A Brief Description of Basic Components of A SAP

Cheng Peng

Contents

1 Introduction

Regardless of the research design, statistics are a crucial component of research since it allows the researchers to summarize the collected data and give it to others for interpretation.

We need a defined analytic plan before we start collecting data. The SAP (statistical analysis plan) will direct us from the beginning to the conclusion, help us summarize and describe the data, and test our hypotheses.

The statistical analysis plan (SAP) describes the intended clinical trial analysis. The SAP is a technical document that describes the statistical methods of research analysis, as opposed to the protocol, which represents the analysis.

2 What is a statistical analysis plan?

Statistical Analysis Plan (SAP) is a detailed document specifying how data will be analyzed, ensuring transparency, reproducibility, and minimization of bias.

- **Data Collection**
  - **Primary Data**: Clearly define the main outcome variables (e.g., clinical endpoints, survey responses).
  - **Secondary Data**: Specify additional variables (e.g., covariates, confounders) and their sources.
  - **Data Handling**: Procedures for missing data, outliers, and transformations (e.g., log-transformation for skewed data).
- **Methods of Analysis**
  - **Descriptive Statistics**: Summarize data (means, proportions, SDs, visualizations).
  - **Inferential Statistics**: Hypothesis testing, confidence intervals, effect sizes.
  - **Software/Tools**: Specify (e.g., R, SAS, SPSS) and versions.
- **Primary Analysis**
  - Define the primary outcome(s) and statistical test(s) (e.g., t-test, ANOVA, regression).
  - State how the primary hypothesis will be tested (e.g., superiority, non-inferiority).

- **Comparisons & Significance Levels**
  - Predefined Comparisons: Subgroup analyses, pairwise comparisons (adjust for multiple testing if needed).
  - Alpha (Significance Level): Typically 0.05, with justification if adjusted.
  - **Power Analysis**: Sample size justification based on expected effect size.
- **Exploratory Data Analyses (EDA)**
  - Unplanned analyses to identify patterns (e.g., trends, interactions).
  - Clarify that results are hypothesis-generating (not confirmatory).
- **Statistical Models**
  - **Primary Model**: Specify (e.g., linear regression, logistic regression, Cox model).
  - **Response Variable**: Clearly defined (e.g., blood pressure, survival time).
  - **Predictors/Covariates**: List included variables and rationale.
  - **Alternative Models**: Robustness checks (e.g., sensitivity analyses, different covariate adjustments).

3 Identifying the need for an SAP

While a study protocol outlines the general research methodology, an SAP provides a deeper, technical specification of statistical procedures. Here's when an SAP becomes essential:

- **High-Risk or Regulated Studies**
  - **Clinical trials** (especially Phase II/III) requiring regulatory approval (FDA, EMA).
  - **Studies with major clinical/public health implications** (e.g., drug efficacy, policy decisions).
  - **Pre-specified analyses to prevent bias** (e.g., avoiding selective reporting).
- **Complex Statistical Methods**
  - **Advanced modeling** (e.g., mixed-effects models, survival analysis, machine learning).
  - **Adaptive trial designs** (interim analyses, Bayesian methods).
  - **Handling missing data** (multiple imputation, inverse probability weighting).
- **Large or Multi-Center Studies**
  - Ensures consistency across sites/analysts.
  - Prevents post-hoc decisions that could introduce bias.
- **Reproducibility & Transparency**
  - Needed for peer-reviewed journals (e.g., ICMJE, CONSORT requirements).
  - Allows an independent statistician to replicate analyses without ambiguity.

4 Study Protocol vs SAP

The study protocol and SAP are complementary documents that guide different aspects of a research study. While they overlap in some areas, they serve distinct purposes and audiences.

The following table gives some comparisons of the two formal documents:

The protocol and SAP are interdependent: The protocol sets the rules; the SAP enforces them statistically. Here are some examples:

- **SAP Expands on the Protocol**
  - **The protocol** states: "We will compare Group A and Group B using regression analysis."
  - **The SAP** specifies: "A Cox proportional hazards model will be used with covariates X, Y, Z. Hazard ratios will be reported with 95% CIs. Missing data will be handled via multiple imputation."
- **SAP is More Rigid (Pre-Specified)**
  - The protocol may allow minor methodological flexibility.
  - The SAP locks in analysis details before data unblinding to prevent bias.
- **SAP is Often a Separate Document**
  - For simple studies, statistical methods may be fully described in the protocol.

    – For complex/high-stakes studies, a standalone SAP is required (e.g., clinical trials for FDA submission).

5 Key Information in the SAP

The SAP should contain various sample size calculations for different statistical procedures to achieve certain statistical power and a thorough explanation of the main and any interim analyses used in the data analysis technique.

The SAP should also thoroughly explain the procedures used to analyze and display the study results.

- **Statistical Significance** – The predefined level of statistical significance (e.g., $\alpha = 0.05$) and whether one-tailed or two-tailed tests will be employed.

- **Missing Data Handling** – Methods for addressing missing data (e.g., imputation techniques, complete-case analysis).

- **Outlier Management** – Approaches for identifying and handling outliers.

- **Estimation Methods** – Techniques for point and interval estimation.

- **Composite/Derived Variables** – Rules for calculating composite or derived variables, including data-driven definitions, with sufficient detail to minimize ambiguity.

- **Baseline and Covariate Data** – How baseline and covariate data will be incorporated into the analysis.

- **Randomization Factors** – Inclusion of randomization factors (if applicable).

- **Multi-sources Data Handling** – Methods for managing data from multiple sources.

- **Multiple Comparisons & Subgroup Analysis** – Methods for adjusting for multiple comparisons and conducting subgroup analyses.

- **Interim/Sequential Analyses** – Details of any planned interim or sequential analyses.

- **Software Specifications** – Identification of the computer systems and statistical software packages used for data analysis.

- **Assumptions & Sensitivity Analyses** – Critical assumptions of the statistical models and methods for conducting sensitivity analyses to validate these assumptions.

- **Data Presentation** – Guidelines for tables and figures to present study data.

- **Safety Population Definition** – A clear definition of the safety population.

- **Model Validation & Alternatives** – Provisions for testing the statistical model and alternative methods if model assumptions are violated.

The SAP must include provisions for testing the statistical model, along with alternative methods to be used if the model assumptions are not met.