Creating an Integrated Summary of Safety Database using CDISC ADaM : Challenges, Tips and Things to Watch Out

Rajkumar Sharma, Genentech Inc.,
A member of the Roche Group, South San Francisco, CA

ABSTRACT

Most individual trials are not powered to identify trends and rare adverse events. By combining trails, rare events can be detected and a better picture can be obtained of the general safety profile of a drug. Creating an integrated safety database by pooling studies together can facilitate this. Pooling information across multiple studies into an integrated safety database generally requires that relevant data from all studies have the same structure and metadata standards. Since, CDISC structure is better defined; it can be used to create integrated safety database with ease. However, there are challenges and issues to watch out for. This paper will cover in detail on how to create integrated safety database, challenges associated with it, things to watch out for and tips.

INTRODUCTION

Pharmaceutical/Biotech industries are widely accepting and implementing SDTM and ADaM standards. The main idea behind the implementation is to harness the potential which CDISC standards have to offer. Because of the wide acceptance/implementation of CDISC standards, programmers now are dealing with more studies which are in SDTM/ADaM format. This could be a real advantage for someone working on integrated safety database especially because less time is spent on standardizing data from different studies and more time to focus on critical analysis. Since ISS and ISE both are very important for drug approval it is very important to build solid integrated database to support for ISS/ISE reports. Main objective of this paper is to provide audience guidelines on how to build reliable integrated safety database using CDISC SDTM/ADaM standards.

Creating an Integrated Safety Database can be very daunting and challenging task especially when you have to deal with large number of studies. A programmer can easily get confused as where to start. Thankfully, life is little better if the studies needed to create ISS database are in CDISC/SDTM format. Datasets in SDTM format are comparatively easy to pool to create integrated safety database. If few studies are in CDISC/SDTM format and few others in non-SDTM format then it might be worth to convert non-SDTM studies into SDTM format based on timeline and resources availability. For this paper, focus is on creation of integrated safety database assuming all studies pooled for integrated database are in CDISC/SDTM format.
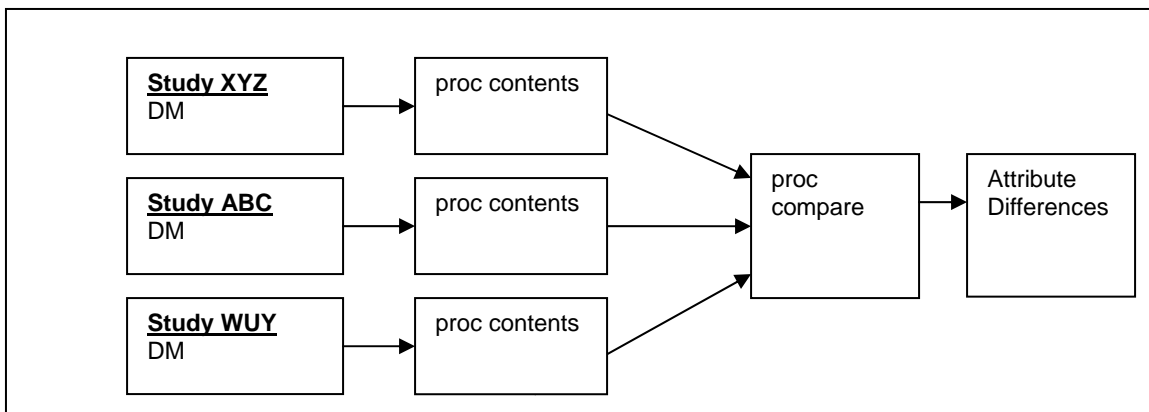
WHERE TO START?

This could be most confusing question for many especially if you are dealing with integrated database for the first time. The very first step should be to determine which studies are needed to generated integrated database. Once studies required for ISS are determined, programmer could follow following steps as guidelines to create integrated safety database.

1. DATA CHECK FOR CONSISTENCIES

If all the individual studies pooled for integrated database are in CDISC/SDTM format then the very first step before pooling the data could be checking the inconsistencies between the studies to determine if there are any issues. Since SDTM standard is pretty well defined, it is easy to do following checks programmatically with little efforts. Following types of checks could be performed.

   i)   **Attribute Check** – Main purpose of attribute check is to get an idea of attribute differences among the studies pooled for ISS. This check gives an idea as what adjustment needs to be made to the variable attributes latter down the road while creating integrated analysis datasets. Attribute check can be done programmatically by running proc contents on SDTM domains from all studies and then doing a proc compare to find the differences in length, type, format, etc. A SAS program/macro could be developed to perform these checks. Below is the schematic diagram using DM domain from three studies to give an idea. Similar checks could

be performed on all other SDTM domains which are needed to create integrated analysis datasets.

Below is an example showing differences in length of AEACN variable across three studies. In this case length of AEACN should be adjusted before integrating AE domain.

Study	Name	Length
Study ABC	AEACN	$22
Study XYZ	AEACN	$16
Study WBC	AEACN	$25

Tip: Attribute check is worth 1) if you are pooling SDTM data from different studies to build an integrated SDTM database because attribute adjustment is needed before creating integrated SDTM database 2) if few SDTM datasets were created by different entities.

ii)   **SDTM mapping Check** – SDTM mapping differences between studies could arise either because data was collected differently among different studies or because older version of SDTM IG was used for few older studies. Also, SDTM IG can be interpreted differently by different individuals which can lead to mapping differences. Before starting with integrated analysis dataset, it is good idea to perform this check. Since there could potentially be many SDTM datasets and many studies to check, it could be confusing as where to start. One approach is to focus on variables which are needed for ISS analysis. Based on scope of analysis, programmer could focus on annotated CRFs and check if the information programmer need for analysis is mapped consistently across studies or not. Programmer could also focus on supplemental qualifier domains since mapping rules are less stringent for supplemental qualifier domains and there could be some differences there. An example shown below illustrates that.

	Study ABC SUPPAE	Study XYZ SUPPAE
Date of Death	**AEDTHDTC**	**DTHDTC**

Mapping differences could be identified by either comparing annotated CRFs across studies manually or by creating a program which could perform these checks. Team working on ISS needs to determine which approach works best for them.
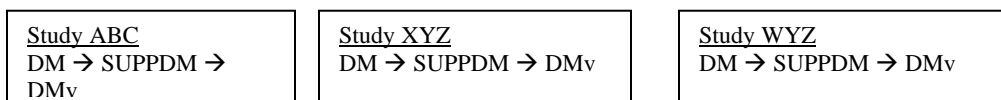
iii)   **Domain Check** – The idea behind domain check is to get an overview of what is available and what is not and to look out for any potential issues. If programmers are expecting certain domains to be present, it is a good idea to perform a domain check. Programmers can check 1) if all the domains they need for analysis are present across all studies and 2) if supplemental qualifier datasets were created for few studies and not for other studies. Based on study design it is possible that there might be supplemental qualifier for one study and not

for the other study. If the programmer is expecting certain information to be present in a suppqual domain across all studies then it is worth to perform this check.
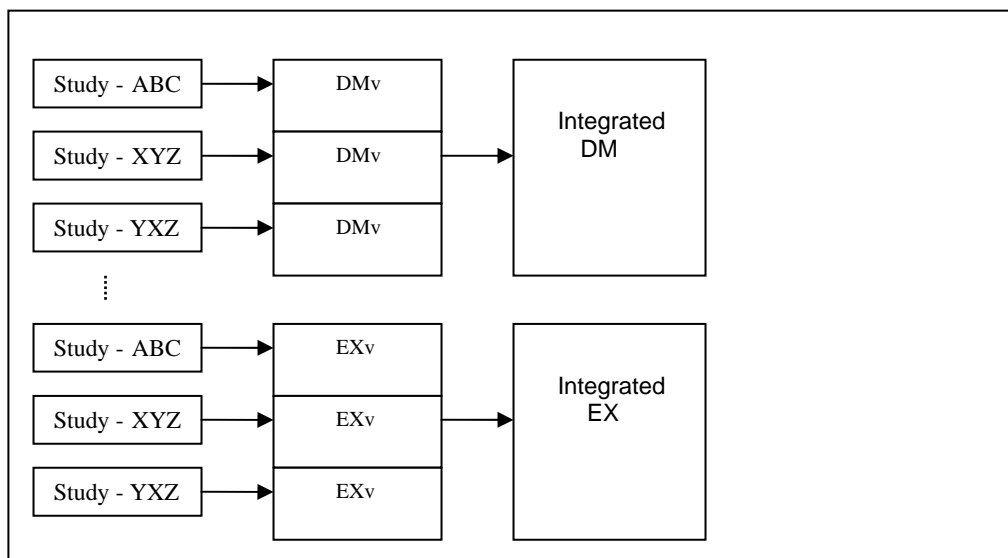
2: INTEGRATED SDTM DATABASE

An integrated ADaM database can be build from an integrated SDTM database. Integrated SDTM database can be created by pooling individual study SDTM domains and stacking them. This can be done following several steps shown below.

    i)   Merge individual study supplemental qualifier (suppqual) with parent domain. Below is an example using SDTM DM domain. Suffix 'v' is added to the resulting domain after the merge to differentiate it from regular SDTM domain. Similar merge should be performed on all other domains needed for analysis. If suppqual domain is not present then there is no need for this step.

Study ABC	Study XYZ	Study WYZ
DM → SUPPDM → DMv	DM → SUPPDM → DMv	DM → SUPPDM → DMv

    ii)   After the first step, the identical SDTMv domains from different studies can be stacked to create an integrated domain. An example below illustrates creation of integrated DM (demographic) and EX (exposure domain). Similarly, other integrated SDTM safety domains can be created.

Tip : Before stacking domains to create an integrated domain, it is recommended that user adjust attributes of the variables if they differ between studies to avoid any data issues (see step 1). After integrating domains it is advisable to drop SDTM variables which are completely missing in values.
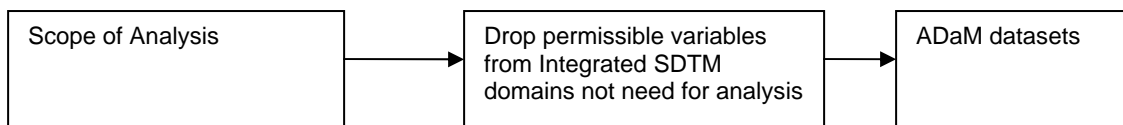
3. MedDRA VERSION LEVELLING

Since new version of MedDRA releases twice every year, it is very likely that not all studies pooled for integrated database are using same version of MedDRA for coding adverse events. It is important though to code all adverse events from all studies using same version of MedDRA to avoid version specific differences

and to correctly identify number of adverse events. MedDRA version leveling could be done before pooling data or after creating an integrated SDTM AE domain.

4. SCOPE OF ANALYSIS

After creating the integrated SDTM database, programmer could focus on creating analysis database for integrated summary of safety, as a next step. Integrated summary of safety reports usually contains adverse events, laboratory results, extent of exposure, demographics, deaths and discontinuation information. More safety information like vital signs, ecg, etc may also be included based on scope of analysis. The best approach is to focus on analysis scope and determine how many ADaM datasets are needed. It is possible that not all the SDTM variables are needed for analysis. Programmer could drop the permissible SDTM variables from the integrated SDTM domains if they not required for analysis. Below is a schematic diagram to outline the process.

Scope of Analysis	→	Drop permissible variables from Integrated SDTM domains not need for analysis	→	ADaM datasets

5. INTEGRATED DATABASE RULES AND GUIDELINES

 Once the scope of analysis is clear then the next step can be designing the integrated ADaM analysis datasets (after designing integrated SDTM database). It is recommended to handle the SDTM data inconsistencies of different studies in integrated analysis dataset (ADaM datasets) rather than at in integrated SDTM datasets. The reason is that ADaM standard is more flexible than SDTM standard as it was developed keeping different analysis needs in mind.

But before jumping on creating integrated analysis datasets, it is worth to document the design of the integrated database and set up general rules before programming. Since ADaM is designed for flexibility, there could be some confusion while defining variables names and roles. Team working on ISS could create a document to set up guidelines to avoid confusion. The document can include details such as

- Brief details about studies pooled for integrated database. For example – list of studies, studies are blinded or open label, phase of the studies (I, II or III), number of patients in each study, etc.
- List of SDTM domains pooled from each study.
- List of integrated ADaM datasets required for ISS reports.
- Naming conventions for ADaM variables. For example – Flag variables should end with suffix xFL, date variable should end with suffix xDT, etc.
- Date imputation rules (if any) for different analysis domains.
- Directory structure and locations of different datasets. Team could include details such location of SDTM datasets from individual studies, location of integrated SDTM datasets, location of integrated analysis datasets, etc.
- Detail about records dropped (if any) for analysis. For example – If patients are not treated with study drug were dropped from ISS analysis then study team can document this information.
- MedDRA version used for AE coding.
- Any other data handling information useful for the team.

The main purpose of creating such document is to have one single guidance document for all the team members so that they can follow the guidelines and avoid inconsistencies. The purpose of this document is not to serve as dataset specifications which should be a separate document.

6. ANALYSIS DATASET CREATION

There could be many approaches for creating integrated analysis datasets for ISS. Since ADSL is a fundamental domain that would be used throughout the analysis and is minimum requirement for CDISC ADaM, programmer can start with it first. Many ADSL variables like age, sex, race, country, etc are straight forward but some other variables might not be straight forward. ADSL treatment variables may need adjustment as per the need of analysis.

Main purpose of ADSL is to contain demographic information, treatment information, population flags, simple disposition information, etc. ADSL can get complicated very easily if lots of information is stuffed in it. If there is also a need to create ISE for a project then it is advisable to handle ADSL with extra care since it can be used for both ISS and ISE reports and it can get fully bloated and complicated. Best approach before adding any variable to ADSL is to carefully consider weather it really belongs in ADSL or in any other dataset. ADaM IG could be very helpful resource to make the decision.

Please note that it is recommended to make any adjustment to the values of variables in the in ADaM dataset itself and not in the integrated SDTM datasets. Please see an example below where differences between values of ARM are harmonized in TRT01P variable in Integrated ADSL dataset and not in DM ARM.

Study ABC – DM	Study XYZ – DM	Integrated - ADSL
ARM	ARM	TRT01P
PLACEBO	Placebo	PLACEBO
DRUG ABC BID	Drug ABC	DRUG ABC

Programming for integrated analysis datasets can be done in different ways but no matter what the programming approach is, it is a good idea to focus on few items shown below.

- **Dates**
  SDTM dates are in character format which could not be used for analysis. For analysis purposes numeric dates are needed. Study team need to determine weather they need all three DATETIME, TIME and DATE variables or just DATE variables. If time is not used for analysis then other two variables may not be needed.

  Tip: It is good idea to create a one SAS macro/program to do character to numeric date conversion instead of handling it in each separated analysis dataset program since it is easy to maintain and make changes to one program. If any date imputation is done then there should be a corresponding date imputation flag.

- **Controlled Terminology**
  There could be some differences in controlled terminology within the values of the SDTM variables in comparison to CDISC SDTM controlled terminology. The differences could be because SDTM controlled terminology was not followed as intended. The best solution in this case is to map any differences to the latest available CDISC SDTM controlled terminology. Few examples are shown below.

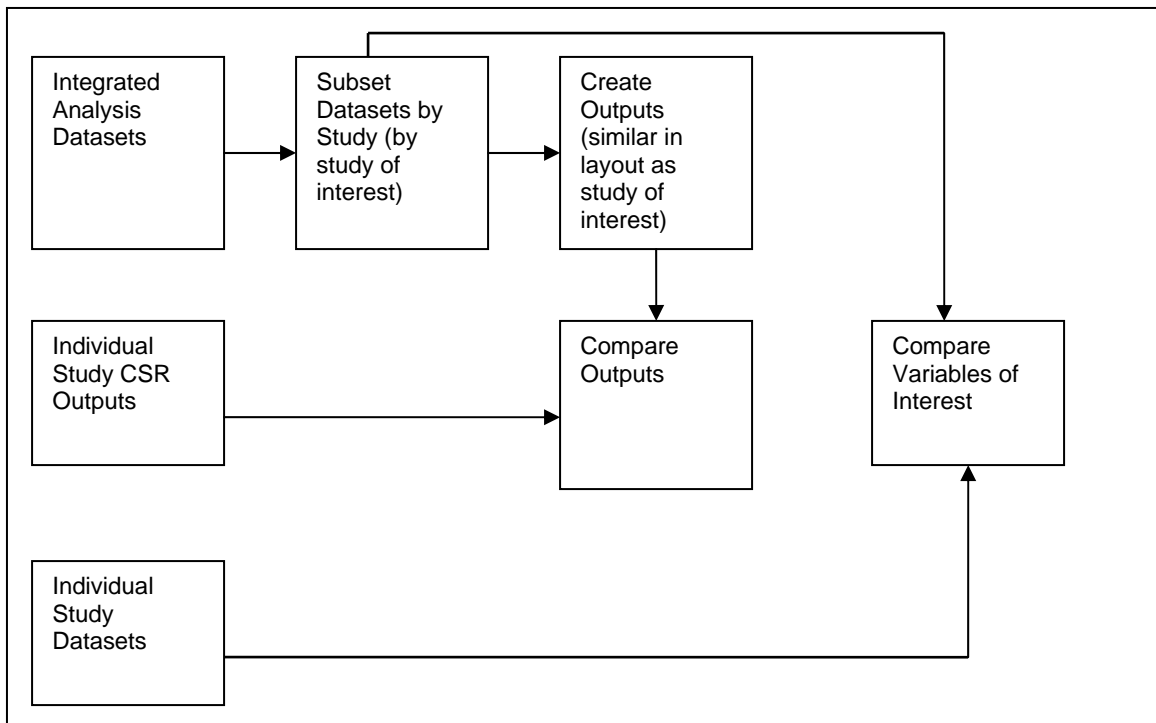|	Study XYZ	Study ABC	SDTM Controlled Terminology
SEX	M	MALE	M
	F	FEMALE	F
COUNTRY	AUSTRALIA	AUS	AUS
	DENMARK	DEN	DEN
AEREL			
	Y	DEFINITELY RELATED	Y
	N	NOT RELATED	N
		POSSIBLY RELATED	
		PROBABLY REALTED	
		UNLIKELY RELATED	

- **Units**

  Units of quantifiable data could be different across studies. It is advisable to adjust any differences in units in analysis datasets. Laboratory data collected from different sites could have this issue. It is advisable to check units for all the variables used for analysis (if they are expected to have units).

- **Baseline values and Flags**

  Baseline flags are part of some SDTM domains but it is possible these flags are not defined consistently across studies or it may not be what you need for ISS analysis. For purpose of analysis, it is a good idea to re-defined baseline flags within ADaM analysis datasets as per the need of analysis. For example within particular SDTM domain baseline may be defined as last assessment before study dose but for analysis you may need average value of two assessments to define baseline.

  Defining baseline values for roll-over is little tricky. If there is a roll-over study as a part of ISS then special care should be taken while defining baseline variables. Baseline values for roll-over studies are generally defined using baseline values from parent study and not from the roll-over study itself.

- **AE coding and missing data**

  There is possibility that some AE preferred terms or intensity might be missing in the AE datasets, especially if some studies are on-going. It is good idea to perform a check on MedDRA terms and AE intensity to see if there are any missing values. It is advisable to keep all 5 levels of MedDRA hierarchy in integrated AE datasets.

7. CROSS CHECKING

After integrated ADaM analysis datasets are ready, team can cross check the validity of the data by comparing the integrated analysis dataset against the individual study data. The idea behind this approach is to make sure that integrated analysis datasets are indeed true representation of individual study data and no relevant information is lost or got distorted.

There are two different possible approaches to cross check the data as shown by the schematic diagram below.

6

If team made any adjustment to integrated analysis datasets for controlled terminology, units, MedDRA version, etc there is possibility that some differences could pop out because of these adjustments while doing the cross checking. For example – if individual studies AE datasets pooled for ISS are in different MedDRA version then ISS AE datasets, some differences might come out while doing the comparison.

CONCLUSION

Task of creating an integrated database could be challenging but it be could be handled easily by developing and following a systematic approach. Since CDISC standards are well defined data checking and data integration could be done efficiently and easily. The approach mentioned in this paper purposes a reliable and systematic step by step solution for building an integrated database. A proactive approach and good planning along with data in CDISC standards can really help to build an efficient integrated database as shown in this paper.

REFERENCES

[1] Analysis Data Model(ADaM) Implementation Guide, Version 1.0 , Final, Published Dec 17, 2009
[2] Study Data Tabulation Model Implementation Guide, Version 3.1.2 Final, Published Nov 12, 2008

AUTHOR CONTACT INFORMATION

Rajkumar Sharma
Genentech, Inc., A Member of the Roche Group
1 DNA Way
South San Francisco, CA 94080
sharmar9@gene.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 registered trademarks or trademarks of their respective companies.