International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

SAS Programming in Clinical Research for DrugDevelopment

Subrahmanya Rajeev Dhulipala
Student
Dept of Electronics and Communication Engineering
R.V. College of Engineering, Benguluru
dsubrahmanyar.ec19@rvce.edu.in

V Dhanush
Student
Dept of Electronics and Communication Engineering
R.V. College of Engineering, Benguluru
vdhanush.ec19@rvce.edu.in

Dr. Geetha K.S.
Professor & Vice-Principal
Dept of Electronics and Communication Engineering
R.V. College of Engineering, Benguluru
geethaks@rvce.edu.in

***Abstract*—In this paper, a study data tabulation in the demographics domain from the vast amount of clinical trial data in accordance to the SDTM Specification sheet provided by CDISC has been generated. SDTM specifies a standard format for study data tabulations that must be submitted to a regulatory body like the Food & Drug Administration (FDA) for drug review. Statistical Analysis Software (SAS) programming was used to extract demographic data from a clinical trial that was double-blinded and randomly assigned. It involves creating datasets for the demographics domain using SAS programming in-line with the SDTM specification. Tables, Lists, and Figures (TLFs) using SAS programming in relation to the study variables were also created. The following data can be used for analyzing a drug with a potential of diabetes prevention. It is used for determining the safety & dosage range, the toxicology and side effects of the investigational drug. These generated datasets& TLFs can beused as a part of drug review for the regulatory agencies.**

***Index Terms*—Study Data Tabulation Model (SDTM), Clinical Data Interchange Standards Consortium (CDISC), Statistical Analysis Software (SAS), Tables Lists and Figures (TLFs), Food & Drug Administration (FDA) .**

I. Introduction

A worldwide, nonprofit group called CDISC (Clinical Data Interchange Standards Consortium) creates and supports data standards for clinical research. CDISC was founded in 1997 with the goal of improving the effectiveness of medical research through the development of standards that allow for the exchange, sharing, and utilization of clinical data.

Data collection, representation, and submission are only a few of the clinical research-related topics that are covered by the standards that CDISC creates and upholds[1].[8]These requirements cover many different types of information, including patient demographics, adverse events, test findings, medical history, and more. The CDISC standards are intended to be platform-independent, vendor-neutral, and compliant with current regulatory requirements [6].

The primary CDISC standards include:

1) Clinical Data Acquisition Standards Harmonization, (CDASH): is a set of standards that ensures consistency and high-quality clinical trial data gathering.
2) Study Data Tabulation Model (SDTM): Clinical trialdata can be organized and presented using SDTM's standardized framework. It specifies particular variables and their forms, allowing for uniform data representation and exchange between studies and systems.
3) Analysis Data Model (ADaM): Standards for the analysis and reporting of clinical trial data are providedby ADaM. It supports statistical analysis and reporting requirements and makes it easier to create analytical datasets.
4) Define-XML:is a standard for specifying the metadata and data structure of clinical trial data. It gives a dataset's content, organization, and relationships be- tween data items in a machine-readable format.

To create and improve these standards, CDISC works with partners from business, academia, regulatory organizations,and technology suppliers.[9] In order to facilitate the adoption and application of CDISC standards, they also provide trainingand tools.

II. Study Data Tabulation Model

The Clinical Data Interchange Standards Consortium (CDISC) created the SDTM, or Study Data Tabulation Model, as a standardized framework for organizing and presenting data in clinical trials. For the uniform representation and interchange of clinical trial data throughout various studies, platforms, and organizations, SDTM offers a structured format. For these data items, it establishes a standardized structure and controlled vocabulary to ensure consistency and comparability across various research and datasets.

Improved data interoperability and easier data analysis and integration in the pharmaceutical business and other healthcare research disciplines are the main objectives of SDTM.[7] Data gathered during clinical trials are represented by a set of

standard domains, variables, and controlled language called SDTM. Irrespective of the particular study or data source, these standards guarantee that data is organized and formatted in a uniform and harmonized manner. Data is arranged by SDTM into various domains that correspond to various facets of a clinical trial. Demographics, Adverse Events, Medical History, Laboratory Results, Vital Signs, and Concomitant Medications are a few of the SDTM domains that are frequently employed. Each domain is made up of preset variables that record pertinent data pertaining to that domain. Examples of possible variables in the Demographics domain are the subject ID, age, sex, and race.

Variable classification under SDTM: Identifier are the variables that identify the study, the subject(s), the domain, and the sequence number. This details the observation's main topic. Topic provides details about observation. [4]Timingdescribes when an observation was made. Qualifier provides supplementary text, values, or outcomes that further explain the observations. Rule this describes the formula or algorithm used to calculate dates, times, or visits. This is mostly utilized in the domain of trial design.

Different forms of clinical trial data are categorized into standardized groupings using SDTM (Study Data Tabulation Model) domains. These domains offer a dependable framework for classifying and displaying particular data kinds gathered throughout a study. A large range of data items frequently found in clinical research are covered by the SDTM domains. The primary SDTM domain classifications are as follows:

1) Interventions: Data from the study's use of investigational products or interventions are gathered in this domain. It contains information about dosage, administration method, intervention length, and timing.
2) Adverse events, medical history, concurrent medications, and protocol deviations are all recorded in the Events domain during the course of the trial. It offers a thorough record of health- and safety-related data.
3) Findings: This area of the study includes several kinds of observations and measurements. It covers areas including electrocardiograms (ECGs), medical gadgets, vital signs, physical examinations, and laboratory test findings.
4) Exposure: Information on subject exposure to research items or other substances is captured in the Exposure domain. It contains information on dosage, exposure time, and any relevant findings.
5) Demographics: This domain contains data on the study participants' ages, sexes, races, ethnicities, and other pertinent factors. It offers a framework for looking at and comprehending the study population.
6) Questionnaires: The data gathered through structured questionnaires or surveys given to research participants falls under the realm of questionnaires. It contains quality of life evaluations, patient-reported outcomes (PROs), and other patient-reported information.

The area of "Trial Design" describes the features and design of the research itself. It contains details about research arms, treatment groups, the randomization process, and the study's goals.

Additional specialized data items that may be unique to particular therapeutic specialties or research types are included in the Special-Purpose domain. Data on immunogenicity, genetics, or specific evaluations of a certain condition or indication are a few examples.

With the use of these domain classifications, clinical trial data may be organized and reported in a consistent manner, assuring consistency and promoting data sharing and analysis between various studies and systems. The SDTM standards define a unique set of variables and data structures for each domain.

III. Tables Lists and Figures

TLF, or Tables, Listings, and Figures, is an important part in reporting and analyzing clinical trials.[9] TLF stands for Tables, Listings, and Figures, which are used to compile, present, and graphically represent the outcomes and conclusions of a clinical research. Numerical data is presented in tables in a tabular manner for simple comparison and analysis. They frequently present important study parameters, demographic data, efficacy results, safety information, and other pertinent trial metrics. Researchers, statisticians, and regulatory bodies can review and evaluate the findings thanks to tables' succinct and organized presentation of the data. Listings, on the other hand, give personal information in a more thorough way. They frequently contain subject-level data such demographics, adverse events, test results, and therapy allocations. Listings are especially helpful for investigating data patterns, spotting anomalies, and doing in-depth studies.[2] Figures are graphical representations of data, such as graphs, charts, and plots, that offer visual insights and support data understanding. They can be used to show comparisons, trends, distributions, and correlations between various variables or treatment groups. Figures are useful for reporting key findings and trends in a clinical study because they may clearly and concisely communicate complicated information.

Tables, Listings, and Figures (TLF) are crucial parts of reporting and analyzing clinical trials.[10] They provide a systematic and visual representation of the trial data, allowing researchers and stakeholders to efficiently analyze, understand, and convey the results. TLFs improve the comprehension of clinical trial findings and promote the use of evidence-based judgement in medical and pharmaceutical research.

Different types of graphs that are used in clinical data analysis using SAS:

1) Histogram: A histogram is a visual that shows how a continuous variable is distributed. The frequencyor number of data points inside predetermined bins or intervals are displayed. Understanding a variable's shape, central tendency, and spread is made easier by histograms.

2) A bar chart uses rectangular bars to display categorical data, with the height of each bar reflecting the frequency or percentage of the data that falls under each category. The distribution of discrete variables can be seen and compared using bar graphs.
3) Line Plot: A line plot, also referred to as a line graph, shows how two continuous variables relate to one another over a continuous x-axis. It displays a variable's trend or change over time or in another continuous dimension. In clinical investigations, line plots are frequently used to visualize time-series or longitudinal data.
4) A scatter plot is a two-dimensional diagram that shows the relationship between two continuous variables as a collection of points. The values of the variables for a certain observation are represented by each point. Scatter plots can help you evaluate correlations, spot outliers, and spot trends in your data.
5) Box Plot: A box plot, also known as a box-and-whisker plot, offers an artistic representation of the distribution of a continuous variable. It shows the data's median, quarterlies, and any outliers. Comparing distributionsand finding potential discrepancies or anomalies between groups or treatments is made easier with the use of box plots.
6) Clinical trials and survival studies frequently use survival graphs, such as Kaplan-Meier plots. They show the likelihood of surviving or the amount of time until an event (like death or the progression of a disease) through time. Censored data are frequently included in survival plots when the important event has not yet happened by the end of the research period.
7) A forest plot is a type of graph that is frequently used in meta-analyses and systematic reviews to compare and summarize the findings of several studies. Along with a broad summary estimate, it shows the point estimates and confidence ranges from several research.

IV. TOOLS USED

SAS Enterprise guide v9.2 is used for creating Study Data Modulation Table for demographics domain .[11] Various graphs such as scatter plot , box plot and line graph were plotted using SAS Enterprise guide v9.2 . SDTM specs were provided for various variables , including description about the variables about their length, format ,type and commentson how to extract the raw data . Raw data files are provided , using SDTM specs the various variables used in Demographics domain are extracted.

Obtain the SDTM Implementation Guide for the study version that applies to you.[5] This manual provides a framework for arranging the data by outlining the common domains, variables, and their descriptions. Recognize the data collection tools used in the study, such as the case report forms (CRFs) or electronic data capture (EDC) platforms. Determine which variables need to be extracted after identifying the data items that were gathered for each domain. In accordance with the SDTM specification, map the collected variables to the relevant SDTM domains and variables. This mapping guarantees that the data is properly transformed and organized for SDTM compliance. Using the mapping, separate the identified variables from the research dataset. To filter and extract the necessary variables, this often entails querying the database or utilizing programming languages like SAS or R.[3] The extracted variables should be formatted according to the applicable SDTM standards. This entails making certain that variable names, labels, and formats adhere to the SDTM specification. If necessary, take into account the controlled terminology (CT) while standardizing coded variables. Verify the accuracy and integrity of the extracted variables by performing data quality tests and validations. Compare the extracted data to the source data to look for any inconsistencies or mistakes. Following the SDTM framework, group the variables into SDTM datasets after they have been extracted and validated. Following the guidelines provided in the SDTM Implementation Guide, each domain should be established as a distinct dataset. Keep records of the extraction procedure, including information on variable mappings, extraction techniques, data transformation stages, and any variations or factors unique to the study. Tablesand graphs are produced using SAS programming after the necessary datasets and variables have been extracted.

V. RESULTS

This section shows the various tables and graphs obtained .Figure 1 shows the Proc means for various treatment with respect to the analysis variable age. In the first table depicts for treatment trt = 0 (DS1234), with respect to analysis variable Age, the total number of observation N = 30 , Mean = 34.6666667 , Standard Deviation = 9.1964511 , Minimum = 20.0000000 and Maximum = 50.0000000.

In the second table depicts for treatment trt = 1 (Placebo) , with respect to analysis variable Age, the total number of observation N = 19 , Mean = 39.0000000, Standard Deviation = 6.8879928 , Minimum = 25.0000000 and Maximum = 50.0000000.

In the third table depicts for treatment trt = 2 (Overall), with respect to analysis variable Age, the total number of observation N = 49 , Mean = 34.3469388, Standard Deviation = 8.5696923 , Minimum = 20.0000000 and Maximum = 50.0000000.

The below figure 2, shows the subject demographics and baseline characteristics Safety Population. It gives more detailed analysis including the mean ,median, min ,max with respect to different analysis variable like age gender and ethnicity for each treatment trt= 0, 1 ,2.

Regarding the analysis variable gender, it provides observations on the quantity and proportion of male and female patients who received which treatment. Similar with respect to analysis variable ethnicity, it gives detailed information on how many and what proportions of white, black or African American, Asian American Indian, native Hawaiian, and other people have received which treatment.

The below figure 3 depicts scatter plot of age versus name Discrete data points make up the graph, each of which

The SAS System

The MEANS Procedure

TRT=0

Analysis Variable : AGE				
N	Mean	Std Dev	Minimum	Maximum
30	34.6666667	9.1964511	20.0000000	50.0000000

TRT=1

Analysis Variable : AGE				
N	Mean	Std Dev	Minimum	Maximum
19	39.0000000	6.8879928	25.0000000	50.0000000

TRT=2

Analysis Variable : AGE				
N	Mean	Std Dev	Minimum	Maximum
49	36.3469388	8.5696923	20.0000000	50.0000000

Fig. 1. Proc Means Table

14.1.3.1 Subject Demographics and Baseline Characteristics
Safety Population

	DS1234 (N=30)	placebo (N=19)	Overall (N=49)
AGE(Years)			
N	30	19	49
Mean(sd)	34.7(9.20)	39.0(6.89)	36.3(8.57)
Median	32.5	40.0	38.0
Min.Max	20 .50	25 .50	20 .50
Gender[n(%)]			
Male	17(17.3)	12(12.2)	29(29.6)
Female	13(13.3)	7(7.1)	20(20.4)
Ethnicity [n(%)]			
Hispanic or Latino	16(16.3)	5(5.1)	21(21.4)
Not Hispanic or Latino	14(14.3)	14(14.3)	28(28.6)
Race[n(%)]			
White	2(2.0)	4(4.1)	6(6.1)
Black or African American	9(9.2)	4(4.1)	13(13.3)
Asian	4(4.1)	4(4.1)	8(8.2)
American Indian	3(3.1)	1(1.0)	4(4.1)
Native Hawaiian	3(3.1)	1(1.0)	4(4.1)
Other	9(9.2)	5(5.1)	14(14.3)

Reference Listing 16.2.4.1
Note: SD=standard deviation,Min=Minimum, Max=Maximum

Fig. 2. Description Statistics

represents a person with a unique name and age. The age numbers would be placed on the y-axis of the graph, and the name labels would be on the x-axis. Each name and age combination is depicted as a separate data point rather than a continuous line or curve due to the graph's discrete nature.

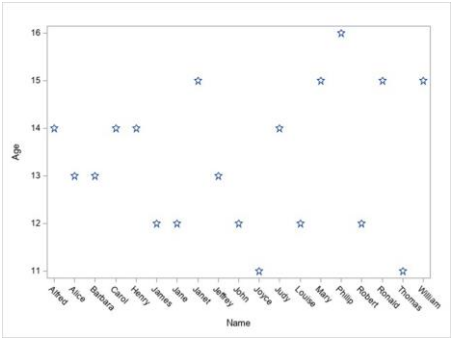

Fig. 3. Scatter plot of age versus name

The figure 4 represents the line graph of age versus name. The graph illustrates the relationship between various names and the corresponding ages associated with those names. The graph displays a gradual transition between ages and names rather than discrete data points. The names are shown on the x-axis, while the ages are shown on the y-axis. On the x-axis are the names, and on the y-axis are the ages, are plotted.To show the connection between names and ages, the graph shows a continuous line connecting the data points. Giventhat it more clearly illustrates trends and patterns, this kindof graph enables a more sophisticated understanding of the relationship between names and ages. This allows us to gain insights such as recognizing typical age ranges for particular names or seeing any age-related trends based on distinct namesby allowing us to visualize the general shape of the age distribution in respect to the names. It's crucial to remember that numerous data points with the same name but different ages may exist when working with continuous data. In this instance, a continuous line or curve rather than discrete points would be used to depict the distribution of ages for each name on the graph.

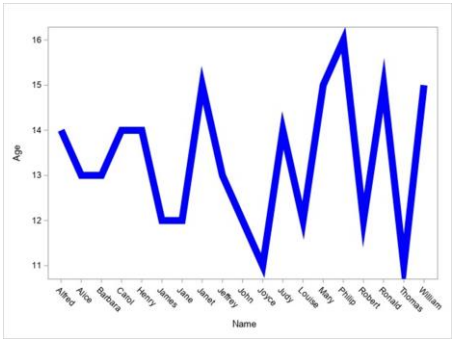

Fig. 4. Line Graph of age versus name

The age versus sex storyline is shown in the figure. The box plot in "Age versus Sex" shows the distribution of ages between the two sexes or genders. It displays important statistical measures and looks for any variations or patterns using a box-and-whisker graphic.

The male and female categories in this plot are represented by the x-axis. The age values are displayed on the y-axis. The interquartile range (IQR), which includes the center 50% of the data, is shown as a box in the plot. The lower quartile (Q1) is represented by the bottom of the box, while the higher quartile (Q3) is shown by the top of the box. The median (Q2), or midpoint of the data, is often represented by the line inside the box.

The variability or spread of the data is represented by the whiskers that protrude from the box. Depending on how the box plot is implemented specifically, these can be determined in a variety of ways. For instance, they might be used to generate a statistical formula or indicate the minimum and maximum values within a given range. The plot may additionally include outliers, or data points that deviate greatly from the normal distribution. Beyond the whiskers, these are often displayed as individual data points or as asterisks.

The box plot makes it simple to compare how the ages of people of different sexes or genders are distributed. You may evaluate the central tendency, spread, and skewness of theage distribution within each group by contrasting the positionsand sizes of the boxes. Additionally, it is simple to detect the presence of outliers. The box plot, as a whole, offers a succinct assessment of the age distribution between sexes or genders, enabling insights into potential

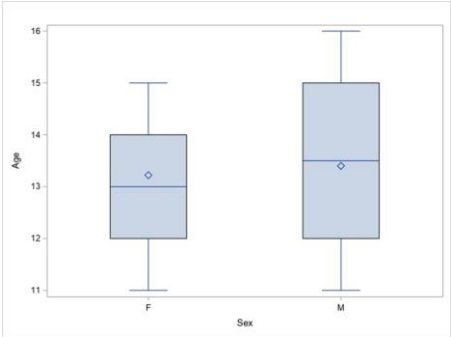

Fig. 5. Box plot of age versus sex

VI. Conclusion

This paper presents a study data tabulation in the demographics domain from the vast amount of clinical trial data in accordance to the SDTM Specification sheet provided by CDISC was created . Various graphs were created like scatter plot , Line Graph , Box Plot were plotted using lists and figures in Sas programming with respect to various analysis variables like age sex and ethnicity . Tables were generated for providing means, median max extra. Descriptive Statistics was provided for Demographics Domain in Study Data Tabulation Model.

References

[1] Sam Hume and Jozef Aerts and Surendra Sarnikar and Vojtech Huser, "Current applications and future directions for the CDISC Operational Data Model standard: A methodological review", Journal of Biomed- ical Informatics : Express Briefs vol. 60, pp. 352-362, 2016, doi: 10.1016/j.jbi.2016.02.016.

[2] G. T. Bartoo, "Design and execution of clinical studies for medical devices," The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Francisco, CA, USA, 2004, pp. 5135-, doi: 10.1109/IEMBS.2004.1404425.

[3] W. Tiancai, L. Baoyan, H. Liyuan, L. Xiaoping, W. Xin and Z. Yenning, "A randomization and trial supply management system for adaptive clinical studies of TCM and its scientific research application in recurrent tuberculosis," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 1917-1921, Doi: 10.1109/BIBM.2018.8621116.

[4] S. M. Reddy and S. Moriyama, "Exploring Multi Feature Optimization for Summarizing Clinical Trial Descriptions," 2020 IEEE Sixth InternationalConference on Multimedia Big Data (BigMM), New Delhi, India, 2020,pp. 341-345, Doi: 10.1109/BigMM50055.2020.00059.

[5] G. Lotz, T. Peters, E. Zrenner and R. Wilke, "A domain model of a clinical reading center - Design and implementation," 2010 An- nual International Conference of the IEEE Engineering in Medicineand Biology, Buenos Aires, Argentina, 2010, pp. 4530-4533, doi: 10.1109/IEMBS.2010.5626032.

[6] L. Fu, S. Ding and T. Chen, "Clinical Data Management Sys- tem," 2010 International Conference on Biomedical Engineer- ing and Computer Science, Wuhan, China, 2010, pp. 1-4, doi: 10.1109/ICBECS.2010.5462386.

[7] K. Ravvaz and J. Weissert, "An Interactive Tutorial on Simulated Pharmacogenomic Clinical Trials," 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, USA, 2018, pp. 445-445, doi: 10.1109/ICHI.2018.00093.

[8] W. Cui, S. Hou and H. Shao, "Design and Implementation of Clinical Trial Collaboration and Management System," 2013 6th International Conference on Intelligent Networks and Intelligent Systems (ICINIS), Shenyang, China, 2013, pp. 219-222, doi: 10.1109/ICINIS.2013.63.

[9] G. B. Laleci, M. Yuksel and A. Dogac, "Providing Semantic Interoperability Between Clinical Care and Clinical Research Domains," in IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 2, pp. 356-369, March 2013, doi: 10.1109/TITB.2012.2219552.

[10] K. Kitamura, M. Irvan and R. Shigetomi Yamaguchi, "Medical knowledge discovery by randomly sampled "patient characteristics" formatted data," 2022 Tenth International Symposium on Computing and Networking Workshops (CANDARW), Himeji, Japan, 2022, pp. 323-329, doi: 10.1109/CANDARW57323.2022.00035.

[11] E. Hughes and E. Lada, "Looking beyond the model: Data input, collection, and analysis with SAS® simulation studio," 2017 Winter Simulation Conference (WSC), Las Vegas, NV, USA, 2017, pp. 4423-4423, doi: 10.1109/WSC.2017.8248151.