INTERNATIONAL RESEARCH JOURNAL OF PHARMACY

www.irjponline.com

ISSN 2230 – 8407

Review Article

# BIOSTATISTICS IN CLINICAL RESEARCH: A REVIEW

Aneesha Chatla [1]*, Bhargavi Neela [1], C. S. Mujeebuddin [2], V. C. Randeep Raj [3]
[1]Intern at Clinosol Research Pvt. Limited, Ameerpet, Hyderabad, Telangana, India
[2]Founder and CEO of Clinosol Research Pvt. Limited, Ameerpet, Hyderabad, Telangana, India
[3]Clinical SAS programmer, Clinosol Research Pvt. Limited, Visakhapatnam, Andhra Pradesh, India
*Corresponding Author Email: aneesha29.chatla@gmail.com

Article Received on: 09/01/21 Approved for publication: 31/01/21

**DOI: 10.7897/2230-8407.1201109**

## ABSTRACT

Statistics is the discipline concerning collection, organizing, analyzing, interpretation and presentation of data as the basis for explanation, description and comparison. In clinical trials and in the drug development process, statistics play a key role, from trial design to protocol development. The credibility of a clinical trial can be upheld and cooperation between physicians and statisticians can be strengthened by providing a fundamental understanding of statistical issues. In any phase of clinical research, including trial design, development of procedures, data management and tracking, data processing, and reporting of clinical trials, biostatistics are involved. Statisticians also have roles in formulate hypothesis, develop statistical analysis plan (SAP), choosing the appropriate test, choose an apt sample size, data collection, perform the tests, generating TLGs (tables, listings, and graphs) and reporting the inferences. It is important that the rest of the research team recognizes the statistical approach suggested by the biostatistician, because statisticians can specialize in study designs, therapeutic areas and statistical methods.

**Keywords:** Biostatisticians, clinical research, statistical analysis plan, sample size, TLFs.

## INTRODUCTION

Statistics is the discipline concerning collection, organizing, analyzing, interpretation and presentation of data as the basis for explanation, description and comparison. It is customary to start with a statistical population or a statistical model to be studied when applying statistics to a scientific, industrial, or social problem[1]. Biostatistics is a statistical branch and is described as the statistical processes and methods used to collect, analyze and interpret biological data, in particular data relating to biology, health and medicine in humans. Biostatistics covers not only health, medicine and nutrition applications and contributions, but also areas such as genetics, biology, epidemiology, and several others. Biostatistics consists primarily of numerous steps, such as hypothesis generation, data collection and statistical analysis implementation[2].

Karl Pearson quoted that "Statistics is the grammar of science".

The individuals who are concerned with the collection and analyzing data from living organisms are specialists called Biostatisticians. They work mainly with data gathered via medical research, which they use to formulate conclusions and make predictions[3].

## IMPORTANT ROLES OF BIOSTATISTICS IN CLINICAL RESEARCH

Biostatistics help in Clinical Research right from its start for: Designing, conducting, analyzing, reporting, minimizing biases, confounding factors, measuring random errors, understanding the research, make suggestions on hypothesis testing & analysis, determine the power of the study, calculating the sample size, ensure continuity throughout the research, assess the statistical significance of the results, efficacy & safety of the drug, line of treatment and therapy.

**PROTOCOL DESIGN:** In the design of protocols in clinical research, the functions and responsibilities of biostatisticians are as follows:

**Objectives:** Biostatisticians must have a simple specification of the hypothesis to be evaluated based on the purpose of the scientific manuscript. They have to provide, in other words, the parameters to be assessed. In clinical study, they are also responsible for choosing and identifying endpoints.

**Study Design:** The biostatistician's study design should provide the data needed to address the goals, such as: Defining processes to eliminate bias in selection; in the case of the Randomized Control Trial, establishing randomization processes such as concealment of sequence generation and allocation, and the duration and frequency of contacts to follow up.

**Sample size:** Biostatisticians responsibilities in calculation of sample size include: the methods used to measure the sample size should comply with the primary data analysis method and should also be suitable for the design, justifying the primary endpoint in power or precision terms, assumptions of proper historical data should be supported and rationale in terms of feasibility.

**Analysis plan summary:** The purpose of analyzing plan summary is to assure objectives to be achieved and to justify design and data collection by ways like provide statistical methodology for the evaluation of primary goals, such as testing procedures, statistical hypotheses, and the discussion of statistical approaches to be used in the intended interim analyses as per the Data Safety Monitoring Board (DSMB).

**Protocol Review:** It is pertinent for the lead study biostatistician to review the full protocol for checking the Clarity, Completeness, Consistency Data quality issues and feasibility.

**Protocol writing:** To write analysis plan with the inputs from objectives, endpoints study design, and allocation concealment, randomization procedures and sample size.

**DATA MANAGEMENT:** Management of design and content of Case report form (CRF), Validation with the specification of error checking and test data and specification of Dataset with CRFs annotation and record layout.

**STUDY IMPLEMENTATION:** It involves implementation of procedures for randomization and sampling selection.

**STUDY MONITORING:** It involves monitoring for safety and efficacy and quality.

**DATA ANALYSIS:** It helps to plan for the reporting, writing of manuscripts and the validity and creditability of the findings and a comprehensive study strategy for all theories to be evaluated along with the study hierarchy.

**REPORTS OR MANUSCRIPT WRITING:** Methodology section with statistical analysis used that are used for interpretation of the results; data overview with endpoints and design, the result section contains details provided in the form of a graph, tables and others and discussion portion with the required explanation of the findings[4].

- Informs methods on data collection by preemptively identifying statistical tests.
- To support or negate a hypothesis with volumes of data communications.
- Seeks out uncertainty, errors and outliners in the data through data visualizations.
- Aids interpretation, summarization, and communication of datasets.
- Multivariate statistics and modeling help deal with our multivariate statistical questions so that hypotheses could be assessed from every possible angle[5].

There are two major branches of the statistical system mainly descriptive and inferential. **Descriptive statistics** describe the distribution of population measurements by presenting data types, central tendency estimates (mean, mode and median) and variability measurements (standard deviation, coefficient of correlation), whereas **inferential statistics** are used to express the degree of certainty about estimates which include hypothesis testing, confidence interval and standard error of mean[6].

## TYPES OF DATA

Data constitutes findings reported during research study. There are three kinds of data, namely nominal, ordinal, and interval data. Statistical analytic methods rely primarily on the form/type of data. In general, data shows the variability and central tendency. Thus, types of data are also very important to understand.

**Nominal data:** is analogous with categorical data where the data is strictly allocated to categories based on the presence or absence of specific characteristics without any ranking between the categories[7].

For example: patients are classified as men or women by gender. It also includes binominal/dichotomous data, which refers to two likely results that may be death or survival, such as in cancer outcomes.

**Ordinal data:** It is also referred to as graded/ordered/categorical data. This data form is usually represented as ranks or scores. There is a natural order between groups, and it is possible to organize them in order[7].

For example: How you feel today, for instance, can be graded as very sad, sad, okay, happy, and very happy.

**Interval data:** This type of data is characterized by an interval between two measurements that is equal and definite. Celsius temperature, for example, since the difference between each value is the same. A measurable 10 degrees is the difference between 60 and 50 degrees, like the difference between 80 and 70 degrees.

The interval data type may either be continuous or discrete. Within a given range, a continuous variable can take on any value. For example, the hemoglobin level can be taken as 11.5, 12.8, 13.2 gm%, while integer values are typically assigned to a discrete variable, which means that it does not have fractional values. Blood pressure values are usually discrete variables.

Often to minimize dissymmetry and make it meet the normal distribution, certain data can be transformed from one form to another form. Drug doses, for example, are converted to their log values and plotted in the dose response curve to obtain a straight line to facilitate the analysis[8].

## ROLE OF BIOSTATISTICIANS

Statisticians analyze and design research studies to reduce bias and achieve study goals, in which they have the expertise and experience needed to collaborate with therapeutic professionals and clinicians. The fundamental research study is established in a quantifiable way to[9]

- Formulate hypothesis
- Develop statistical analysis plan (SAP)
- Choosing the appropriate test
- Choose an apt sample size
- Data collection
- Perform the test
- Build TLGs (tables, listings, and graphs); and report the conclusion, finally.

## FORMULATING HYPOTHESIS

Hypothesis implies an unproven assumption or opposition that provisionally states certain facts or phenomena. Hypothesis testing is a collection of logical and statistical rules that are used to make choices regarding population characteristics from sample statistics. The object of hypothesis testing is to analyze two opposing speculations formally, i.e. Null hypothesis ($H_0$) and Alternative hypothesis ($H_A$)[10]. In hypothesis testing, the basic concepts are:

- Null hypothesis & alternate hypothesis
- Level of significance
- Critical region
- Decision rule (Test of hypothesis)
- Type I & type II errors
- Power of Study
- Two tailed & one tailed tests
- One sample & two sample tests

The main objective of statistical analysis is to decide if the effect produced by a compound under study is actual and is not accidental. A test of statistical significance is also typically added to the analysis. In a test like this the first stage is to state the null hypothesis.

### Null hypothesis

We assume that there are no differences between the two groups in the null hypothesis (statistical hypothesis). A new drug 'A' is claimed to have neuroprotective effects, for instance, and we want to test it with a placebo. The null hypothesis in this study is that drug A is no better than placebo.

### Alternate hypothesis

The alternative hypothesis (research hypothesis) states that two groups differ. There would be a distinction between a new drug 'A' and a placebo.

When the difference between two groups is not significant, the null hypothesis is accepted. This indicates that both samples were taken from a single population, and the difference between the two groups was due to chance. If alternative hypothesis is confirmed, i.e. null hypothesis is rejected, then the difference is statistically significant between the two groups. The difference between the drug 'A' group and the placebo group that would have occurred by chance is less than 5% of the cases, which is less than 5 of the 100 times is labelled statistically significant ($P < 0.05$). There is the possibility for two errors to occur in any experimental procedure[6].

### Level of significance

We mean it is not unusual or not rare if the probability (P) of an outcome or event is high. But if there's a low probability, we assume it's unusual or rare. A rare outcome is considered significant in biostatistics, whereas a non-rare event is called non-significant. The 'P' value at which we consider an outcome or event, sufficient to be considered relevant is called the level of significance[11]. In clinical studies/research, less than 0.05 or 5% of the P value is most usually regarded as a significant value.

### Critical Region

It is the area of rejection. The null hypothesis is rejected if the mean value falls within this area. The critical value is the value of the test statistic above which it is possible to reject the null hypothesis[10].

### Decision Rule (Test of hypothesis)

The Decision Rule is the rule by which the null hypothesis is accepted or rejected[10].

### Type I & type II errors

**Type I error (false positive):** It is also known as α error. In a statistical hypothesis testing, it is the probability of finding a difference; when no such difference actually exists, which results in the acceptance of an inactive compound as an active compound. Such an error, which is not rare, can be accepted because the compound would be exposed as inactive in subsequent trials and thus eventually rejected[12]. In fact, the Type I error is fixed in advance by choosing the level of significance used in the test[13].

**Type II (false negative) error:** This is often referred to as the β error. It is possibility that the difference will not be detected when it actually exists, leading to the dismissal of an active compound as an inactive compound. This error is more serious than the type I error because there is a chance that no one will try it again after we have classified the compound as inactive. An active compound would therefore be lost[12]. By taking larger samples and using a sufficient dose of the compound under study, this type of error can be reduced. These two types of error rates are traded against each other: an attempt to minimize one type of error usually results in an increase in the other type of error for any given sample set[14].

### Power of a test

It is a possibility that studies will display a difference between the groups, if the difference actually exists. To pick up the higher chances of existing differences, a more powerful study is needed. By subtracting the beta error from 1, power is calculated. Power, therefore, is (1-β). Power of study is very significant in the calculation of sample size. Any research study should have at least 80% power to be scientifically valid. If the research power is less than 80% and the difference between groups is not significant, then we can assume that difference between groups could not be detected, rather than no difference between the groups. If we increase the study power, then the sample size increases as well. At the initial level of research, it is often better to determine the power of study[6].

### One-tailed and Two-tailed Test

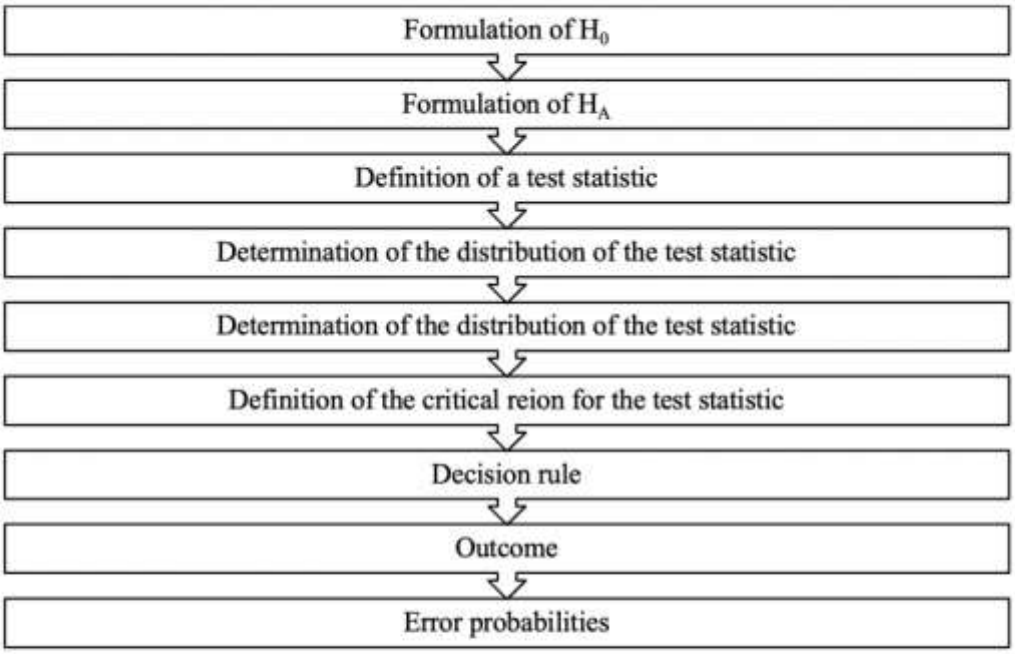
The null hypothesis when comparing two groups of continuous data is that there is no real difference between the groups (A and B). The alternative hypothesis is that the groups have a real difference. This difference could be in either direction, e.g. $A > B$ or even $A < B$. If there is a certain way to know beforehand that the difference can only be in one direction, e.g. $A > B$ and when only one possibility is considered, the test is called a one-tailed test. The test of significance is known as a two-tailed test if we consider both of the possibilities.

For instance, if we know that English boys are taller than Indian boys, the result would be at one end, which is a one tail distribution, so one tail test is used. If the direction of difference is not completely certain, which is common, it is often easier to use the two-tailed test[15].

### One Sample & two sample tests

On the basis of a given sample, one sample test is when we draw conclusions about the population and two sample test is when we try to compare and draw conclusions about two populations on the basis of a given samples[10].

## STEPS INVOLVED IN HYPOTHESIS TESTING


Formulation of $H_0$
Formulation of $H_A$
Definition of a test statistic
Determination of the distribution of the test statistic
Determination of the distribution of the test statistic
Definition of the critical reion for the test statistic
Decision rule
Outcome
Error probabilities

## STATISTICAL ANALYSIS PLAN (SAP)

The SAP is a scientific document explaining the mathematical methods of research analysis, while the protocol is the one that describes the analysis. SAP gives statistical analysis of clinical trials it determines all of the statistical outputs to be used in the report of the clinical trial. The most widely used documents for mathematical programmers to build their deliverables are the SAP and the annotated CRFs. The SAP offers important data to the statistical programmer. They are carefully revised by statistical programmers. It gives clarity and comprehension. It helps in data consultation to prepare TLF. The expected review of clinical trials is given in the statistical analysis plan (SAP)[16].

This Statistical Analysis Plan (SAP) will include more detailed explanations and comparisons of the endpoints in the report. In certain organizations, statistical analysis plans can also be known as data analysis plans (DAP) or reporting analysis plans (RAP). In order to support the programming of analytical data sets along with analysis and presentation of study results, statistical programmers must build document review skills that facilitate reading material, thorough comprehension and critical review of the SAP.

### Purpose of statistical analysis plan

All the statistical output that will be included in the clinical study report is determined by the SAP. While they should not officially be part of the SAP, Shell tables, figures and occasionally listings are generally added to the SAP. The SAP and the annotated CRF are the documents that are most commonly used to construct their deliverables by statistical programmers. In general, the clinical production of a compound involves four distinct forms of study plan:

**Statistical analysis plan for a clinical study** – describes the expected statistical analysis of a study. **Interim statistical analysis plan** – The proposed statistical analysis of an interim analysis for a sample outlines the interim statistical analysis strategy and therefore needs to discuss the handling of partial unblinding problems in the case of blinded studies. It also explains the potential effect on action and the complete final review, such as the possible modification of the degree of significance.

**Data Monitoring Committee (DMC) statistical analysis plan** – Modification of interim analysis used by DMCs and defines routine (e.g., monthly) safety or effectiveness data tracking procedures. The DMC SAP also includes the DMC Charter, which specifically explains the names and roles of the parties concerned.

**Integrated statistical analysis plan** – describes the planned analysis that is used in submissions, for example. It typically specifies the programming performance information for the ISS and ISE in one paper[16].

Typically, using a template, the SAP is written by the test or project statistician. The SAP should usually include more information than the protocol about the expected statistical analysis. The following material should be included in the SAP, at least:

Brief overview of the research and intent, such as the description of details of actions that are important for analysis, goals of the study and variables.

The statistical methods to be used include summary statistics of the subject data (means, standard deviations, extreme values, percentage counts, etc.), statistical tests (variance analysis, t-tests, etc.). Definition of populations of study-e.g., protection set, per protocol set, complete analysis set. Data handling rules -e.g., imputation rules, derived variables algorithms. Full table of contents with all the TFLs to be generated as an attachment to the SAP text by the statistical programmer (i.e., not as an appendix or any other official and formal part of the SAP) Shell TFLs that the statistical programmer produces to describe the layout of the

TFLs as an attachment to the text of the SAP text (i.e. not as an appendix or any other official and formal part of the SAP).

The clinical trial statistician therefore needs to ensure that the SAP is carefully reviewed and accepted before the research is unblinded. SAP should be tested and accepted for open label studies prior to database lock approval. The SAP is a paper addressed to the regulatory authorities as part of a set of submissions. The SAP is also part of a clinical research report appendix. For all the expected statistical analysis, the SAP is therefore critically essential to record. The SAP is also stored in the trial master file and is used during audits to verify if the definitions in the SAP were exactly followed by statistical programming. The SAP is intended to be a stand-alone report. In addition to the technical statistical data, short explanations and summaries of the protocol should be included. It does not only apply to the protocol itself[16].

### Review by statistical programmer

There should be several phases of the SAP analysis. If these stages can be performed at the same time, it depends on the statistical programmer's expertise. Phase by step, a less experienced programmer might do this analysis. Within a single reading of the text, a more advanced programmer may do all the stages. Nevertheless, the following points should be kept in mind by any statistical programmer:

- Correctness
- Quality
- Ensure protocol consistency
- Ensuring quality in relation to project standards

**Completeness**- check whether all TFLs listed in the SAP text are listed in the TOC and shells are eligible for specific TFLs, check whether all TFLs listed in the TOC and shell TFLs are mentioned in the SAP text.

**Degree of Details**- all required information (e.g. baseline information, algorithms for derived variables, is identified and defined in necessary details.

**Appropriateness**- check if the proposed statistical study is suitable for the function (e.g. tables to describe populations' baseline characteristics, tables to describe primary and secondary variables comprehensively[16].

## HOW TO CHOOSE AN APPROPRIATE STATISTICAL TEST

In biostatistics, there are a variety of studies, but the preference depends primarily on the functionality and type of data processing. Often the difference between means or medians or the relationship between the factors needs to be figured out. The number of groups used in a study can differ, so the nature of the study often differs. Therefore, in such a case when choosing the suitable test, we would have to make the choice more accurate. Inadequate checking can result in invalid assumptions. Statistical tests can be categorized into parametric and non-parametric tests (Table 1). If variables obey regular distribution, data should be subjected to parametric testing and non- parametric testing can be used for non-parametric distribution[6].

**Table 1: Statistical tests applied for different types of data[6]**

| Type of Groups | Parametric Test (Gaussian distribution) | Non-Parametric Test (Non-Gaussian distribution) |
|---|---|---|
| • Comparison of two-paired groups<br>• Comparison of two unpaired groups<br>• Comparison of three or more matched groups<br>• Comparison of three or more unmatched groups<br>• Correlation between two variables<br>• Only two possible outcomes | • Paired 't' test<br>• Unpaired 't' test<br>• Repeated measures ANOVA<br>• One way ANOVA<br>• Pearson correlation | • Wilcoxon- Signed Ranks test<br>• Mann-Whitney test<br>• Friedman test<br>• Kruskal-Wallis test<br>• Spearman correlation<br>• Chi-Square test / Fisher's test |
| ANOVA = Analysis of variance | | |

## CALCULATION OF SAMPLE SIZE

Sample size estimation plays a crucial role in conducting any research. The following five points
should be considered very carefully before calculation of sample size.
First of all, we have to assess the minimum expected difference between the groups.
Then, we have to find out standard deviation of variables.

**Standard deviation (SD)** defines the variability of the observation of the mean[17]. SD is the most useful attribute of variability to describe the population distribution. Summary measures of variability of individuals (mean, median and mode) are also required to be assessed for the reliability of statistics based on individual population variability samples. We need a square of SD called variance to calculate the SD i.e., SD = √variance. The variance is the average square deviation around the mean and the formula to calculate variance is = Σ(x-x-) 2/n OR Σ(x-x-) 2/n-1. SD lets us predict how far the given value is away from the mean, so we can predict the coverage of the values. SD is more fitting only if the data is normally distributed. If individual findings are distributed around the mean sample (M) and are uniformly scattered around it the SD helps to calculate a range that includes a given percentage of the observations.
Now, set the level of significance (alpha level, generally set at $P < 0.05$) and;
Power of study (1-beta= 80%); after deciding all these parameters,
we have to select the formula from computer programs to obtain the sample size.

Various software's are available free of charge for calculation of sample size and power of study. Finally, for non-compliance and dropouts' appropriate allowances are given, and this will be the final sample size for each group in study[6].

### Sample Size Determination and Variance Estimate

The formula includes the knowledge of standard deviation or variance, to determine sample size, but the variance of the population is unknown. Therefore, standard deviation has to be estimated. Sources commonly used for estimation of standard deviation are:
It is possible to draw a pilot[18] or preliminary sample from the population, and to use the variance measured from the sample as

an estimate of standard deviation. As a part of the final sample, Observations used in pilot sample may be counted.

Estimates of standard deviation may be accessible from the previous or related studies[2], but they may often not be accurate.

**Importance of Sample Size Determination**

A fraction of the universe is Sample. Studying the universe is the best parameter. However, a measurement is taken when it is possible to obtain the same result by taking a fraction of the universe. Applying this, we save time, manpower, cost and maximize productivity at the same time. In biomedical studies, thus a sufficient sample size is of prime importance. If sample size is too small, accurate finding will not be given to us and validity is doubtful in such a situation, so the whole analysis will be a waste. In addition, large sample needs higher costs and manpower. It is a waste of money to enroll more subjects than required. A good small sample is much better than a huge, poor study. Therefore, to obtain correct data the required appropriate sampling size would be ethical[6].

**Factors Influencing Sample Size Include**

Prevalence of particular event or characteristics- If the prevalence is high, it is possible to take small sample and vice versa. If prevalence is not established, then a pilot study may be used to achieve it.

Probability level considered for accuracy of estimate- If we need more security about conclusions on data, we need a larger sample. Hence, the sample size will also be larger when the security is 99% than when it is just 95%. If only a minor difference is predicted and if even that small difference needs to be observed, then we need a large sample.

Availability of money, material, and manpower.

Time bound study curtails the sample size as routinely observed with dissertation work in post graduate courses[6].

## DEVELOPING TLGs (TABLES, LISTINGS AND GRAPHS)

In the clinical field, the data are represented and analyzed in the form of tables, listings and figures/graphs (TLGs or TLFs) that are usually generated using SAS[19]. TLFs are usually generated as part of the normal reporting process. In clinical trial studies, the trial data is analyzed and summarized as tables and figures[20]. There are some of the widely acknowledged challenges in the traditional approach to programming TLFs for inclusion in clinical study reports. Some of the well-known issues associated are:

- inefficiency caused by permitting modified programming, rather than reusing code
- resource intensive double-programming for validation
- inconsistency of outputs across and within studies
- errors caused by misinterpretation of analysis
- output specifications[21].

TLFs are also used to answer the regulatory questions and to support publications based on the trial data and one of the critical aspects of any summary result is precision and transparency. Over the years, SAS systems have been the cornerstone of transparency and R and other languages have also recently been used to generate summary results used in trial, regulatory and publication reports. By using a program, log and output, the transparency is recorded, with data going into the program and the output being summary results that are presented in an ASCII, PDF or RTF file in a tabular format[20]. Plotting data and presenting statistical information using SAS ODS and SAS procedures is not novel for the biometrics team. However, for the most part, the statistical outputs produced using SAS ODS and SAS procedures are static and only give end users the potential to look at the outputs without giving them the potential to communicate/interact with them[19]. The goal of achieving TLF automation, however, remains obscure. Common constraints to the wide-ranging adoption of solutions exist and include:

- lack of flexibility
- complexity of use
- the need to learn a new syntax in some instances
- difficulty in validation
- resistance among parts of the programmer community to automation[21]

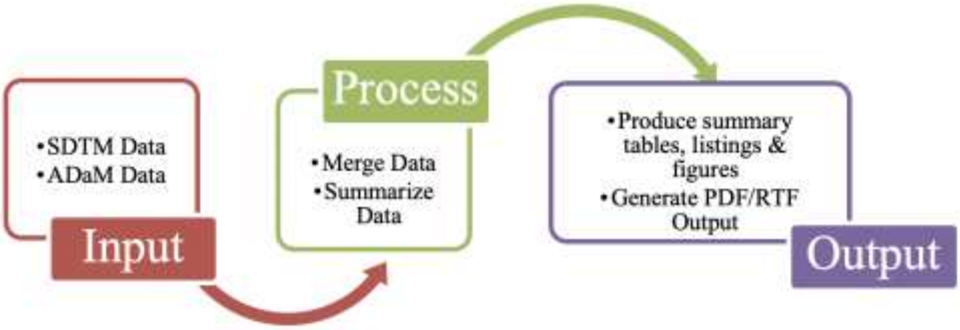


**Figure 1: The output programs sequence for generating summary tables, lists and graphs/figures without any programs[20]**

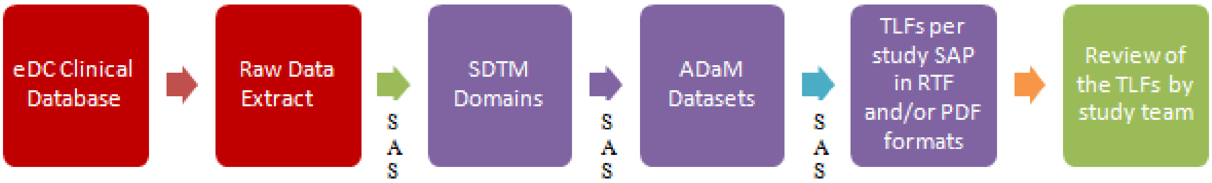

eDC Clinical Database
Raw Data Extract
SAS
SDTM Domains
SAS
ADaM Datasets
SAS
TLFs per study SAP in RTF and/or PDF formats
Review of the TLFs by study team

**Figure 2: A typical data flow in which all TLFs are generated using SAS[20]**

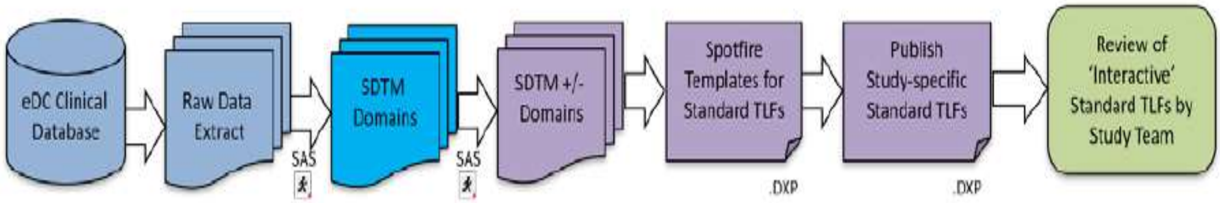

eDC Clinical Database
Raw Data Extract
SAS
SDTM Domains
SAS
SDTM +/- Domains
Spotfire Templates for Standard TLFs
.DXP
Publish Study-specific Standard TLFs
.DXP
Review of 'Interactive' Standard TLFs by Study Team

**Figure 3: Flow of data using Spotfire to generate standard TLFs for interactive analysis[19]**

### Process

In order to generate summary tables, lists and graphs/figures without any programs, the method currently in use must be analyzed and the limitations should be established. The output programs currently follow the sequence (Figure 1)[20]. A typical data flow in which all TLFs are generated using SAS and the statistical outputs are in RTF and/or PDF formats (static TLFs) (Figure 2) [20].

Other analytical tools such as TIBCO spotfire (Figure 3) can also be used to provide the research teams with the ability to see' their data more interactively, identify trends, visualize patient profiles and review outcomes at a high level, while also being able to drill down to get a full picture. The intention is to not challenge the process and data flow of the standard procedures, but rather to challenge the status quo of doing things the traditional way and educate stakeholders regarding the technology solutions. This can help the clinical study teams make their Clinical study reports (CSRs) TLF review and monitoring activities faster and better

The biometrics function will still need to generate TLFs using SAS, while regulatory submission purposes and publishing of the TLFs in the CSR[19].

## RESPONSIBILITIES OF BIOSTATISTICIANS

Data management, monitoring quality, safety, and efficacy
Provide statistical and data related leadership across the business
Responsible for product release and performance monitoring.
Oversee data related aspects of DV and product performance testing for regulatory submissions.
Design, analyze and formally report on studies/experiments
Liaise as statistics subject-matter expert in preparing and participating in internal and external audits of the quality management system (QMS) for all data analysis.
Support products throughout their life cycle, from concept, throughout development and launch, to post-market surveillance
Communicate the interpretation of experiments, clinical studies and fitness evaluations to management
Ensure data is compliant with the native protocols, regulations, and standards[9].

## A BIOSTATICIAN SHOULD POSSESS THE FOLLOWING SKILLS

Strong interpreting skills in statistics.
The ability to choose the correct statistical analysis depends on the ability to identify the right types of variables to be analyzed; therefore, know-how on measurement scales is crucial.
Identify missing information, inconsistent data, outliers and unforeseen lack of protocol variability and deviations.
Perform the transformation of data, estimation, confidence intervals, testing of hypothesis, adjusting confidence intervals and significance.
Understanding of ISO 13485, CE and FDA requirements, along with skills in supporting regulatory submissions and regulatory statistical issues.
Examine data patterns within and across sites in terms of consist ency, range and data variability.
The development of mock tables, list of tables (LoT) and Statisti cal Analysis Plan (SAP).
Awareness of the regulatory protocols of India, China and South America would be an advantage.
Experience in monitoring of test strip systems and in the release of product batch.
Familiarity with different methodologies, drug indications, experimental design and analyses and clinical studies to promote development of new products.
Guide and validate TLGs for programmers.
Compare various treatments: Intent to treat review (ITT), several primary variables (e.g., Dunnett's, Bonferroni Corrections, closed test procedures, and single primary treatment comparison); treatment through centre interaction, dose response analysis, and magnitude effect.
Evaluation of significant or systemic data collection mistakes and on-site and across site reports. Report issues about data integrity or the possible misuse of data.
Evaluate characteristics of sites and performance metrics.
Create a statistical analysis plan: thorough and technical elaboration of the key features set out in the protocol.
Interim review of the comparison of treatment arms with respect to efficacy or safety.
Proficiency in basic programming for statistical analysis (SAS, R, AMOS, Eviews, Strata, etc.).
Respond to regulatory requests and inquiries[9].

**Table 2: Advantages and Disadvantages of Statistical Data**

| Advantages | Disadvantages |
|---|---|
| • Data feasibility for large size samples.<br>• Patterns and correlations projects data visualization.<br>• Possibility of multiple interpretations for different variables<br>• Draw reasonable and accurate data inferences<br>• Prevents numerous errors and biases | • Discarding unfavorable data.<br>• Loaded questions and over generalization.<br>• Misleading graphs or misunderstanding of estimated error.<br>• Confusing statistical significance with practical significance.<br>• Data manipulation and ambiguous averages. |

Biostatistics has become critical to seek and understand better, for not only the clinical trials currently under development but also for various other applications, however there are advantages and disadvantages with statistical data (Table 2).

### Biostatistics Application in various fields of Clinical Research

Clinical research uses biostatistical approaches to formally account for variability sources in the response of patients to treatment. It also helps researchers to draw fair and reliable conclusions from knowledge obtained in periods of uncertainty in order to make exceptional decisions. The biostatistics has applications in various fields such as clinical trials, population genetics, epidemiology, and system biology, in which data can be collected and further be analyzed, presented and interpreted[4].

### Biostatistics in Evidence-Based Drug Development and Clinical Practice

Biostatistical analysis is fundamental to current clinical research and one of the foundations of evidence-based clinical practice. Previous research results are analyzed and applied explicitly to new research. In all phases of the drug development process, state of the art statistical approaches now plays a pivotal role. With fewer than 10% of new compounds reaching the market, the need for advanced biostatistics is rising every day. That is because it decreases deadlines, decreases costs and reduces risks by improving the quality of submissions[4].

### Biostatistics prevent fraud in Clinical Research

In recent years, there have been several allegations of fraud in clinical trials. It includes cheating on the inclusion criterion for the disqualified individuals to join the trial, as per the recorded instances. In addition, the requisite information for such incorrect inclusions is made not missing by fabricating the data. Fraud also occurs due to the manipulation of the data or fabrication of the data values or falsification, which changes data values. In clinical research, biostatistics often avoid fraud or accidental errors[4].

## SOFTWARES FOR BIOSTATISTICS

Statistical computing is now very conceivable due to the availability of computers and relevant programming languages. Now a days, computers and laptops are often used to perform various statistical tests, as it is very cumbersome to perform them manually. Commonly used softwares are SAS, SPSS, NCSS, MINITAB, STATA, Instant, MS Office Excel, Graph Pad Prism, Dataplot, Sigmastat, Graph Pad Instat, Sysstat, Genstat and Sigma Graph Pad. Free statistical software websites are also available. Statistical methods are required to draw a rational inference from the data[6].

## CONCLUSION

Throughout the research, including data analysts, mathematical programmers and medical writers, the biostatistician works closely with the remainder of the biometrics team and management. The biostatistician will assist with CRF creation and dataset requirements with regard to data management. Methodological biostatisticians, collaborating with statistical programmers, ensure that data formatting is accurate and choose data to be pooled. The biostatistician will define research endpoints, sample size estimation, interim analysis preparation and the hypothesis and testing procedures in the Statistical Analysis Plan (SAP). Perhaps the biostatistician's most well-known responsibility is the concept of sample size, which includes multiple variables influencing the studies size, schedules, and budget requirements. Biostatistics provide a base for evaluation, assessment and improvement of efficiency and quality by disentangling data and draw valid inferences. It helps in taking data driven decisions to deliver effective evidence-based research.

## REFERENCES

1. Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? Statistics in medicine 1987; 6:3-10.
2. Rao KV. Biostatistics: A manual of statistical methods for use in health, nutrition and anthropology. 2nd ed. New Delhi: Jaypee Brothers Medical Publisher (P) ltd; 2007.
3. Learn About Being a Biostatistician [Internet]. [cited 2020 Dec 9]. Available from: https://www.indeed.com/career-advice/careers/what-does-a-biostatistician-do
4. The Importance and Role of Biostatistics in Clinical Research [Internet]. [cited 2020 Oct 18]. Available from: https://pubrica.com/academy/2019/08/17/the-importance-and-role-of-biostatistics-in-clinical-research/
5. The importance of Statistics in Scientific Research and Development [Internet]. [cited 2020 Oct 15]. Available from: https://www.studypug.com/blog/the-importance-of-statistics-in-scientific-research-and-development/
6. Dakhale GN, Hiware SK, Shinde AT, Mahatme MS. Basic biostatistics for post-graduate students. Indian Journal of Pharmacology 2012; 44:435-42.
7. Nanivadekar AS, Kannappan AR. Statistics for clinicians. Introduction. The Journal of the Association of Physicians of India 1990; 38:853-6.
8. Bland JM, Altman DG. Transforming data. British Medical Journal 1996; 312:770.
9. Role of biostatisticians in clinical trials. [Internet]. [cited 2020 Oct 18]. Available from: https://www.pepgra.com/role-of-biostatisticians/
10. Formulating hypotheses. [Internet]. [cited 2020 Dec 9]. Available from: https://www.slideshare.net/aniket0013/formulating-hypotheses
11. Nanivadekar AS, Kannappan AR. Statistics for clinicians. Introduction. The Journal of the Association of Physicians of India 1990; 38:853-6.
12. Ghosh MN. Fundamentals of experimental pharmacology. 3rd ed. Kolkata: Bose Printing House; 2005.
13. Mahajan BK. Sample variability and significance. Methods in biostatistics. 7th ed. New Delhi; Jaypee Brothers Medical Publisher (P) ltd; 2010.
14. Smith RJ. Bryant RG. Metal substitutions incarbonic anhydrase: a halide ion probe study. Biochemical and biophysical research communications 1975; 66:1281-6.
15. Ludbrook J. Analysis of 2x2 tables of frequencies: Matching test to experimental design. International journal of epidemiology 2008; 37:1430-5.
16. Training Statistical Programmers on SAP Review Skills. [Internet]. [cited 2020 Oct 18]. Available from: https://www.yumpu.com/en/document/read/35311418/training-statistical-programmers-on-sap-review-skills-phuse-wiki
17. Medhi B, Prakash A. Biostatistics in pharmacology. Practical Manual of Experimental and Clinical Pharmacology. 1st ed. New Delhi: Jaypee Brothers Medical Publisher (P) ltd; 2010.
18. Ludbrook J. Statistics in Biomedical Laboratory and Clinical Science: Applications, Issues and Pitfalls. Medical principles and practice: international journal of the Kuwait University, Health Science Centre 2008; 17:1-13.
19. Bhavin B. PharmaSUG 2019: Paper AD-326. Interactive TLFs- A Smarter Way to Review your Statistical Outputs [Internet]. [cited 2020 Oct 18]. Available from:

https://www.pharmasug.org/proceedings/2019/AD/PharmaSUG-2019-AD-326.pdf

20. Shafi Ch, Zobair MD. Paper TT09. Generating tables, listings and figures without any programs. [Internet]. [cited 2020 Oct 18]. Available from: https://www.lexjansen.com/phuse-us/2018/tt/TT09.pdf
21. Lyon Neil. TS08. PhUSE 2015. The Automated Metadata-driven Table Generation Process (TFLGen) at Amgen Inc. [Internet]. [cited 2020 Oct 18]. Available from: https://www.lexjansen.com/phuse/2015/ts/TS08.pdf

**Cite this article as:**

Aneesha Chatla *et al*. Biostatistics in clinical research: A review. Int. Res. J. Pharm. 2021;12(1): 1-9. http://dx.doi.org/10.7897/2230-8407.1201109

Source of support: Nil, Conflict of interest: None Declared

Disclaimer: IRJP is solely owned by Moksha Publishing House - A non-profit publishing house, dedicated to publishing quality research, while every effort has been taken to verify the accuracy of the content published in our Journal. IRJP cannot accept any responsibility or liability for the site content and articles published. The views expressed in articles by our contributing authors are not necessarily those of IRJP editor or editorial board members.