other treatment by tossing a coin. Basing on the boundary points only, termination of the trial as well as preferences regarding treatments are
drawn. If the difference between two treatments is not much appreciable, then the trial can be terminated.

1.4 Determination of Sample Size

Determination of sample size is an important aspect in the design of a clinical trial. In many studies particularly those related to long term treatment of patients, the researcher is constrained to censor the observations. There are two popular methods to determine the sample size as given below. (Elisa T. Lee, 1992)

a) Sample size calculation for comparing two response rates:

If A and B are two treatments for comparison, then the expression for sample size \( n \) is

\[
 n = \frac{(z_{a}+z_{\beta})^2}{2(\sin^{-1}\sqrt{p_{1}} - \sin^{-1}\sqrt{p_{2}})} \tag{1.4.1}
\]

where \( p_{1} \) and \( p_{2} \) are the response rates to treatments A and B.

\( Z_{\alpha} \) and \( Z_{\beta} \) are upper percentage points of standard normal distribution corresponding to type I and type II risks \( \alpha \) and \( \beta \), respectively.

b) Sample size calculation for comparing two survival distributions:

For comparing the two survival distributions the total sample size needed is

\[
 N = \frac{(z_{a}+z_{\beta})^2}{(\lambda_{1} - \lambda_{2})^2} \frac{[\Phi(\lambda_{1})Q_{1} + \Phi(\lambda_{2})Q_{2}]}{[\Phi(\lambda_{1}) + \Phi(\lambda_{2})]} \tag{1.4.2}
\]

where \( \Phi(\lambda) = \frac{\lambda^{2}T}{\lambda T + e^{\lambda T} - 1} \) and

\( Q_{1} \) and \( Q_{2} \) are the sample fractions for the two treatment groups and in case of equal sizes, the

\[
 N = \frac{(z_{a}+z_{\beta})^2}{(\lambda_{1} - \lambda_{2})^2} \frac{[\Phi(\lambda_{1}) + \Phi(\lambda_{2})]}{[\Phi(\lambda_{1}) + \Phi(\lambda_{2})]} \tag{1.4.3}
\]
1.5 Randomization and Blinding

Results of a statistical analysis depend on the nature of the data and the technique used. In a comparative trial these results depend on how patients were assigned to treatments. Randomization is the process of assigning subjects to treatment groups and this technique gives each subject equal chance of being assigned to any of the groups. Successful randomization requires that assignment of groups cannot be predicted in advance. The process of randomization aims to ensure similar levels of all risk factors in each group; not only known but also unknown characteristics which are comparable results in similar numbers or levels of outcomes in each group, except for their the play of chance or a real effect of the interventions. Randomization is one way to achieve comparability. The process of random allocation is done by a statistician who is usually a member of the clinical trial Team.

In this section we outline the techniques of randomization, like Simple Randomization, Permuted Block randomization, Stratified allocation, dynamic random allocation method.

a) **Simple Randomization:** This is the commonly used method of random treatment assignment and simple to apply. There are many informal ways of assignment such as flipping a coin, use of tables of random numbers and shuffling numbered cards.

b) **The Permuted Block Randomization:** This is used to maintain good balance in assignments and also whenever the trial is small. In this technique, subjects are grouped into several blocks of equal size according to their chronological entry time. Within each group of subjects, the treatments are assigned so that
there is a balanced allocation for each treatment. In choosing a particular block to be allocated, a random number sequence will be used. This technique ensures that treatment group numbers are evenly balanced at the end of each block.

The methodology behind this technique is totally different when we compare with simple and permuted block randomization methods. Simple and Permutated block randomization methods are defined and allocation sequences are set up before the start of the trial.

c) **Dynamic Randomization:** Here the allocation of subjects to treatment groups is done by checking the allocation of similar patients already randomized and allocating a new case to next group so as to balance the treatment groups across the stratification variables. The following terminology is popular among medical researchers.

Blinding is a scientific method used in allocation of treatments to patients aimed at preventing the observer's bias in the research outcomes. It is a very important tool in almost all fields of research. Here are some commonly used blinding methods.

1. **Single Blinding:** In this method only the investigator knows whether a patient is taking the standard treatment or the new treatment being tested.

2. **Double Blinding:** In this method neither the investigator nor the patient knows which of several the possible therapies/drug is administered. In this method neither the investigator nor the patient knows the identity of the drug given to a patient. However, the statistician will keep a confidential document about the blinding protocol.
3. **Triple Blinding**: Triple-blind trials are double-blind trials in which the statistician interpreting the results also does not know which investigation has been given. It is used to mean that multiple investigators are all blinded to the protocol.

### 1.6 Analytical Tools for Clinical Trials

The data collected from a clinical trial has to be analyzed with the following objectives

a) Estimating the effect of treatment/procedure

b) Testing the statistical significance of the observed effects

c) Exploring functional relationship between dose and response

d) To develop rules for discriminating between healthy and diseased subjects.

Commonly used statistical tools include summary statistics like mean and standard error, skewness and kurtosis for presenting data. Correlation and Regression analysis are used to study functional relationship among the variables. The effect of treatments and prognostic factors on the response is studied using experimental designs/response surface methodology. Special tools like Stepwise regression, Logistic regression, Poisson regression are used.

Classification problems arise when new therapies or diagnostic tools are proposed by the researcher. The ability of a therapy/tool to discriminate between healthy and diseased subjects is often analyzed with the help of tools like Logistic regression and Discriminant analysis. Another interesting tool used in medical research for classification is called **ROC Curve analysis**. With the advent of software, the ROC analysis is highly adopted in Medicine and Radiology apart from non-medical...
applications like Signal Detection Theory. In the following section, the basics of ROC curves are outlined. Chapter II, however contains mathematical details of ROC analysis.

1.7 The ROC Curve and Its Ramifications

The word ROC analysis origins from Statistical Decision Theory as well as Signal Detection Theory (SDT) and was used during II World War for the analysis of radar images. In these studies, the objective is mainly to distinguish between the two possible outcomes of a dichotomous event like signal/noise or diseased/healthy. In medical diagnosis the term noise refers to Healthy status and signal refers to the presence of Disease. In the theory of statistics, this is the well known problem of classification.

The distinction between the two groups of cases is normally made basing on a threshold or cutoff. Assuming that there is a cutoff, any value exceeding cutoff indicates the existence of signal (presence of disease) and any value less than or equal to the cutoff represents noise alone (absence of disease).

A valid measure of correctness of classification with a cutoff value can be obtained from a tool called Receiver Operating Characteristic Curve (ROC Curve). This measure is found suitable in almost all kinds of systems such as weather forecasting, flaws in manufacturing products, aptitude testing, psychometry, data communication etc.

During 1960’s same measure was applied in the medical field for discriminating the subjects (patients) regarding the presence or absence of a disease. ROC analysis is
also applied to compare the efficiency of two or more statistical tools for a classification problem.

The ideal classification:

Suppose a cutoff value or a statistical model is used to discriminate between healthy and diseased patients. In most cases there exists a criterion called gold standard which unambiguously divides the patients into two groups. If we assume that the diagnostic test result follows normal distribution in both the groups the gold standard produces two normal curves without overlap shown in figure 1.3

![Figure 1.3 Probability distributions of an ideal system](image)

The gold standard is a valid and widely accepted set of procedures, which leads to perfect classification. Clinical experts and biomedical researchers concentrate on cost effective and patient-friendly procedures that are capable of classifying the patients as closely as possible to the classification made by the gold standard. Such an alternative procedures are often known as Biomarkers.
1.8 Brief Review of ROC Curves and Related Concepts

Much work in the area of ROC curves was reported by Green and Swets (1966, 1979). Leo Lusted, a Radiologist suggested using ROC analysis in Medical decision making in 1967 and his original description of ROC is, "Plotting false – positive diagnosis and true – positive diagnosis on the X – axis and Y – axis respectively". Metz (1978) stated that ROC analysis is useful to determine the discriminating ability of a diagnostic test. In later years, eventually ROC analysis made its way into other areas of medicine.

ROC curve analysis gives the accuracy of a diagnostic test. Diagnostic accuracy is the most fundamental characteristic of the test itself as a classification device. It measures the test ability to discriminate among alternative states of health.

The result of a diagnostic test could be either discrete or continuous.

a) When the outcome is ordinal like normal, moderate and severe, then X would be a discrete random variable.

b) When the test result X is a measurement, like the blood cholesterol level of a patient, then X would be a continuous random variable.

While comparing a test result with the actual diagnosis, there will be four possible states as follows:

True Positive: If both diagnosis and test are Positive
False Positive: If the diagnosis is negative but the test is positive
True Negative: If both the diagnosis and test is Negative
False Negative: If the diagnosis is positive but the test is negative
These states are shown below, which is often known as *confusion matrix*.

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Test Result</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Negative</td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td>TOTAL</td>
<td>Q</td>
<td>Q'</td>
</tr>
</tbody>
</table>

Table: 1.1: Confusion Matrix

The above four possibilities are similar to the four states in the classical test of hypothesis scenario. Let $X$ denote the test result and the classification is positive if $X > c$ and negative otherwise. Let $D$ denote diseased and $\overline{D}$ denote Healthy states of a patient. Then we get the following cases.

<table>
<thead>
<tr>
<th>Status</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>$[X &gt; c / D]$</td>
</tr>
<tr>
<td>FP</td>
<td>$[X &gt; c / \overline{D}]$</td>
</tr>
<tr>
<td>TN</td>
<td>$[X &lt; c / \overline{D}]$</td>
</tr>
<tr>
<td>FN</td>
<td>$[X &lt; c / D]$</td>
</tr>
</tbody>
</table>

Table: 1.2

When the value of ‘c’ changes, this classification gets modified.

Define $n_{TP}$ = number of True Positive cases

$n_{FP}$ = number of False Positive cases

$n_{TN}$ = number of True Negative cases

$n_{FN}$ = number of False Negative cases
The following fractions or proportions are essentially used in the analysis:

True Positive fraction (TPF) = \( \frac{n_{TP}}{n_{TP} + n_{FN}} \)

False Positive fraction (FPF) = \( \frac{\gamma}{\gamma + \gamma_{TN}} \)

In the following section, some measures of performance are discussed.

1.9 Specificity, Sensitivity and the ROC

ROC curves focus on the parameters like sensitivity and specificity, whose values lie in the interval \([0, 1]\). When the cutoff value for a continuous diagnostic variable is increased, both the TPF and FPF will decrease. We define the following parameters which are fundamental in ROC analysis.

Sensitivity: Sensitivity is the probability that a test result will be positive when the disease is actually present. In other words, it is the probability of having a positive test among the patients who have a positive diagnosis. This is denoted by \( S_n = P(X > c | D) \), which is estimated from the available data as a the relative frequency given by

\[
S_n = \frac{n_{TP}}{n_{TP} + n_{FN}} = \text{TPF} \tag{1.9.1}
\]

The change in TPF with respect to the threshold/cutoff is shown below.

![Figure 1.4: Shows the influence of the threshold value on sensitivity](image)
Specificity: Specificity is the probability that a test result will be negative when the disease is not present. In other words, it is probability of having a negative test among the patients who have a negative diagnosis, denoted by $S_p$

$$S_p = \frac{N_{TN}}{N_{TN} + N_{FP}}$$ (1.9.2)

![Diagram showing specificity](image)

Figure 1.5: Shows the influence of the threshold value on Specificity.

It is obvious that $S_n + S_p = 1$ or $S_n = 1 - S_p$ (1.9.3)

If a variable has no diagnostic capability, then a test based on that variable would be equally likely to produce a false positive or a true positive.

This equality is represented by a diagonal line from (0, 0) to (1, 1) on the graph of the ROC curve. Any ROC curve starts at (0, 0) on the XY-axis. A typical ROC curve is obtained for each cutoff value.

A graph of $S_n$ against $(1 - S_p)$ is called the Receiver Operating Characteristic (ROC) curve. Each cutoff $C_i$ corresponds to a point $(1 - S_p, S_n)$ on a ROC curve. The value of ROC lies between 0 and 1. If the test is perfect then $1 - S_p = S_n = 1$, if $1 - S_p = S_n = 0$ the ROC curve becomes a straight line from (0, 0) to (1, 1).
1.10 Shape of the ROC Curve

When the test result is discrete we get a pair of value $S_n$ and $1-S_p$ for each cutoff value.

Since the number of such cutoffs is at most denumerable, we get a step function for the ROC.

Let $ROC(C_i, u)$ denotes the ROC value corresponding to $u = 1 - S_p$ at a cutoff value $C_i, i = 1, 2... k$. Then the jump at $C_i = ROC(C_i, u) - ROC(C_{i+1}, u)$

The shape of the ROC curve is shown in Figure 1.7

![ROC Curve](image)

Figure 1.7: ROC Curve

On the other hand when $X$ is continuous, it is possible to have 'infinite' number of possible cutoffs.

1.11 Area Under The Curve (AUC)

The accuracy of a diagnostic test can be explained by using the Area Under the Curve (AUC) of an ROC curve. AUC describes the ability of the test to discriminate between diseased and non-diseased. If, $AUC = 0.80$ one will be able to correctly identify the diseased patient in 80% of the cases. AUC gives us information about the
general "goodness" of a test and not the interpretation of a test result. ROC curve
starts at the point (0, 0) and ends at (1, 1) and the diagonal line separates the area into
two halves having 0.5 units in each half. The AUC lies between 0.5 and 1. Value
AUC = 0.5 for a test indicates 50% chance for correct classification and hence a good
test shall have AUC \( \geq 0.5 \). Test for which AUC < 0.5 need not be considered at all.
Higher the AUC, better will be the diagnostic test. AUC can be interpreted as a
probability that a randomly selected subject, with disease, will have a higher test result
compared to a control subject.

The categorization/rating of the test efficiency, is often linked to the AUC and here are
some guidelines used in practice.

<table>
<thead>
<tr>
<th>AUC</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50 to 0.60</td>
<td>Bad</td>
</tr>
<tr>
<td>0.60 to 0.70</td>
<td>Poor</td>
</tr>
<tr>
<td>0.70 to 0.80</td>
<td>Fair</td>
</tr>
<tr>
<td>0.80 to 0.90</td>
<td>Good</td>
</tr>
<tr>
<td>0.90 to 1.00</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

The above are the rough guidelines for classifying the accuracy of a diagnostic test.

1.12 Methods for Estimating AUC

When the diagnostic tests result is categorical, then it is easy to evaluate the AUC with
the help of a non-parametric approach and if it is of continuous type, then there is a
need of mathematical solutions. Two popular methods are in use, one basing on
logistic model and the other based on the assumption of normal distribution.

a) In the logistic model method, classification of subjects is based on a cutoff
value supplied into the logistic model. Each cutoff produces a different set of
false and true classifications of the subjects, originally classified by the gold standard. Hence, this method is categorical itself. It is interesting to note that logistic model becomes the diagnostic tool (instead of general biomarkers). Thus the efficiency of the cutoff used in logistic model, can be assessed with ROC curve.

b) The second method assumes that the test result is continuous and follows normal distribution. Two normal distributions with some overlap when combined, produces a model called Binormal distribution.

Dorfman and Alf (1969) gave a Maximum Likelihood (ML) Estimation methodology for an ROC curve. Later Hanley (1988) showed the robustness of the Binormal assumption. Egan (1975) proposed the following power law for fitting the ROC curve, when the distribution of results is negative exponential in both groups.

\[ TP = FP^k \]  

(1.11.1)

where \( k \) is the ratio of two means of normal and abnormal population. The ROC curve fitted by this method is called EROC (Exponential ROC Curve). England (1988) proposed exponential equation for identifying the optimum threshold of ROC.

In 1990, Goldard and Hindberg recommended using a non-parametric approach, called Trapezoidal rule to calculate AUC. Bamber (1975) first noticed the relationship of Trapezoidal rule with the Mann-Whitney U statistic. This allows calculating both areas and standard error. They showed that the nonparametric area has equivalent statistical power for comparing ROC curves as that of parametric method. Ogilive and Creelman (1968) proposed a ML Estimation method for fitting ROC using Logistic
distribution. They have shown that the difference between logistic distribution and normal distribution is indistinguishable.

Focus of the Thesis:

In this thesis a new method of estimating the Area Under the Curve (AUC) associated with the Receivers Operating Characteristic (ROC) curve of a diagnostic test using a Binormal model is proposed. A new estimator is proposed which takes into account the confidence interval of the estimated means while evaluating the AUC. A weighted average of the AUC based on the worst case differences between the means, of the two distributions is used as an estimator of the true AUC and three different types of weights are proposed. The performance of the estimators is numerically demonstrated and the sensitivity with respect to sample size is illustrated. Focus is also on developing a spreadsheet solution to derive ROC and AUC, as an alternative to standard software solutions. The utility of excel paste functions is explored to produce ROC curves and to run simulated trials.

In the next chapter various methods of estimating the ROC curve and the corresponding AUC are reviewed.