PharmaSUG 2020 – Paper DS-023

Data Transformation: Best Practices for When to Transform Your Data

Janet E. Stuelpner, SAS Institute, Inc.
Olivier Bouchard, SAS Institute, Inc.
Mira Shapiro, Analytic Designers LLC

ABSTRACT

When is the best time to create the CDISC standard data? This has been debated for many, many years. Some say that it should be done at the very end of the study before the protocol is submitted. Some say to transform the data at the very beginning of the study as subjects start to enroll. And some do it as needed as the study is enrolling, the data is being cleaned and the shells of the tables, listings and figures are in the process of creation. This is a great forum for experts in the field to give their opinion as to how and when to perform the transformation into CDISC format.

INTRODUCTION

All submissions must be in CDISC format. We all know that. However, when is the best time to transform your raw or CDASH format data into SDTM and ADaM? We will be exploring this during this roundtable discussion. This paper will raise issues surrounding the different definitions of transformation in the life science industry.

TRANSFORMATIONS DEFINED

According to the Merriam Webster dictionary "Transformations are usually applied so that the data appear to more closely meet the assumptions of a statistical inference procedure that is to be applied, or to improve the interpretability or appearance of graphs. Nearly always, the function that is used to transform the data is invertible, and generally is continuous." [1] The transformation of which we speak is from the raw data into standard formats. In the life science industry, CDISC is the standard which is now required for a regulatory submission.

There are many thoughts about when the transformation should take place. Table 1, 2 and 3 below, show the pros and cons of transforming the data at various times during a study. But it is even bigger than the actual programming of the data, who is going to do it and when it should be done. There are many factors to consider. Time to complete, cost and having the personnel to finish the transformations are big considerations. But when to start the process is a crucial decision that will influence all of the components.
.

PHASE OF STUDY

As we all know, in the process of identifying safe and efficacious solutions, it takes the discovery of thousands of compounds or devices. Many of the Phase I and Phase II trials do not create a compound that will go into a Phase III trial for submission and approval. So when is the optimal time to start performing transformation with regard to the phase of the study?

Some say that it isn't worth the money nor the time to do any transformations until the compound starts in Phase III. It is no guarantee, but there is a better chance that you aren't wasting your time and resources because the trial is more likely to be included in a submission. If this is the plan of attack, then the team will need to go back to provide the CDISC domains for all of the studies up to and including Phase III.

There is the other possibility that there is a great deal of evidence that the compound will go all the way to Phase III. If this is the case, maybe it is better to start programming the SDTM domains right away and then you don't need to go back to Phase I and Phase II studies to perform the transformations.

CDASH, SDTM, AdAM, DEFINE.XML

When CDISC started to define the various standards, which was the first one that they tackled? Believe it or not, the first one was SDTM.[2] They actually started in the middle and worked out to the other ends. CDASH was created after SDTM (one would think that it was the other way around). Now, we look at CDASH as the beginning. When we look at the metadata for CDASH, we see that the variable names are the same as some of the names in SDTM. That makes the transformation to SDTM domains easier and quicker. But we also need to look at the fact that, in an ideal world, we should create the AdAM domains directly from the SDTM domains. In order to do that, we need to do some manipulation to the date and time fields in SDTM. All of the SDTM date and time fields are in character format. In order to perform calculations (e.g., days from baseline), we need to convert the dates and times back into a number in a date and/or time format. When the date is incomplete, this is not always an easy task. There are nuances to all of this programming and subject matter experts are required to work on these issues. The question then becomes, when to start programming the analysis domains so that they can be used for the tables, listing and figures.

Lastly it is the DEFINE.XML (the Define) that needs to be created for submission. The Define is mostly metadata. It also includes the value level metadata and references to external controlled terminology. When is the best time to create this document? Some say that the best time is at the beginning of the study. Once all of the variables are set (especially the permissible ones), you can start to create the Define. You need the value level metadata from the data values as you are enrolling. It doesn' generally change that much, but it could. The controlled terminology dictionary needs to be decided upon. Which version is the best to include for a study? This all will be included in the Define.

PROS AND CONS OF TRANSFORMATION METHODOLOGIES

There are many pros and cons as to when to perform all of the needed programming to transform the data from its raw state (most likely CDASH) and SDTM and then from SDTM into AdAM. We have created a table of all of the pros and cons for each step of the way. It has been our experience that there are some main themes here. Of course, you need to consider the number of staff devoted to doing this task. Another issue may be the knowledge or expertise of the staff. And of course, the total time that it might take to provide the transformation.

Ultimately, it is up to the number of resources and the cost that determines with is the best scenario to create the domains that are needed for your studies.

Pros	Cons
Programs reflect standards requirements	Investigators are still providing comments and changes to CRF and potentially protocols
Organized according to standards from the start	Changes may be more extensive due to shifting priorities, datasets and intermediate outcomes
Easy when data collection is close to CDASH	If SOPs don't exist, they need to be completed early
Availability of resources is greater	Need to build trust and lines of communications between all parties and it may be expensive to make changes
Have more time to gain CDISC expertise	Staff may create a better product after working on the project for some time after taking a deeper dive into the data and working with sponsors
Ability to be proactive with any changes	
Can run diagnostics and when final run at the end of the process when all the data is available, it should be clean	
Team gets in the habit of adhering to standards and can build	

programs/documentation in a modular manner to facilitate changing requirements	
Mistakes, omissions and incorrect inclusions can be caught earlier requiring less time and effort to correct	

Table 1. Early Transformation

Pros	Cons
Data Format will probably stay the same at this point	Reinventing the wheel if there are a great number of changes
Still can make changes to transformation code	Extra expense if an outside entity is preparing the FDA package for submission
Team has more knowledge of project specifics, sponsors and data sources	
Mid-course corrections of omissions, errors or changes to requirements can be reflected in the building of domains, tables and documents	

Table 2. Middle of the study

Pros	Cons
More time to formulate SOPs	Time is of the essence to submit
Any data anomalies would have shown up for more consistent programming	Last minute changes including items that are difficult to map and need quick decision-making
Team will have complete knowledge of requirements, data inconsistencies and complete package details including what needs to be included in the reviewer's guidelines	Need to make metadata changes to create Define.XML
Potentially lower cost if passing the creation of the final package on to an external entity	Less efficient

Table 3. After enrollment is complete

CONCLUSION

There aren't any right or wrong ways to accomplish these tasks. There are a great number of factors to take into consideration when tackling this topic. It all comes down to expertise, staffing, timing and cost. When providing the data, documentation, and reports that are needed for a submission, getting everything done in a timely manner can save money for a company. Hopefully, in this paper, we have provided some considerations that will help your team decide the best way to complete these requirements . We recognize that these decisions are difficult, but when approached with a great deal of thought and consideration, it can make all of the difference to the company, the regulatory agency, and the public.

REFERENCES

1. Merriam-Webster's collegiate dictionary (10th ed.). (1999). Springfield, MA: Merriam-Webster Incorporated.
2. www.cdisc.org

ACKNOWLEDGEMENTS

From Janet: This paper would not have been written if it hadn't been for the support given to me by my husband, Robert Stuelpner. Bob diligently read this paper to correct obvious errors and keep me on the right track. His criticisms were constructive and his support never ending.

From Mira: Special thanks to Janet Stuelpner for including me in this project. I always enjoy collboarations where I can provide my expertise and also enhance my knowledge.

CONTACT INFORMATION

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

Janet Stuelpner
SAS Institute, Inc.
Janet.Stuelpner@sas.com

Olivier Bouchard
SAS Institute, Inc.
Olivier.Bouchard@sas.com

Mira Shapiro
Analytics Designers LLC
Mira.Shapiro@gmail.com