PharmaSUG 2017 - Paper DA07

Data Integrity: One step before SDTM

Pavan Kathula, Sonal Torawane, Ephicacy Lifescience Analytics

ABSTRACT

In clinical research, errors occur despite careful study design, conduct, and implementation of error-prevention strategies. Data cleaning intends to identify and correct these errors or at least to minimize their impact on study results. Little guidance is currently available in the peer-reviewed process on how to set up and carry out cleaning efforts in an efficient and ethical way. With the growing importance of Good Clinical Practice guidelines and regulations, data cleaning and other aspects of data handling will emerge from being mainly subjects to being the focus of comparative methodological studies and process evaluations.

We would like to present an overall summary of the scattered information, integrated into a conceptual framework aimed at assisting investigators with planning and implementation. Our paper will explain the suggestions on using unique specifications, processes and specific methods along with SAS® to maintain data integrity and cleanliness. With this suggestion, the scientific reports might describe data-cleaning methods, error types and rates, error deletion and correction rates. Utilization of simple tips and techniques along with the proper documentation will impact not only the study results but also time, effort and cost. However, this paper will not throw light on the SDTM techniques, but SDTM terminologies / tools are used for explanations wherever needed.

INTRODUCTION

Data integrity is maintaining and assuring the accuracy and consistency of data. Have you ever come across any issue between “Informed consent date” and “Visit date” after SDTM dataset production? The answer might be **yes** in some cases. Is this a significant issue? **No**, but identifying the same at the initial stage helps in saving efforts or time. It labels too simple but ends big.

We have tools to perform SDTM mapping. We’re having automated systems to run, check on the data issues. And organizations started investing on the automated tools to get the data in the desired format (e.g.: SDTM). It is good that the technology providing ample opportunities to save time and produce quality output as per the available standards. Technology too has its limitations and the chance of missing out data issues is high, if software tools are deployed for the purpose. This is the area where human intervention through programming adds value and specifically helps in resolving data issues.

Hence, handling data in a right way results in better outcome, meaning spending time on reviewing the raw datasets before SDTM dataset production will help in finding and reporting data issues at the initial stage of programming.

Figure 1. Basic idea of implementing the plan.

One step before SDTM – includes:

- Creating a specific process
- Specification development

OUTLINE FOR "ONE STEP BEFORE SDTM" APPROACH:

- LEVEL 1: Pre-ONE STEP
- LEVEL 2: ONE STEP BEFORE SDTM
- LEVEL 3: Post-ONE STEP

LEVEL 1: PRE-ONE STEP

In this step, the focus is on reviewing the data at a basic level using elementary programming techniques and filtering methods. This is part of the good programming practice probably not followed religiously in the industry. This programming/filtering concepts aims at figuring instances such as – does the data extract comprise all the datasets? Are there any missing records? Data review is further dissected in the steps below.

DATA REVIEW

The following steps can be followed to perform data review:

1. Understanding the data: Have a detailed look of the data and check all the variables in the data extract. Treating each variable as they all are equally important. One of such examples is described below.

SUBJECT NUMBER	MEDICATION START DATE	CHARACTER MEDICATION START DATE
72601	03/01/2016	03/01/2016 NUL:NUL:NUL
72601	03/01/2016	03/01/2016 NUL:NUL:NUL
72601		UNK/UNK/2000 NUL:NUL:NUL
72601		UNK/UNK/1994 NUL:NUL:NUL
72601		UNK/UNK/1994 NUL:NUL:NUL
72601		UNK/UNK/1994 NUL:NUL:NUL

Figure 2: Snapshot of example on understanding and checking the data thoroughly.

This dataset contains the variables SUBJECT, MEDICATION START DATE in two different formats, do not forget to include all the related variables. Here by extracting the known values from "CHARACTER MEDICATION START DATE" will result in filling the "MEDICATION START DATE" with a value. Missing records can be reduced by considering all the related variables.

SUBJECT NUMBER	SYSTOLIC BLOOD PRESSURE	DIASTOLIC BLOOD PRESSURE
10101	128	85
10102	130	80
10103	140	90
10104	90	120
10105	120	80

Figure 2a: Snapshot of example on data issue.

This dataset contains the variables SUBJECT, SYSTOLIC and DIASTOLIC BLOOD PRESSURE and the record #4 is one of the examples for data issues.

2. Simple data check: Performing checks on the data.

SUBJECT NUMBER	DATE OF FIRST DOSE	TIME OF FIRST DOSE	DATE OF BIRTH	TIME OF BIRTH
178	29JAN1989	14:00	29JAN1989	1:20
179	30JAN1989	13:15	29JAN1989	19:00
180	30JAN1989	12:30	30JAN1989	4:10
183	01FEB1989	12:30	01FEB1989	8:30
378	12APR1990		12APR1990	9:16
722	24SEP1991	9:20	29SEP1991	3:18

Figure 3: Snapshot of example on data issue.

This dataset contains the variables SUBJECT, FIRST DOSE DATE and TIME, BIRTH DATE and TIME. There was an issue with the last record whereas, the Date of Birth is after the Date of First Dose. Writing a simple programming check with respect to dates results in identifying the issues at the early stage.

3. Finding the potential data issues: Initial glance of the data might not aid us find all the issues, instead a simple SAS® code on each dataset for finding the potential data issues will be very helpful.

```
DATA _null_;
  LENGTH allvars $1000;
  RETAIN allvars;
  SET sashelp.vcolumn end=eof;
  WHERE libname = "RAWDATA" & memname = "Dataset Name";
  allvars=trim(left(allvars))||','||left(name);
  allvars1=substr(allvars, 2);
  IF eof THEN call symput('varlist', allvars1);
RUN;
DATA mis_val;
  SET "Mention desired dataset name to find the missing values for variables";
  chk=cmiss(&varlist.);
RUN;
```

The above simple SAS® program provides variables from any dataset of a specific library into one variable. One could run this program on any raw data library and will find the number of variables with missing values. The program performs a check on the data variables and identifies missing values, which could further lead us to the possible probability of discrepancy – erroneous data collection/entry or issue with the sample collected.

LEVEL 2: ONE STEP BEFORE SDTM

IS THERE ANY SPECIFIC DOCUMENTATION AVAILABLE?

Apart from the current tools, processes & open resources, several CROs, service based organizations and sponsor companies are following their own set of standards and documentation to clean the data with different processes like edit check specifications, offline listings, early stage data analysis and data review listings etc.

The various systems followed by Clinical Data management for capturing, managing and reporting clinical data are Medidata RaveX®, Medrio, Oracle Clinical, ClinPlus® and Open Clinica. These are the systems

having the features of efficient data collection, inbuilt edit check process and data validation to ensure cleanliness. When the data comes to SAS® programming to report, based on the design structure of those sources and format of data, it's often to notice issues or challenges like missing values, special characters in free text variables, inconsistency in the date values etc.

Although, these tools or systems having their own set of edit checks, they are not much effective when it comes to different data standards, comparison between more than one forms.

Patient profiles and narratives are the other aspects where the industry is following to clean data and to find data anomalies. They are used for the medical review such as why the patient took a concomitant medication, why the adverse event ended in a serious adverse event and to verify all the dosing records for an individual subject. This analysis can be carried out at any stage of the study as per the requirement by sponsor or the investigator. Most of the times it will be conducted post ADaM or along with the CSR generation and sometimes at the initial stages.

JReview® is a Web-based data review tool which is another approach for cleaning data and providing patient profiles. The clinical trial teams, data managers and safety monitors use it to identify data issues, data integrity, site performance, safety monitoring and efficacy responses. However, it has some limitations. For example, no calculations and derived variables can be made. It can't merge multiple datasets.

E.g. If we want to compare adverse events (AE), medical history (MH) and concomitant medications (CM) to find any inconsistences, JReview® can't do this directly.

We can overcome such situations with the help of SAS® programming at the initial stages of the study.

With this paper, we intend to suggest a process with specific documentation using SAS® suiting the needs of all different organizations at the early stage for providing better stand for SDTM.

WHY DO WE HAVE A PROCESS?

The delivery of the clinical study program is the expensive, time-consuming component of the drug development process. Delivering the clinical study program successfully is getting harder as the operating environment has become more complex. There is increased competition for and continued geographical spread of clinical study sites; more complicated clinical study protocols; increased regulatory agency expectations; stringent post marketing commitments and continuing adoption and adaption of industry standards and best practices. If we have a standard process at the early stage (data level) one can achieve quality results and can save lot of time, cost and effort.

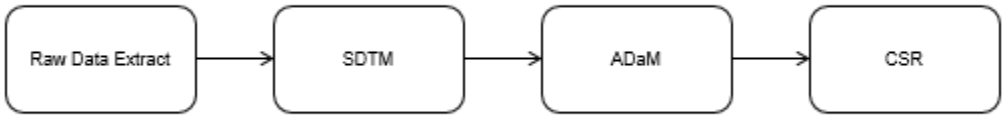

Figure 4: Traditional way of processing the data.

PROCESS THE DATA

We consider data from the extract and will create output listings in the excel format (section: Output format) with filter provision on various aspects based on the study design. We also perform the common checks on the safety data like Adverse Events, Demographics and Disposition etc. The common checks include finding the missing values for the topic variables, chronological order of the visit dates etc.

Following is the step-by-step process:

1. Create a specification to provide specific listings as per the study design.
2. Have the programs ready for standard domains with common data issues to be checked. Here are few common examples:
    - Check for the partial or missing dates.
    - End date should not be less than start date.
    - Visit dates should be always incremental.
    - Topic variables should not be missing.
    - Ensure diastolic blood pressure is always lesser than the systolic blood pressure.
    - Ensure “Other, specify” is filled for all values of “Other”.
3. Check for the data issues which are specifically related to study objectives.
    - E.g.: Action taken from Adverse events related to dose change.
    - Adverse event date is always after informed consent date and before last participation date
    - Concomitant medications cross reference with either Adverse events or Medical history.
4. Report the data issues to concerned team in the specific format mentioned.

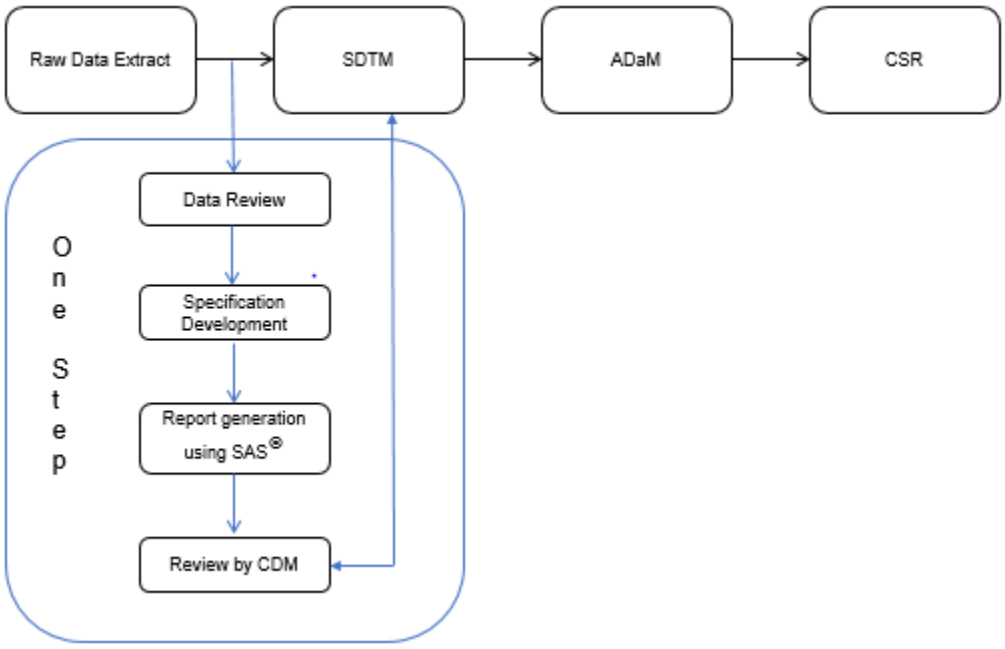

Figure 5: ONE-STEP approach.

SPECIFICATION DEVELOPMENT

It is important to spend time developing the specification in detail to ensure:

- consistency on accuracy, data quality to reduce the noise, time and cost. This is also a fair process for clinical industry to ensure they are following the right process without deviating from the guidelines. With this paper, we would like to suggest documentation to maintain the data quality and to ease further process of analyzing the data. This provides consistency, accuracy and quality in the process.

The specification includes unique way of creating detailed document for data checks. Following are some important points to remember while creating this document:

1. Provide a unique listing number & title for tracking purpose.
2. Mention the CRF form name involved in creating the specific listing.
3. Ensure the required visits are included to the list.
4. Ensure list of variables to be reported in the final listing (e.g.: AE.ACNTKN, EX.DOSE).
5. Explain the logic to create the specific listing. (e.g.: Report all the Adverse Events which received Concomitant Medications CMSTDTC <= AESTDTC <= CMENDTC).

Listing Number	Title of listing	Form name	Visit Number	List of items	Logic / Restriction	Comments
Provide the serial number	Provide the title as per the performing check	Provide the form name from the annotated CRF	List down the visit numbers where the check to be performed	List all the variables from contributing datasets to perform the respective check	Mention the rules, restrictions, conditions and logic to apply	Space for reviewer to provide his / her comments

Table 1. Sample Specification Format

The above table briefly describes the specification which is a primary document for the programmers. It provides information regarding the checks to be performed with exact logic to be applied depending on the study design and the structure. We recommend following double programming to perform this task and even specification to be reviewed by two programmers which helps in preventing issues in developing the specifications as per the CRF and the Study data.

The specification document can be included with various checks as follows:

- Screen failure subjects with missing CRF pages
- Subjects with Adverse events / SAE's cross-reference with study drug administration
- Subjects with overlapping or duplicate AE/CM records
- Subjects deviating from Primary objective of the study
- Subjects who had study drug interruption for longer than two weeks
- LAB / VITAL SIGNS with changes from baseline to highest or lowest values on trial
- List of Deaths

PROGRAMMING ASPECTS:

It is important to setup a proactive and systematic approach to start with any task / program. While the initial setup is in place then the programmer's task would be easy to proceed further. One of the basic and very important programming aspect is to have double programming approach. The approach not only helps in providing the quality output but also helps you in finding the potential data issues as we have mentioned earlier.

OUTPUT FORMAT:

Report the issues in the following excel format with filtering provision to subset the data. This will provide ease of access for Data Management colleagues to track and address their resolutions or clarifications in a separate column named as 'comments'. With the excel format one can easily visualize the data and can clarify any updates by dropping a note or comment. It helps in describing the errors or issues meaningfully. The format is designed in a way that it represents the data subject-wise, which will be easy to review and find the issues. The efficient way we report the data in a clear format results in better understanding and analyzing the same.

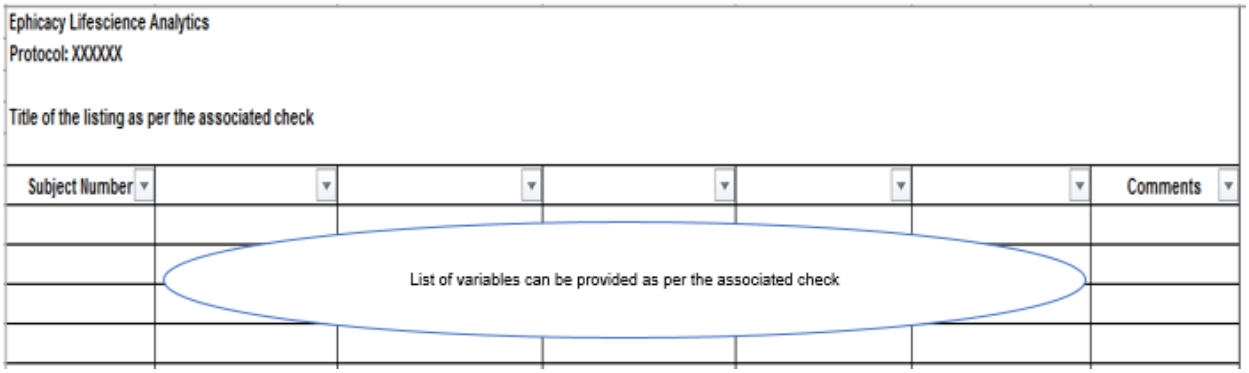

Figure 6: Snapshot of output format.

The above output resembles the patient profile, but the format in the excel that we are providing can be checked at the early stage before database lock and can be provided to the CDM to make the changes to data if required. This will ensure providing a clean data for the SDTM programmers.

LEVEL 3: POST-ONE STEP

Levels 1 and 2 describes the process. Level 3 talks about the advantages of the same.

- Simplicity: The process is simple and easy to implement and replicate.
- Awareness about common issues: Increased awareness about erroneous data could help in resolving issues at a faster rate.
- Traceability and accessibility: Traceability of flawed data and controlled accessibility is another advantage.

The positive aspects of 'ONE STEP before SDTM' approach can be studied using the below mentioned cases:

Case 1: SDTM programming will be easier.

Case 2: Protocol Amendments / Change in study design.

Case 3: Possibility of terminating the study well-in-advance by reviewing this data.

CASE 1: SDTM PROGRAMMING WILL BE EASIER:

SDTM programming has a standard process which includes steps like double programming, running CDISC reports, check on the CDISC issues, report for any data issues, etc. If CDISC report is throwing an error because of data issue, then entire process needs to be repeated which can ends up increasing overall project duration.

E.g.: If calculated age for a subject is negative in SDTM demographics, one of the reason for this is data issue i.e. reference date is less than birth date. In such case CDISC report will through an error after which we need to raise an issue to CDM. This entire process would consume time.

Hence with the presented process we can find data issues at early stage of the study and might reduce the error rate while SDTM programming.

CASE 2: PROTOCOL AMENDMENT / CHANGE IN STUDY DESIGN:

Protocol amendment will happen when there are changes in the existing protocol that significantly affect safety of subjects, scope of the investigation, or scientific quality of the study. Based on the checks performed, we might conclude or expect dose interruption of the study drug. This can be achieved from the listing related to AE and Drug dosing.

E.g.: changes requiring an amendment may include:

- Any increase in drug dosage or duration of exposure of individual subjects to the drug beyond that described in the current protocol, or any significant increase in the number of similar Adverse events.

CASE 3: POSSIBILITY OF TERMINATING THE STUDY:

Trials terminate for a variety of reasons, not all of which reflect failures in the process or an inability to achieve the intended goals. Primary outcome data were reported most often when termination was based on data from the trial.

Hence to have a check on data is a necessity. With the help of suggested process, we can check number of serious adverse events or number of death cases at the early stage of study to avoid more fatal outcomes.

CONCLUSION

The process of data cleaning proposed in this paper will help to have good control on data and entire study work at early stage. SAS® programming also takes a good part in the process of data cleaning. A well implemented data cleaning process needs planning and a good start. By following the process presented, hassle free SDTM programming resulting in good study conclusions would be the outcome.

The way forward could be automation; automating the process by developing macros which could read data from the SAS® files and directly point the discrepancy in data. A standard library of programs ready to be tweaked based on the study design could also improvise the process.

Data integrity being the key in this entire game of drug development, there should be continuous & conscious efforts to ensure subject data safety and integrity, thus working towards a heathier world.

REFERENCES

A Strategy for Managing Data Integrity Using SAS®, Brett J. Peterson

http://www2.sas.com/proceedings/sugi31/103-31.pdf

Website: Data Integrity, Anonymous author, Available at

https://en.wikipedia.org/wiki/Data_integrity

Website: SAS, Multiple authors, Available at

http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a003121095.htm

Why clinical trials are terminated, Multiple authors

http://biorxiv.org/content/biorxiv/early/2015/07/02/021543.full.pdf

Clinical operations comes of age, Available at

http://www.kinapse.com/media/1107/clinical-operations-comes-of-age.pdf

ACKNOWLEDGMENTS

We would like to acknowledge our managers and organization, without their support we would not have been here.

CONTACT INFORMATION

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

Name: Pavan Kathula
Enterprise: Ephicacy Lifescience Analytics
Work Phone: +91 80 41463195 / 96
E-mail: pavan.kathula@ephicacy.in
Web: www.ephicacy.com

Name: Sonal Torawane
Enterprise: Ephicacy Lifescience Analytics
Work Phone: +91 80 41463195 / 96
E-mail: sonal.torawane@ephicacy.in
Web: www.ephicacy.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.