PharmaSUG China 2015 - Paper 11

Why Statistical Analysis Plan (SAP) should be comprehensive?

Riddhi Merchant, PPD (I) Pvt. Ltd., Bengaluru, India
Ranjith Prayankotveettil, PPD (I) Pvt. Ltd., Bengaluru, India

ABSTRACT

A Statistical Analysis Plan (SAP) is one of the key Regulatory documents which comprises of detailed Statistical methods and techniques used for data analysis. It serves to inform readers about clinical study design, the execution of statistical methods/techniques, data analysis and reporting.

But do we think, "Are the Statistical Analysis Plan always comprehensive enough to proceed with the Data Analysis?" As a Statistician do we write SAP, keeping in mind all the statistical methods, in depth that best correspond to the research hypothesis?

This paper will discuss study statistician's responsibilities for preparation of SAP in which he should describe detailed procedures for executing the statistical analysis of the primary and secondary variables which is required in conveying statistical results obtained from stated statistical methods used.

INTRODUCTION

A statistical analysis plan (SAP) describes the planned analysis for a clinical trial. SAPs must be carefully prepared by statistician working on the project for clarity and comprehension in order to construct analysis data sets and prepare planned Tables, Figures, and Listings. SAP writing is a challenging task that requires both statistical expertise and abstract thinking skills, and is often completed in the absence of available trial data.

The SAP brings the team together on the same page. For biostatisticians, all inferential analyses and statistical methods are to be delineated. For programmers, the SAP provides explicit guidance on SAS® programming and the presentation of the tables, figures, and listings. For the sponsor, the SAP ensures that they get the outputs in the way they need for the clinical trial report. The SAP adds another layer of specificity to the clinical trial.

The SAP describes the planned statistical analysis of a clinical study as outlined in the protocol. In contrast to the protocol which outlines the analysis the SAP is a technical document which describes the statistical techniques for study analysis in detail. The SAP defines all the statistical output which will be included in the clinical study report. Mock shells of tables, listings and figures are usually attached to the SAP although they should not be formally part of the SAP. The SAP and the annotated CRF are the documents which are most often used by statistical programmers to create their deliverables. In general there are four different types of analysis plan in the clinical development of a compound:

- **Statistical analysis plan for a clinical study** – describes the planned statistical analysis of a study
- **Interim statistical analysis plan** – describes the planned statistical analysis of an interim analysis for a study and therefore needs to address handling of partial unblinding issues in case of blinded studies. It also describes possible impact on the conduct and the complete final analysis like the possible adjustment of significance levels.
- **Data Monitoring Committee (DMC) statistical analysis plan** – modification of interim analysis used for DMCs and describes regular (e.g. monthly) data monitoring procedures for safety or efficacy questions. The DMC SAP also contains the DMC charter which clarifies exactly the names and responsibilities of the involved parties.
- **Integrated statistical analysis plan** – describes the planned analysis for an integrated analysis which is used for example in submissions. It defines the details of programming output for an ISS and ISE usually in one document.

While writing SAP we generally follow ICH E3 and E9 guidelines. This gives us an idea of the body content of individual sections of SAP. But E3 and E9 does not specify about specific statistical techniques.

This paper is an assortment of lessons learnt by biostatistics professionals while writing SAP. The examples cited below are in general describing some statistical methodologies which vary depending on study design, study phase or therapeutic area.

Below mentioned are some sections of SAP which a biostatistician should take care of while writing SAP.

THE EXPERIMENTAL DESIGN:

The experimental design should be explained very clearly in SAP because if we do not mention it in detail then it becomes difficult for statistical programmers and reviewers to develop and validate the datasets and TLFs. It is very important for any individual who is involved in data analysis that he/she is having few basic ideas about

study design on their fingertips, e.g. definition of baseline value, missing value imputation for baseline value or any other time-point etc.

Sometimes, we may have more than one observation as pre-dose value or before randomization value. In such a case, it is important to mention the definition of baseline value so that statistical programmer can understand the requirement and can derive the baseline values as required.

E.g. suppose we are having three pre-dose values of heart rate based on ECG report and we need to derive the baseline values based on these three values. There may be a chance of considering the mean or median value of these three observations as baseline value. Moreover, we can have one separate screening value. Now if any of the three observations is missing then we can also use the screening value as imputed value and can calculate the baseline value. This should be specified clearly in the SAP.

Now one may have a quick question of why do we need to impute the baseline values? The baseline value plays an important role in the primary or secondary analysis.

Suppose we want to compare the test and reference drug affecting heart rate. As we know this comparison can be done through the two sample t-test, but it will be better to do this comparison on change from baseline values. In this case, if we are not having baseline value then we will miss some of the patients due to missing change value. In addition to this if we are doing some statistical modelling on the primary outcome and we need to include the baseline covariate in this model then missing value of baseline covariate can result in exclusion of few observations.

For example, if we are trying to apply PROC LOGISTIC on the primary outcome of subject will have stroke or not (0 = No and 1 = Yes). For this analysis, we are capturing few affecting factors like; Age, Gender, Disease History, heart rate, blood pressure, left ventricular mass and Aspirin use. In our study we were having 35% missing data for any of the baseline covariates then ultimately we are losing 35% of our actual data due to incompleteness of the data. This is a huge amount of data and can lead to completely change the study results.

FAILURE TO SPECIFY DATA TRANSFORMATION / STATISTICAL METHOD IN SOFTWARE

I would like to present one of the common things that we statistician generally miss, while writing SAP for the analysis of time to event data. This is required for more information and for obtaining accurate statistical results.

For time to event analysis, while testing the hypothesis whether the survival distribution functions for two treatments are equal or not, we often use Kaplan-Meier analysis with log rank test and 95% CI over median survival time. Here, we generally specify in the SAP, “For the time-to-event endpoint; the distribution for each arm will be estimated by the Kaplan-Meier method and will be compared by the Log-rank test. The median time along with the 95% confidence limits will be presented for each treatment arm. Simultaneous confidence bounds for the Kaplan-Meier curve will be computed for both the treatment groups.”

Moreover, for the ease of the programmer, we will mention the SAS® code for this as;

```
PROC LIFETEST DATA = <DATASET NAME>;
      TIME TIME*CENSORING VARIABLE;
      STRATA <>;
RUN;
```

In the above paragraph, we have not mentioned anything about the transformation method used for the calculation of 95% confidence interval of median survival time. We know there are different types of transformation methods available for this, e.g. log, liner, log-log, arcsine-square root, logit etc.

But sometimes, we don’t know, by default which statistical method is used in a SAS® procedure for analysis. For example, SAS® V9.1.3 – uses linear function for calculation of CI for median survival time, whereas SAS® V9.2 uses LOGLOG function for calculation of CI for median survival time.

Now even though, if we have mentioned the SAS® code in SAP, we have not specified anything about transformation type so if any SAS® programmer, uses this code in SAS® V9.1.3, he/she will get the 95% CI based on the Liner transformation because this version of SAS® uses Linear transformation by default. But if the same code will be used in SAS® V9.2, the 95% CI will be different. This difference is due to the transformation type used in version 9.2 which is Log-Log by default. Transformation (e.g. log, linear, log-log etc.) should be specified in the SAP and a rationale provided especially for the primary variable(s).

Hence, this leads to the inference that mentioning the transformation type in the SAP is essential.

DEFINING CENSORING RULES IN DETAIL FOR TIME-TO-EVENT ANALYSIS

In clinical domain, we often come to a situation where we need to analyze the time-to-event data. In such type of analysis our point of interest is occurrence of specified event at a particular time. But, there will always be few subjects who does not experience the specified event throughout the study or few subjects who were lost to

follow-up or terminated the study before the end of the study. Such subjects who do not experience the specified event are called censored subjects. In general, every time to event analysis always defines the censoring rule based on which the subject will be classified as censored subject or subject experiencing event. But sometimes it becomes difficult to determine whether the subject experienced the event or not. Moreover, the statistician should define the censoring rules in consultation with clinical experts.

E.g., suppose, in an oncology trial analysis our event of interest is progression free survival of a subject. Here, we may consider multiple conditions to censor the subject as mentioned below:

1. If there is documented Progressive Disease (PD) during the study
2. If death during the study before PD
3. If discontinued due to PD, but no documented PD
4. If no baseline assessments
5. If treatment discontinuation for other than PD or death, and no post-baseline disease assessment
6. Treatment discontinuation for other than PD or death with post-baseline disease assessments
7. If new anti-cancer treatment started prior to disease progression
8. If death or PD after 1 or more missed disease assessments
9. Subjects still on treatment without PD as of data cut-off

Often we miss to detail out these multiple rules in a simple language and so for programmers it becomes difficult to create the censoring variable for primary efficacy analysis of PFS survival.

MODEL ASSUMPTIONS TO BE SPECIFIED IN SAP

There are always some or the other issues faced by programmer or reviewer while working on efficacy tables where mostly model selection is involved.

Let's understand this with an example. Recently I came across to a situation while reviewing the analysis report that the results provided through a model which was specified in SAP were not appropriate since the model was not running successful. In SAP, the model was specified with dependent variable as subject had stroke (Yes/no) and the independent variables as BMI, blood pressure, history of cardiovascular disease and physical activity (mild, moderate, access). This model was designed prior having any idea of data. But when it was run on actual data, it was not running properly.

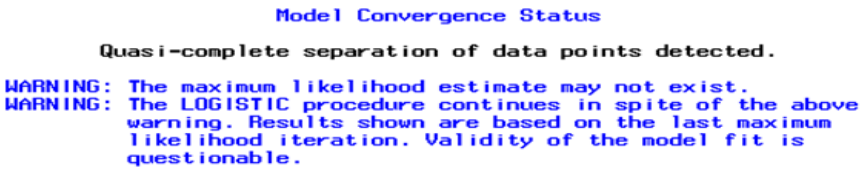
Based on this example, we generally notice that while writing about logistic or proportional Hazard model in SAP, we mostly specify the perfect model in SAP prior having any knowledge about data.

Instead of writing the perfect model in SAP prior to having any idea about data, we should specify all the factors or covariates which are clinically affecting the dependent variable. Then we can specify about a model selection methods, like; backward, forward, score and stepwise, and significance level to be considered for selection of factors and covariates.

For example, suppose we are doing a study in which our outcome variable is vital status (dead or alive). Based on the experience of clinical experts, we can initially identify eleven covariates; Age, Sex, Heart Rate (HR), BMI, History of cardiovascular disease (CVD), Atrial Fibrillation (AFB), Cardiogenic Shock (SHO), Congestive heart complication (CHF), Complete heart block (AV3), MI order (MIORD) and MI type (MITYPE), which can affect the model outcome variable. Here, we can first specify the model selection method as FORWARD or BACKWARD along with the significance level at which we will decide whether to include or exclude any variable from the model. Once we get the statistically significant variables to be included in the model, we can discuss the model with clinical experts to have their suggestions on model. It is always necessary for a statistician to remember that no variable can be excluded from the model just because of the statistically non-significance; we must consider the clinical importance or significance as well. Hence, after having discussion with clinical team, we should decide the model which is actually running well on the data and which is statistically as well as clinically important. But when we used the selection above mentioned methods in PROC LOGISTIC, we found that out of these eleven factors only 5 factors are statistically significantly related to outcome variable which were AGE, SHO, HR, MITYPE and AV3.

Now in consultation with clinical team we can identify if any other clinically significant factor we should include in the model which we have removed based on statistical observation.

Now let us consider the second case of same procedure, i.e. PROC LOGISTIC. Quite often we come across to a situation where we have a warning message in LOG file in SAS® as shown below:



We all know that in general such a case arises when the sample size is inadequate for the number of explanatory variables or the event of interest is rare to occur. But there may be a case when the sample size is large enough and still such a situation arises. Moreover, when such a separation is there at least one parameter estimate is "infinite" (i.e., maximum likelihood estimate does not exist). Based on the large standard error, we can also identify which variable is having separation in the observations.

Of course we cannot predict anything about this separation in our SAP prior having any idea about data, that any particular variable will have this separation. But we can specify in our SAP about the methods; like drop the particular variable which is having separation in observations or instead of LOGISTIC method, use CMH test or use Firth's Penalized Likelihood, to handle this kind of situations if they arise.

It is difficult to identify those variables which need to be dropped from the model, since we don't know which variables will have the separation prior looking at the data. Moreover, it may possible that more than one explanatory variable will be having separation of data. Hence, we can not specify the method of dropping any variable from the model.

Also, we cannot keep CMH as an alternative to LOGISTIC regression in such data separation issue because of several reasons. First, while performing CMH test, we can keep only one explanatory variable other than treatment variable, and hence we cannot draw some inference about other explanatory variables at a time. Moreover, it will add the corrections to the zero cell counts while calculating odds ratio estimates but cross-tabulation a row or column of zero will not be included in computation.

But we can include the Firth's Penalized Method by adding a keyword "FIRTH" in the model statement in PROC LOGISTIC. This method splits each original observation into a "response" and "non-response" so issue of separation gets eliminated. Moreover, eradicating the problem of data separation, this method produces less biased estimates [1] than the conventional maximum likelihood approach. Additionally, it allows investigation of the effect of the explanatory variable that data is separated about. Thus, in such cases where separation of data occurs, the penalized likelihood function becomes very asymmetric.

Hence, while analyzing the binary outcome variable with PROC LOGISTIC, we can give detail description about model selection and data separation so that our model can be selected as per clinical and statistical judgment and the data separation issue can be handled effectively without any loss of data.

BASIC STATISTICAL METHODS

In routine life, as a statistician, many times we are using some basic statistical tests so often that if we are asked any question at any time regarding those tests, we can answer them. But as a non-statistician, have we ever thought whether these simple (for those statistician who use them quite often) tests are really very simple for other readers as well?

For example, suppose we want to compare two insulin to check how they are behaving in controlling the fasting sugar level in two different samples. Here, subjects are randomly given insulin A or insulin B and we are comparing the change from baseline in both the samples. To test the hypothesis H_0: $\mu a = \mu b$. For this comparison, we generally write in our SAP as two sample t-test will be used for comparison of mean and a two-sided p-value will be reported. But we generally miss out to mention anything for equality of variance test. As a SAS programmer, many of our colleagues are not from statistical background and they are not aware of which p-value to be reported in table; whether "Pooled" one or the "Satterthwaite". To make the programming part easy for programmers, we just need to mention about one more p-value in our SAP. The p-value of equality of variance test, based on which we should decide to choose "pooled" or "Satterthwaite" method. If the equality of variance test is non-significant, i.e. p-value > 0.05 then we will use p-value calculated using "Pooled" method or if equality of variance test is significant, i.e. p-value < 0.05 then we will use the p-value calculated through "Satterthwaite" method.

Similarly, we generally do not mention many things about chi-square test and fisher's exact test to differentiate between the two tests. Moreover, while analyzing frequency data through Chi-square test in SAS®, we get four types of p-values. For a SAS programmer it becomes difficult to identify which p-value should be presented in the analysis report. If we can mention about this in SAP, then the programmer can easily prepare the reports without seeking any further help.

WRITING ABOUT CONVERGENCE CRITERIA:

Unlike LINEAR regression or LOGISTIC regression, the computations for longitudinal and hierarchical data models can be problematic. Sometimes the algorithms used to calculate estimates do not converge. Like logistic regression estimation, the algorithm is iterative. The iteration history tells us whether the algorithm converged and how it behaved along the way to convergence.

It is important to check that PROC MIXED says "Convergence criteria met." If this doesn't happen, then you cannot depend on the results of your analysis.

The following paragraph/points will help you out if a particular criteria is not satisfied then which other criteria are supposed to be considered.

Generally statisticians give only one method to converge the data. Programmers are not sure, how to proceed if it fails. For example, statisticians provide option of TYPE=UN (Unstructured) in PROC MIXED syntax, but programmers are clueless about what all are the other options they can try if this model fails and what is the order to follow which should specify in SAP. Please see below different options which programmers can try for TYPE based on the requirement. These options can be used which can be used if for a particular model, the Convergence criteria is not met.

Structure	Description
ANTE(1)	Ante-dependence
AR(1)	Autoregressive(1)
ARH(1)	Heterogeneous AR(1)
ARMA(1,1)	ARMA(1,1)
CS	Compound Symmetry
CSH	Heterogeneous CS
FA(q)	Factor Analytic
FA0(q)	No Diagonal FA
FA1(q)	Equal Diagonal FA
HF	Huynh-Feldt
LIN(q)	General Linear
TOEP	Toeplitz
TOEP(q)	Banded Toeplitz
TOEPH	Heterogeneous TOEP
TOEPH(q)	Banded Hetero TOEP
UN	Unstructured
UN(q)	Banded
UNR	Unstructured Corrs
UNR(q)	Banded Correlations
UN@AR(1)	Direct Product AR(1)
UN@CS	Direct Product CS
UN@UN	Direct Product UN
VC	Variance Components

CONCLUSION

We have tried to touch base on statistician's approach while writing SAP which requires well understanding of the study along with the deep knowledge of various statistical methodologies. A detailed SAP helps programmers to understand the requirements of the study prior starting with programming activities. However, it is difficult to prepare a perfect SAP without having any prior knowledge of data, this paper is just a mere effort to have comprehensive SAP for understanding of the process to complete the analysis task more efficiently.

REFERENCES

[1] Heinze, G., Schemper, M. (2002) *A solution to the problem of separation in logistic regression*, Statist. Med., 21, 2049-2419

[2] http://magazine.amstat.org/blog/2013/04/01/writingtipsfunding2013/

[3] SaschaAhrweiler, (2011) Review of Statistical Analysis Plan: http://www.phusewiki.org/wiki/index.php?title=Review_of_Statistical_Analysis_Plans

[4] SAS® Help for PROC LIFETEST: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_lifetest_sect004.htm

[5] International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use 'E9: Statistical Principles for Clinical Trials', (ICH E9) http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf.

ACKNOWLEDGEMENTS

We would to thank PPD for providing us this opportunity to present this paper. We would also like to show our gratitude to all the biostatisticians from the industry for sharing their pearls of wisdom with us during the course of this research.
I (Riddhi Merchant) would also like to especially thank my husband- Mr. Vimal M. Dave, who provided insight and expertise that greatly assisted me in writing this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Name: Riddhi Merchant and Ranjith Prayankotveettil
Enterprise: Pharmaceutical Product Development Pvt. Ltd
Address: PPD, 9th Floor, Valance Block, Prestige Technology Park 3, Outer Ring Road, Marathahalli
City, State, ZIP: Bengaluru, Karnataka, India – 560 103
Work Phone: +91 (0) 8015188639
E-mail: Riddhi.merchant@ppdi.com and Ranjith. Prayankotveettil@ppdi.com