PhUSE US Connect 2019

Paper SI11

Simplifying the Integration Riddle

Deepak Ananthan, Zifo RnD Solutions, Chennai, India
Arvind Sri Krishna Mani, Zifo RnD Solutions, Chennai, India

ABSTRACT

Have you ever wondered "How many ways the SDTM standards can be interpreted and implemented"? That is exactly what me and my colleagues were wondering when we recently started working on integration of 9 studies for a sponsor. In these 9 studies there were multiple partners for EDC, DM and CDSIC conversion in all sorts of permutations and combinations. Huge studies (1500 subjects) with very bad data quality to add to the fun. And for the cherry on top of the cake, one those 9 studies are an extension study of another study. And we have written a separate paper about challenges & questions in converting the extension studies to CDISC. From our experience the need for integrated analysis is only increasing, and the situation explained above is most common. In line with this trend and to promote collective learning we would like to share our experience in this paper.

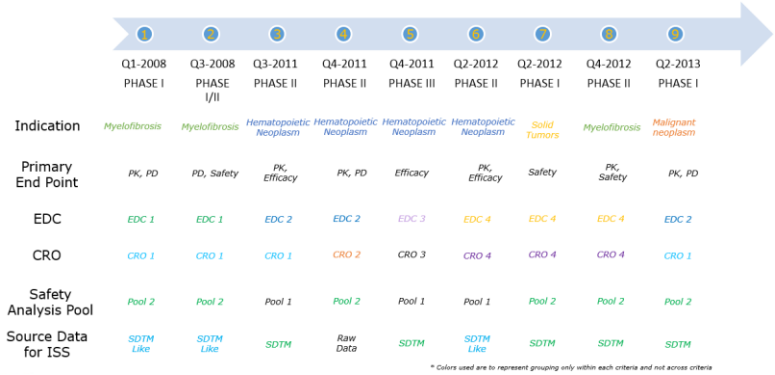
INTRODUCTION

FDA guidance states that an Integrated Summary of Safety (ISS) and Integrated Summary of Effectiveness (ISE) should be included in the Common Technical Document (CTD) for a regulatory submission.

An ISS is an 'Integrated Summary of Safety' combining the safety results from different studies conducted in a compound; whilst an ISE is an 'Integrated Summary of Effectiveness' combining the efficacy results from different studies conducted in a compound. In a CTD, the ISS and ISE are included in Module 5, specifically under Section 5.3.5.3 'Reports of Analyses of Data from More than One Study'. Section 2.7.4 'Summary of Clinical Safety' contains the results of the ISS. In other words, the results from all clinical trials performed on the study drug are pooled together and analyzed, producing combined statistical results. Presenting the FDA with a coherent and clear integration of the data from your product development program is imperative and requires a sound strategy and a skillful approach. As the scope of the paper is to share from our experience we will focus only on ISS in this paper from this point on.

OUR ISS EXPERIENCE– SUMMARY OF STUDIES INCLUDED IN THE ISS

The Sponsor had selected 9 studies conducted in a compound as part of this ISS. The studies were conducted between 2008 to early 2013. Not all the studies conducted on the compound was for the same indication and so the analysis and endpoints in those studies were also different. For these 9 studies the sponsor had partnered with 4 different CROs. And in many cases the CRO for data management and CDISC conversion was also different. All these differences pose different challenges when executing an ISS study. Below picture is a summary of variety involved in those studies.

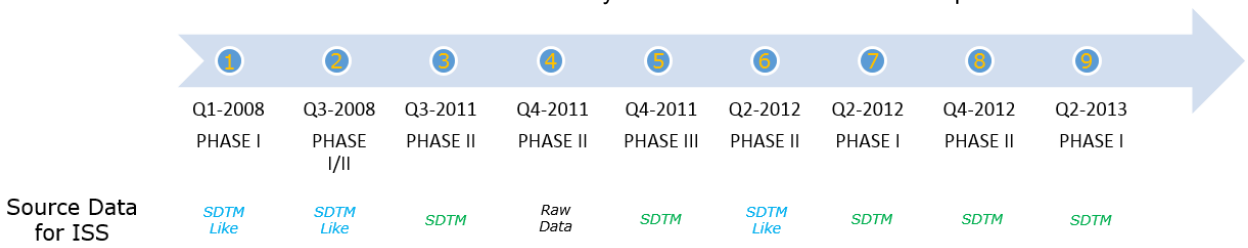


Picture 1

The most common approach in ISS is consolidating all the SDTM datasets created for those studies and then creating analysis datasets on top of that to support the analysis needed. But it is also common practice where we see that the analysis datasets for ISS are directly created from Raw data itself. In this project our approach was to consolidate and standardize the SDTM datasets available and then creating analysis datasets needed for the ISS. We have summarized some of the very interesting observations / challenges we had while creating the ISS.

1. VARIETY IN THE SOURCE DATA FOR ISS

It's rare that all the studies in the ISS have SDTM compliant datasets (irrespective of what the sponsor claims). And that is the case with this ISS as well. About three studies had datasets that follow the basics principles (what we call SDTM-Like) but cannot be considered as SDTM compliant. And 1 study had only raw CRF data. The rest of the studies had data in SDTM format. You can see the variety in the source data in the below picture.



Picture 2

And if you think that the studies that are already in SDTM are the easiest and will take the least time compared to one created directly from Raw data then you could not be more wrong. Big surprises like some of the CRF data missing in the SDTM data is more frequent than you think. And when that happens it takes more time to identify the reason for the missing records by contacting the sponsors. This usually would require more follow-ups and time than fixing the issue as the knowledge on those old studies within the sponsor will also be limited.

Pinnacle validation Issues: SDTM datasets does not mean SDTM compliant datasets. If you are not surprised by this statement or understand what we mean by this statement, then we know you have been part of the ISS studies. The most frequent and biggest gap we notice is not running pinnacle21 validation or not fixing the issues identified in the pinnacle21 validation. Below are some examples of issues that were left unattended in the source SDTM datasets.

SD0012	FDAC106	EXSTDY is after EXENDY	Limit	Error
SD0012	FDAC106	EXSTDY is after EXENDY	Limit	Error
SD0013	FDAC107	EXSTDTC is after EXENDTC	Limit	Error
SD0013	FDAC107	EXSTDTC is after EXENDTC	Limit	Error

Picture 3

SD0040	FDAC085	Inconsistent value for LBTEST within LBTESTCD	Consistency	Error
SD0040	FDAC085	Inconsistent value for LBTEST within LBTESTCD	Consistency	Error
SD1043	FDAC086	Inconsistent value for LBTESTCD within LBTEST	Consistency	Error
SD1043	FDAC086	Inconsistent value for LBTESTCD within LBTEST	Consistency	Error

Picture 4

Compliance to Controlled Terminologies: One of main reasons many prefer to consolidate the SDTM datasets, before creating the analysis datasets, is standardization of the terms used. But it would not be prudent to assume that all the studies in the package would have used the correct terminologies. In our case as you can observe from the example presented below the controlled terminologies that were used were not compliant. Identifying such differences ahead of time saves a lot of time in the conversion process.

QSTESTCD in source SDTMs	Corresponding QSTEST in Source SDTM	Actual QSTESTCD as per CT
FILLING	Filling Up Quickly When You Eat	EARLYSAT
PAINRIBS	Pain Under Ribs On Left Side	RIBSPAIN
ECOG	ECOG Performance Status	ECOG101
BNEPAIN	Bone Or Muscle Pain	BONEPAIN
LEFTPAIN	Pain Under Ribs On Left Side	RIBSPAIN

Picture 5

For identifying these differences, we had created a macro which would match the values in the SDTM datasets against the CT and provides a report of values that match with the CT and values that do not match with the CT. The values that do not match in the CT can then be run against our database of CTs and their corresponding values from older studies. Following a similar approach would save a lot of time. Even identifying only the direct and complete matches will result in significant savings in time.

Next challenge will be differences in the terms used / data collected between studies. Here as well having a macro to identify the frequency of the unique data presented across studies will be very useful. This macro helped in identifying inconsistencies between studies as well. We are sure that the below screenshot is something you would notice very frequently.

Example: AE relationship to study drug

Study 1	Study 2	Study 3	Integrated
No	Unlikely	None	None
Yes	Possible	Possible	Possible
	Probable	Definite	Probable
			Definite

Picture 6

The most tedious part of an Integration process is reconciling the individual studies to be consistent before moving on. Often, studies are created with a lack of foresight on the potential implications they would have on the integration process.

2. SDTM CREATED BY DIFFERENT CROS

It is common, we observe differences in the SDTM interpretation and implementation between datasets created by different people even within the same team / CRO. And when multiple CROs are involved for the CDSIC conversion then the problem only increases. We have seen examples where the same data was mapped to even different domains.

For e.g. MRI related information is collected was mapped to FA domain in one study and to a custom domain called MR in another study.

This problem is compounded when there is no additional information (define.xml, specification etc.) available to understand logic/ reason behind the mapping decisions. So, it is recommended to request for this information and dedicate some time to review the mapping logic used in the source SDTM datasets before proceeding with the mapping / consolidation for the ISS.

3. PROGRAMMING DIFFERENCES

It is easy to assume that while creating a submission ready ISS/ISE, all you need to do is stack the datasets of the component studies and apply the Statistical tests on the result, and voilà, your submission is ready. But the reality is far from that. When you set all your datasets together, you will notice data truncation issues. This is because in some older datasets the length of the variables might not be updated to the maximum length and would be retained with the maximum value (200). Even worse sometimes they are assigned some arbitrary value. And if these differences in length assigned are not taken into consideration then while consolidating the datasets from different studies some data getting truncated is a real possibility. So, we recommend creating a macro to review the length of the same variable in all the source datasets and adjusting the length with the maximum length in all the datasets. A classic example is when in one study DTC variable is presented with time and another study it has only the date part. Like in the below picture SVENDTC is different in different studies.

NAME	LENGTH
domain	2
studyid	10
svendtc	10
svendy	8
svstdtc	10
svstdy	8
usubjid	18
visit	17
visitnum	8

NAME	LENGTH
domain	2
studyid	10
svendtc	17
svendy	8
svstdtc	113
svstdy	8
usubjid	18
visit	17
visitnum	8

Picture 7

And it is not only the length that vary between studies the datatype can also be sometimes wrong. We have even seen scenarios where the VISITNUM was in character format and VISIT was in numeric format. So, focus should also be identifying these differences before the programming starts.

NAME	TYPE
domain	2
studyid	2
svendtc	2
svendy	1
svstdtc	2
svstdy	1
usubjid	2
visit	2
visitnum	1

NAME	TYPE
domain	2
studyid	2
svendtc	1
svendy	1
svstdtc	1
svstdy	1
usubjid	2
visit	2
visitnum	2

Picture 8

The above picture shows the differences in the datatype between the studies. Note that VISITNUM, SVSTDTC and SVENDTC are with differing in their data types.

4. EXTENSION STUDIES

In the set of studies for this ISS, one of the studies is a long-term extension study of another study. And in this case the subjects who were part of the parent study will only be able to join the long-term extension study. These types of studies will pose some specific questions.

How do we treat the same subjects in the parent and the extension study? Are they considered as one subject or different subjects for the analysis? What clarity exist in the SAP regarding handling these extension studies? The SAP should clearly explain this in alignment with the analysis requirements.

5. PROBLEMS POSED BY VOLUME OF DATA

Another frequent challenge in extension studies is posed just by the sheer volume of data in the domains. Especially data intensive domains like EG, LB where datasets with more than million records is common. The biggest we had was LB with little more than 1.5 million records. Apart from the other issues discussed in this paper there is also a real possibility that the volume of the data poses a problem in terms of program execution times.

Here is where the adherence to good programming practices can provide extraordinary rewards. Basic hygiene like retaining only the variables needed in the datasets can have a huge impact in the execution times and disk space. And we all know how frequently SAS programmers use PROC SQL Joins. When the number of records is in the millions then using more efficient HASH objects can save a lot of time. In our study we were able to bring down the processing time by 50% by following good programming practices and making the programs efficient.

This table here explains a very large dataset join. Data step and proc sql requires multiple small join steps for such operation whereas the hash takes just a single data step for that to do. Shown below is the time comparison for the operations to happen.

	Real Time (seconds)		
Step	Standard DATA Step	PROC SQL	DATA Step HASH
1.1 (1st Sort*)	0.08	N/A	N/A
1.2 (2nd Sort†)	1,257.10	N/A	N/A
1.3 (3rd Sort‡)	0.13	N/A	N/A
2.1 (1st Pre-join†)	82.49	3,004.86	N/A
2.2.1 (2nd Pre-join‡)	1.91	N/A	N/A
2.2.2 (2nd Pre-join‡)	0.04	3.10	N/A
3 (Join)	81.74	104.37	275.36
Total	1,423.49	**3,112.33**	275.36

	# of Obs	# of Vars	Size (KB)
* Lookup	86	5	128
†Lab Results	10,620,791	21	26,552,200
‡Cancelled	786	19	2,156
Final	10,621,577	23	2,580,480*

* Compressed using binary option

Picture 9

And here is a sample how it helped reduce some time with HASH objects. Even though the 33% reduction in execution time here in one small step might not be significant. When this can be achieved in many steps it will result in a significant reduction in total execution time.

HASH object

```
NOTE: There were 36556940 observations read from the data set WORK.STUDY1.
NOTE: The data set WORK.ES has 36556940 observations and 7 variables.
NOTE: DATA statement used (Total process time):
      real time           1:15.89
      cpu time            1:00.91
```

Picture 10

Proc SQL

```
NOTE: Table WORK.STUDY2 created, with 36556940 rows and 6 columns.

55   quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           1:26.97
      cpu time            1:30.99
```

Picture 11

The volume of data poses a greater problem during debugging. The most common practice in the industry is to QC by parallel programming. Imagine debugging the differences when there are million records. In cases like these we would mostly be forced to split the datasets to fit into the file size limits in the submission. But even without those external factors it is very wise to split these huge datasets into smaller chunks during development and QC to be more efficient in identifying and fixing the issues quickly. Splitting datasets at Test code level or Category level is most frequent option we would recommend during development.

ANALYSIS CONSIDERATIONS

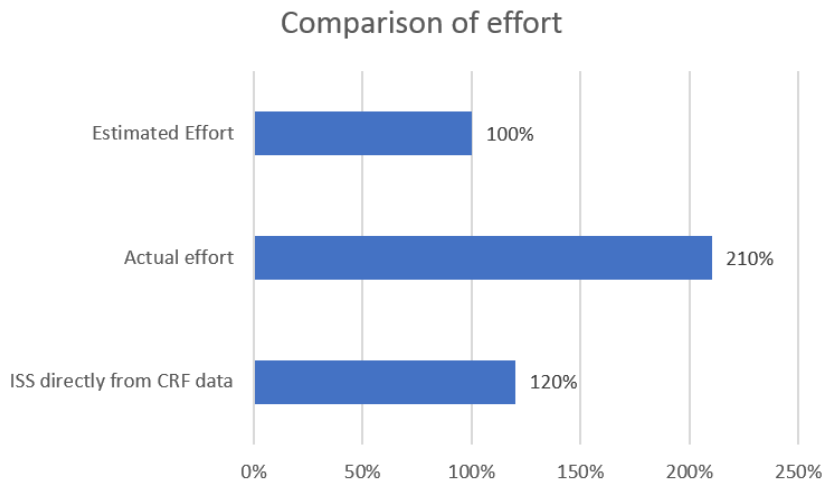
The topics discussed so far in this paper is more focused on the first step of consolidating the SDTM datasets from the source studies. Here we focus on some of the important aspects to consider when creating the Analysis datasets as well as analysis of the data.

The focal point of any integration study is the analysis of its endpoints. Identifying data points that need to be analyzed, and their source, plays a vital role. Combining the data points and analysing them is not straightforward and depends on the basis for data collection. Some of the examples that we came across are listed below:

1. Studies usually do not have a homogeneous study population as the inclusion and exclusion criteria may differ. In such cases, the efficacy measures may be inconsistent and may mean that including certain studies in the ISS/ISE could be inappropriate. Therefore, the rationale for including or excluding certain studies should be justified and documented prior to analysis.
2. Dose ranges will be different in different studies based on study design. Subject pooling strategy can be done in these situations, for dividing the population into groups that reflect the endpoints. For example, Study A may dose patients for 3 weeks, compared to Study B, that doses patients for 6 weeks. A combined summary of the number of patients who reported an AE may not be appropriate, as one group of patients could be ‘at risk’ and followed up for a greater period. The statistician can help identify the appropriate methodology for addressing such issues. In this example, a solution may be to present AEs using a denominator that adjusts for the time at risk.
3. It is also very common that the SAP will not be final before starting the ISS. In the ISS we did there were lot of realizations on seeing issues in the data, inferences from the data where the SAP had to be updated to provide clarity on how to handle certain scenarios. These updates can be avoided if the issues discussed earlier are identified early in the project.

PLANNING THE EXECUTION

One of the biggest challenges in executing such projects is estimating the effort needed for the project. And when estimating during the proposal stage with limited information and seeing that almost all the studies are already in SDTM format it is very easy to imagine this a simple case of stacking the datasets together and underestimate the effort. In fact, with little access to the data for these studies the effort we estimated was about 35% of what we spent at the end of this project. After the project was completed, when we did an analysis on the effort spent on the ISS we were surprised to see how seemingly small issues described above had contributed to such a deviation in the study budget. To knock us out, there was another ISS project executed by another team during the same time. The difference between these projects was, one was with data in SDTM format (the easier way) and the other was the more complex one where we did the analysis datasets for ISS directly from the CRF data (without ever converting to SDTM standards). And the comparison of efforts spent in those projects was another great surprise for us.



Picture 12

This does not mean we are recommending not converting the data to SDTM. It simply points to the fact how a seemingly simpler studies when the issues are not identified earlier can lead to budget overruns. The experience we gained from these projects helped us create checklists, macros, review strategies etc. and we were able to execute subsequent projects with greater control.

CONCLUSION

Even though the challenges posed by each ISS/ISE is unique, we believe with careful planning and anticipation of the problems and analyzing the studies for those problems before deep diving into the programming can save a lot of effort and there by improve the quality of the project.

CONTACT INFORMATION

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

Author Name: Deepak Ananthan
Company: Zifo RnD Solutions
Work Phone: +91 9500061270
Email: deepak.a@zifornd.com
Web: www.ZifoRnD.com