PharmaSUG 2021 - Paper SS-171

A Lead Programmer's Guide to a Successful Submission

Pranav Soanker, Covance by Labcorp
Santhosh Shivakavi, SCL IT Technologies

ABSTRACT

A successful NDA (or other submissions) depends upon a robust Biostatistics/Biometrics department which usually consist of few statisticians and many statistical programmers among others. It requires great co-ordination with all the stake holders (CDM, External vendors, Medical Writing, CROs etc.,) and great understanding of core competencies of each group to achieve a high-quality submission ready package. The programmers usually have a wide and diverse educational background, with overwhelming majority without a statistical degree, and particularly with junior team members the clinical trial knowledge and experience can be limited. Thus, the role of a Lead Programmer is very critical as a liaison between programmers and statistician. In this this paper we walk through in detail some of the checks a Lead Programmer should do as part of Senior Review (for Raw Data, Protocol/SAP, SDTM & ADaM data, TLFs, Define package, ADRG & SDRG) in addition to the CDISC compliance checks and independent validation. This significantly enhances the quality of the submission and instills great confidence to the statistician and minimizes rework prior to a submission.

INTRODUCTION

While roles may vary in various organizations as who is responsible for the Senior Review (or High-Level Review) of the final package, often it mainly rests with Lead Programmers who are usually involved in end-end programming activities. A Lead Programmer needs to quickly develop/incorporate best standard practices and encourage the team to religiously follow the spirits of First Time Right (Six Sigma concepts to reduce waste).

This paper will visit the scope of High level review a lead should perform to achieve a high-quality output after validation of all items for the deliverable.

This paper is divided into five Sections.

Section 1: Review of Raw Data, CRF & Other Data Metrics

Section 2: Review of SDTM data

Section 3: Review of ADaM data

Section 4: Review of TLF & CSR

Section 5: Review of Submission Packages (Define, ADRG & SDRG)

SECTION 1: REVIEW OF RAW DATA, CRF & OTHER DATA METRICS

Review of RAW Data:

The Lead should verify if the raw data extracted is accurate and has to check with external vendors for any vendor data issues. Review the eCRF specs and create 1:1 mapping of Raw Data to SDTM domains which establishes traceability.

A standard DCO (Data cut off) Metrics document which establishes the DCO date must be developed to ensure the statistician agrees with the data to be included in the interim delivery.

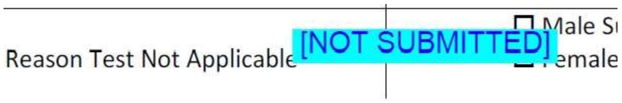
In early data snapshots of a study, A lead should review the efficacy raw data to check for normality, for example use proc univariate Q-Q plots (NORMAL option) to check the data distribution. Proc univariate can also to be used on other safety data to document the outliers and extreme observations (five lowest and five highest). Discuss these findings with statistician to model the data and re test the normality in Analysis level again to ensure the data is normal.

```
proc univariate data = eff_source;
 var eff_var3;
  qqplot eff_var3 / NORMAL;
run;
```

CRF Review:

Programmers face challenges when CRF design has issues, so a proactive CRF review during development is essential for a lead. Example: CRF Visits page – Visit Date inconsistencies etc.,

For data that is Not Submitted a lead should double check with Data Management to confirm such pages are indeed not collected from the start of the trial and should be adequately annotated consistently across all such pages/sections as “NOT SUBMITTED”.



Reason Test Not Applicable [NOT SUBMITTED]

□ Male S

□ Female

[Figure 1: CRF section]

SECTION 2: REVIEW OF SDTM DATA

Raw to SDTM traceability and Mapping:

1. Verify the Controlled Terminology version to be used and ensure all metadata is mapped in SDTM specifications document (define.xml compliant) based on actual raw data values.
2. Code & Decode values: Many times, sponsors maintain lookup tables for code lists and any changes in these standards may affect a previous study CRF. Prepare standard checks so that all Code/Decode are consistent (Perform Proc Freq for the coded values of Raw Data against SDTM & ADaM).
3. Protocol Amendments – Check and track any new pages that are added/modified in CRF and also if SDTM specifications have been updated to account for all CRF modifications and programmers have been alerted for the changes and ensure such changes are incorporated. Need to make sure eCRF specifications is consistent and all the raw datasets (including any that are new) have been accounted.
4. For Multiple Screening subjects, considerations must be evaluated (for example previous Subject ID is to be saved in SUPPDM).

5. Subject Status Tracker – Databases like RAVE provides the stages of each subject (Subject Visit, End of treatment (EOT), Screen Failures (SF), Discontinued subjects etc.,). It will be helpful for lead to prepare a similar tracker from SDTM data and compare against the RAW data which will confirm the raw data is accurately mapped to SDTM at pre-Database (pre-DB) lock.
6. Special ASCII characters in data impacts the dataset programming. These should be processed to replace the special character symbol with its closest appropriate characters. Also any invisible characters should be processed to ensure data is machine readable.
7. Numbers of observations in each SDTM dataset must be assessed against the raw/source datasets to ensure all the data is adequately captured in SDTM.

Verification of SDTM data based on SDTM IG & FDA Technical Rejection Criteria for Study Data:

1. Pinnacle 21 outputs needs to be checked for any issues and also ensure appropriate version of SDTM IG, Controlled Terminology, Dictionary version is used. P21 reports must be annotated for any unresolved issues.
2. Check if any TAUG (Therapeutic Area User Guide) is available for the study indication and follow the guidelines (example Ophthalmology OE domain, Viral studies VR, PF etc.,) so that the therapeutic relevant fields are mapped accordingly.
3. Trial Design domains (TDM) ideally should be created at the start of the study and any protocol amendments need to be incorporated immediately as they are updated. Any changes in Inclusion/Exclusion should be updated to ensure TITEST/TITESTCD are unique based on TIVERSION.
4. A lead programmer should ensure the correct Trial Summary Dataset is included which is one of the most common reason for FDA rejection (refer to FDA Rejection criteria example) and ensure the Study Start Date (SSD) is present for each study.

SDTM consistency checks:

Cross consistency checks to be performed on SDTM data to find any possible data or programming issues.

1. Visits out of order example Visit 1 date is after Visit 2 etc.
2. Record count checks between source data and SDTM data based on key variables.
3. Any free text comments length split based on unique words versus limiting the split variables to 200 length should be evaluated.

SECTION 3: REVIEW OF ADAM DATA

1. After DB lock merging of actual treatments to various datasets needs to be checked 100% accurately.
2. Ensure ADaM data is traceable to SDTM and for derived parameters/variables or derived records indicate in specs the source(s) and all the additional traceability variables to be used (DTYPE and/or PARAMTYP) and traceability variables (SRCDOM, SRCSEQ, SRCVAR).
3. Lead should create a flow chart of ADaM datasets based on data dependency and/or derivation of any composite end points for example ADSL should be the first dataset to be programmed followed by other datasets. Check the batch run files to ensure order is appropriately followed.

4. Ensure all SAP specific guidance related to visit windowing, multiple maseline(s) handling, missing data or incomplete data are included in ADaM specifications document and are implemented in the dataset programming.

5. Pinnacle 21 outputs needs to be checked for any issues and also ensure appropriate version of ADaM IG, Controlled Terminology, Dictionary version is used. P21 reports must be annotated for any unresolved issues.

6. For Lab bi-directional tests make sure toxicity grading has been handled accurately and when applicable use the appropriate directionality variables (ATOXGRL, ATOXGRH, BTOXGRL, BTOXGRH, SHIFT1, SHIFT2, ATOXDSCL, ATOXDSCH).

7. For Analysis that require visit windowing, lead can create an excel spreadsheet with all the possible visits based on SAP analysis visits. Similarly, also ensure visit windowing variables are used AWTARGET, AWTDIFF, AWLO, AWHI, AWU.

8. Key efficacy endpoint variables or values must be rederived from raw to ensure the robustness of data. Example OS, PFS etc., should be derived from RAW data and compared against the TFL outputs. Consistent values reflect great quality of the ADaM and SDTM datasets. Any missing values must be verified against the raw to ensure all data points are adequately captured.

SECTION 4: TLF & CSR REVIEW

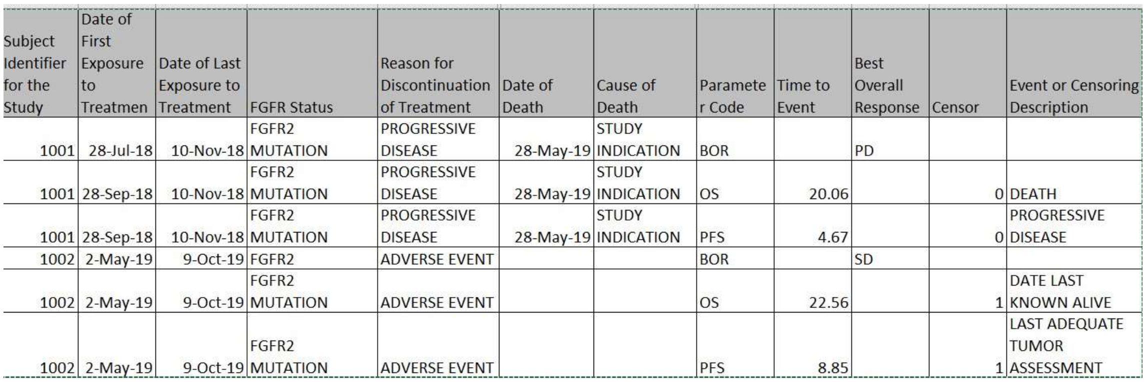
1. Lead should ensure the team follows the General Conventions for TLFs provided by the sponsor. Also, should verify the time stamp of all the datasets are prior to the outputs generation, example SDTM should be run prior to ADAM (including QC programs), Ensure all logs are clean and all datasets or TLFs are completely validated and visually inspected for correctness.

2. Lead should inspect carefully the final package and verify if the title matches the titles noted on table of contents and in the shells. Also check if the data presented in TLFs reflects what the title indicates (including any population subset or any subgroup analysis) and ensure formatting is consistent across all the outputs.

3. Lead Programmer should verify all subjects are represented in TLFs (based on different sets of population). Ensure that Big N's or denominators include all patients in the sample population. Verify consistency between the Big N's for the requested sample/population are consistent across all TLFs.

4. Reconcile with frequency counts (or small n's) on disposition table/baseline characteristics table to ensure the supporting data presented in multiple reports are consistent.

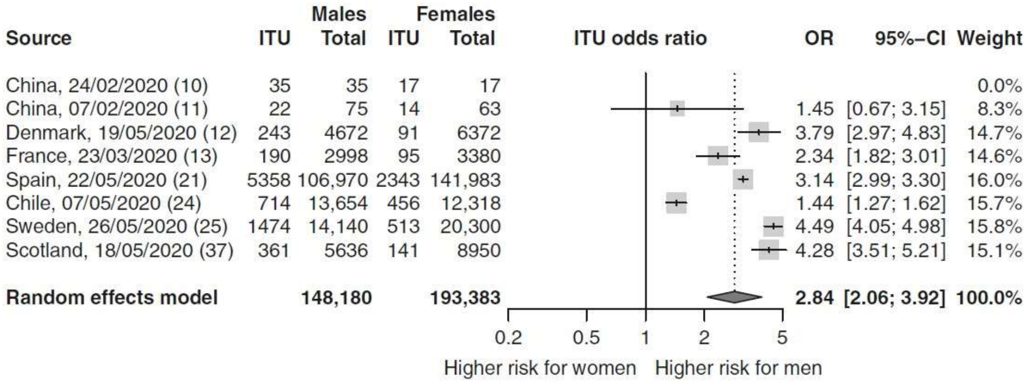
5. It is a good practice for Lead Programmers to derive as many variables of interest as possible and save in spreadsheets (Figure 2) for handy reference. This can be used to compare against the TLF content.



Subject Identifier for the Study	Date of First Exposure to Treatmen	Date of Last Exposure to Treatment	FGFR Status	Reason for Discontinuation of Treatment	Date of Death	Cause of Death	Parameter Code	Time to Event	Best Overall Response	Censor	Event or Censoring Description
1001	28-Jul-18	10-Nov-18	FGFR2 MUTATION	PROGRESSIVE DISEASE	28-May-19	STUDY INDICATION	BOR		PD		
1001	28-Sep-18	10-Nov-18	FGFR2 MUTATION	PROGRESSIVE DISEASE	28-May-19	STUDY INDICATION	OS	20.06		0	DEATH
1001	28-Sep-18	10-Nov-18	FGFR2 MUTATION	PROGRESSIVE DISEASE	28-May-19	STUDY INDICATION	PFS	4.67		0	PROGRESSIVE DISEASE
1002	2-May-19	9-Oct-19	FGFR2	ADVERSE EVENT			BOR		SD		
1002	2-May-19	9-Oct-19	FGFR2 MUTATION	ADVERSE EVENT			OS	22.56		1	DATE LAST KNOWN ALIVE
1002	2-May-19	9-Oct-19	FGFR2 MUTATION	ADVERSE EVENT			PFS	8.85		1	LAST ADEQUATE TUMOR ASSESSMENT

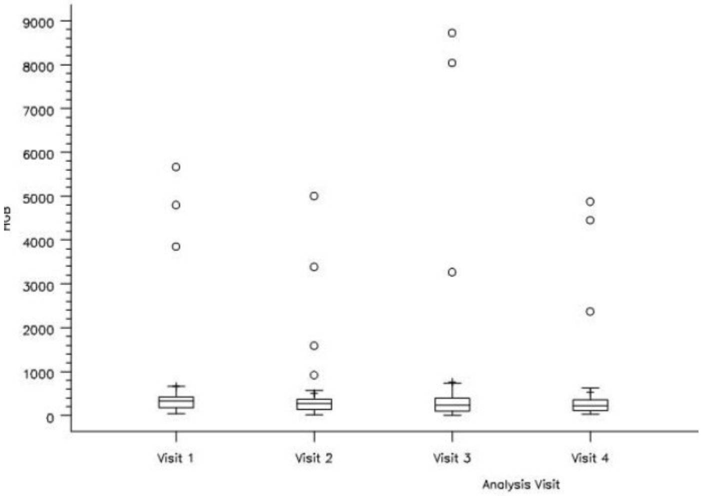
[Figure 2: Variables of interest for cross compare]

6. Investigate Listings with “No Data to display” and Tables with all 0 values to ensure that it is not symptomatic of an erroneous programming logic or inaccurate data source.
7. Verify the percentages reported in TLFs add to 100%. If percentages reportedly are greater than 100% for example in cases of Relative Dose Intensity value being >100% ensure that the reason for discrepancy is noted (example: Subject taking extra cycle/dose etc.,).
8. Ensure programming team has adequate understanding of efficacy data and for the tables that include Hazard’s ratio or Odd’s ratio check to inspect if numeric value of treatment variable is changed based on reference treatment (example Placebo=0 and Study Drug = 1).
9. For pooled analysis of multiple studies make sure that the coding dictionaries are consistent, variable attributes and population flags derivation are consistent based on integrated SAP. Programming lead should verify the usage/algorithm of ANLxxFL variables in individual studies and ensure it is consistent.
10. Ensure to eliminate the possibilities of unintended unblinding by using appropriate communicating policies. Example include in subject line **** UNBLINDED INFORMATION ****.
11. Inspect the figures against the Tables and Listings to observe the trends and any outliers. For instance, in a recent study of global data on COVID[1] the data suggests while there is no difference in the proportion of COVID-19 cases between sexes, men have a higher risk of ITU (Intensive Treatment Unit) admission (OR 2.84) and death (OR 1.39)[1]. In the Figure 3, the overall section (diamond sign) crosses the line of no effect which graphically illustrates the striking variance among sexes. To confirm these observations the Lead can verify other data points (ex: Death/Hospitalization outputs) and cross check to see if the variance is due to inaccurate data source or any issue with programming logic.



[Figure 3: Forest Plot of global data and risk of ITU/Death] [1]

12. In the Figure 4, the Box plot is compressed due to the extreme outliers, is not an ideal way of presentation. Lead should discuss with statistician or sponsor to clip some of the extreme observations for a better readable graph.



[Figure 4: Box plots with extreme outliers] [2]

SECTION 5: SUBMISSION PACKAGES (DEFINE, ADRG & SDRG)

Define Document:

1. Most of the common errors that might occur in Define creation are due to metadata inconsistencies, missing study specific metadata or derivations and comments that are not meaningful or sufficient.
2. A Lead programmer should review the Define comments/derivations sections to ensure no programming code is used. Also need to proof read it to see if there are any typos or mistakes and also check to ensure consistency is maintained across the metadata. Need to check if all the hyperlinks are working correctly and pointing out to the right file or place. Need to check the XSL style sheet if it is working and can be also be able to open in Web browser.
3. CRF page number references are most common errors made and the lead should check if all the CRF page numbers mentioned in origin points to right page from CRF. Origins should be checked for all variables to see they are correct, for example a derived variable cannot be having origin as CRF/missing.
4. Detailed computational methods needed to be given so that it makes reviewers understand the derivation. Lead needs to check the derivation and ensure it is complete and meaningful.
5. Check for any missing study specific code lists for variables for example --CAT, --SCAT etc.
6. Check for any Merged code lists when individual code lists need to be given for each unique variable in domain and check if the code list contains all the possible values as per CRF rather than only those present in data. Free text field can have missing code lists.
7. Check if the key variables have been defined properly and the structure of the domain is consistent with the key variables.

SDRG:

1. Ensure latest standard template has been used and naming convention for clinical studies should be csdrg.pdf.
2. Ensure right version of Pinnacle 21 is used and list the appropriate version in the SDRG.
3. Ensure explanations given in the conformance issues section are sufficient and meaningful.
4. Check the content of the document if it has consistent font and size, all hyperlinks are working correctly, consistent formats across tables, and if any ASCII characters present in the text.
5. Content of SUPPQUAL domain must be described.
6. Check for any study specific details have been given properly including any Note to file sections or assumptions in mapping need to be explained properly. Need to explain any specific TAUG has been used and explain if any important aspects of the TAUG needs to be considered.

ADRG:

1. Ensure latest standard template has been used and naming convention for clinical studies should be adrg.pdf.
2. Ensure Split Datasets, Data Dependencies are described and ensure all P21 unresolved issues are included in Issue Summary section.
3. Check if the common information between SDRG and ADRG are consistent. Ex: Versions of MEDDRA, WHODD etc.
4. Ensure right version of Pinnacle 21 is used and list the appropriate version in the ADRG.
5. Ensure important information from source documents such as Protocol, SAP etc., should be reported in ADRG so that reviewers can have single place to review from analysis datasets perspective.
6. Ensure that adrg.pdf has been placed in appropriate section as per eCTD module 5.
7. Check if all the conformance issues have been reported and has given meaningful justification for any false positive messages as per P21 reports.
8. Ensure Links to the Analysis datasets and TFL programs have been given and all the links are working properly. Also check the order of the TFL is consistent as per TOC.
9. Ensure all the programs are using standard Header as per company SOP and make sure the programs are following GCP guidelines and no log issues and include programs as .txt files.

CONCLUSION

Throughout the programming process (from project Kickoff to the deliverable submission) the team faces many challenges. The only way out is to navigate around them in the first place by anticipating them. An experienced Lead is a vital cog of the team who can proactively have many checks in place so that the team follows through a well-guarded barrier. Also, with well-established processes, standards and with usage of standard programs, Macros or Formats a lead can minimize the errors or inconsistencies prior to statistical review.

[Figure 5: Road Image with well-guarded barrier[3]]

REFERENCES

1. Peckham, H., de Gruijter, N.M., Raine, C. *et al.* Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission. *Nat Commun* **11,** 6317 (2020). https://doi.org/10.1038/s41467-020-19741-6

2. Mary Rose Sibayan, Thea Arianna Valerio. Clip Extreme Values for a More Readable Box Plot. *PharmaSUG China 2016* Paper 72

   https://www.lexjansen.com/pharmasug-cn/2016/DV/PharmaSUG-China-2016-DV08.pdf

3. Image source https://interestingengineering.com/this-new-korean-rolling-barrier-system-could-save-millions-of-lives

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Pranav Soanker
Covance by Labcorp
pranav.soanker@covance.com

Santhosh Shivakavi
SCL IT Technologies, UK
shivakavi.santhosh@gmail.com