PharmaSUG 2024 - Paper DS- 305

Guideline for Creating Unique Subject Identifier in Pooled Studies for SDTM

Vibhavari Ajay Honrao, Ephicacy Lifescience Analytics

ABSTRACT

Demographic (DM) data set is the parent data set which includes a set of essential standard variables that describe each subject in a clinical study. One of these key variables is Unique Subject Identifier (USUBJID). SDTM IG does not provide any guidance on creation of USUBJID for pooled studies. Hence it becomes necessary to understand programming steps involved for statistical programmers.

In clinical trials, there are cases wherein subjects are re-enrolled for different studies for the same compound, and it can be difficult to identify the subject while maintaining CDISC compliance. For ISS analysis, pooling of studies becomes challenging due to multiple SUBJID, RFICDTC, RFSTDTC, RFENDTC etc. within same USUBJID from different studies.

This paper demonstrates various steps and programming logics involved to develop DM data set by taking hypothetical examples from multiple studies and creates pooled data sets.

INTRODUCTION

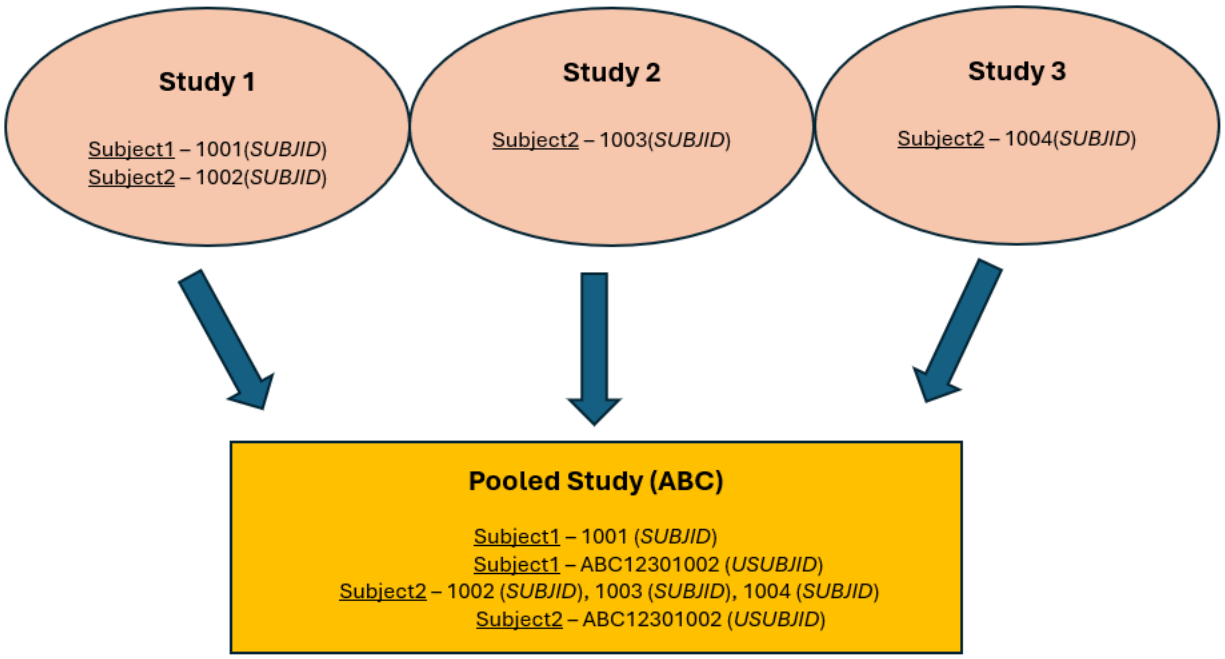
Pooling refers to combining data from multiple studies into a single data set, so that analyses can be run on derived analysis data set. Nevertheless, one must be cautious about the differences between studies which can affect the validity of and ability to interpret pooled analyses. One should also consider studies differences with respect to:

- Important demographic or disease characteristics (e.g., duration, severity, specific signs and symptoms, previous treatment, concomitant diseases and treatments, prognostic, or predictive biomarkers)
- Treatment practices, including methods of assessing effectiveness, specific test procedures (e.g., measurements of biomarkers and clinical endpoints)
- Study design features (e.g., study duration, study size, doses studied, visit frequency)

BACKGROUND

When a subject can re-enroll or re-screen across multiple studies or within same study, it becomes challenging to keep track of this one subject while maintaining CDISC compliance. In case of re-screening in same study, it is expected to make it clear that which subject identifiers (SUBJID) were recorded pre- and post-rescreening. When there is the case of enrollment in multiple studies, we can see from below example (Display 1) how subject identifier (SUBJID) recorded for each study.

Display 1. Pictorial representation of Pooled Study



The SDTM IG 3.3 states that for every subject "In Demographics (DM), only one record should be submitted for the subject". To support this argument, the FDA Technical Conformance Guide December 2023 v5.6 states the three different scenarios to handle the multiple enrollment or screening of the same subject.

1. For subjects with multiple enrollments in a single study, the primary enrollment should be submitted in DM. Additional enrollments should be included in a custom domain with a similar structure to DM. User can add the clarifying statements in the Reviewer's Guide (RG).
2. For subjects with multiple screenings and no subsequent enrollment, include the primary screening in DM with additional screenings in a custom domain with a structure similar to DM.
3. For subjects with multiple screenings and subsequent enrollment, include the enrollment in DM with screenings in a custom domain with a structure similar to DM.

The re-enrollment which has been discussed so far is all about the same study or extended study. However, in pooled studies it is observed that the subject is discontinued in one study and enrolled for second study of the same compound. Likewise, the subject may complete, discontinue second study, and get enrolled for third study. In such cases, the first or the last enrollment based on the study requirement should be submitted in DM data set. All available enrollments should be included in a custom domain with a similar structure to DM.

DM VS CUSTOM DM

According to Italian CDISC User Group Network Annual Meeting. 2020. "Updates on Handling Multiple Enrollments and Screenings Subjects in SDTM." Accessed July 25, 2021. "*Multiple Subject Instances (MSI) team proposes to create an additional special purpose domain called Demographics as Collected (DC) to act as the custom domain with similar structure to DM".*

The approach begins by standardizing the value of USUBJID in the beginning of pooling and subject identifiers that a given subject received. If the subject is enrolled once, the subject ID associated with screening is kept as SUBJID in DM, and as a SUBJID variable in custom demographic data set – Demographics as Collected (DC) for easy reference. Otherwise, the assigned subject ID is kept in DM and all other subject identifiers are stored in DC as Previous Subject Identifiers under SUBJID variable. The

custom domain DC will have more than one record per subject and all the subjects as mentioned in DM data set.

MULTIPLE RE-ENROLLMENT - EXAMPLE 1

Display 2. Custom Domain DC

Row	STUDYID	DOMAIN	USUBJID	SUBJID	DCSEQ	RFSTDTC	RFENDTC	RFXSTDTC	RFXENDTC	RFICDTC	RFPENDTC	SITEID	INVNAM
1	ABC123	DC	ABC12301001	1001	1	2021-02-25T11:50	2021-03-31	2021-02-25T11:50	2021-03-31	2021-01-30		1	JOHNSON, M
2	ABC123	DC	ABC12301002	1002	1	2020-02-27T11:50	2020-03-25	2020-02-27T11:50	2020-03-25	2020-02-11		1	JOHNSON, M
3	ABC123	DC	ABC12301002	1003	2	2020-09-11T11:50	2020-09-14	2020-09-11T11:50	2020-09-14	2020-08-15		1	JOHNSON, M
4	ABC123	DC	ABC12301002	1004	3	2021-02-11T11:50	2021-03-02	2021-02-11T11:50	2021-03-02	2021-01-15		1	JOHNSON, M

Row	BRTHDTC	AGE	AGEU	SEX	RACE	ETHNIC	ARMCD	ARM	ACTARMCD	ACTARM	COUNTRY	DCDTC	DCDY
1	1955-03-22	50	YEARS	M	ASIAN	NOT HISPANIC OR LATINO	Category1	ABC245-category1-IP	Category1	ABC245-category1-IP	USA	2021-01-30	
2	1955-03-22	50	YEARS	M	ASIAN	NOT HISPANIC OR LATINO	Category1	ABC245-category1-IP	Category1	ABC245-category1-IP	USA	2020-02-11	
3	1955-03-22	50	YEARS	M	ASIAN	NOT HISPANIC OR LATINO	Category1	ABC245-category1-IP	Category1	ABC245-category1-IP	USA	2020-08-15	
4	1955-03-22	50	YEARS	M	ASIAN	NOT HISPANIC OR LATINO	Category2	ABC246-category2-IP	Category2	ABC246-category2-IP	USA	2021-01-15	

Display 2. Custom Domain DC

Row 2 to 4: Shows multiple enrollments for the subject ABC12301002. This subject is re-enrolled in the same study and then re-screened and re-enrolled in another study. This information about the subject ABC12301002 will be captured in the custom domain DC.

For these two subjects DM data set will look as shown in Display 3.

Display 3. Main Domain DM

Row	STUDYID	DOMAIN	USUBJID	SUBJID	DMSEQ	RFSTDTC	RFENDTC	RFXSTDTC	RFXENDTC	RFICDTC	RFPENDTC	SITEID	INVNAM
1	ABC123	DM	ABC12301001	1001	1	2021-02-25T11:50	2021-03-31	2021-02-25T11:50	2021-03-31	2021-01-30		1	JOHNSON, M
2	ABC123	DM	ABC12301002	1002	1	2020-02-27T11:50	2021-03-02	2020-02-27T11:50	2021-03-02	2020-02-11		1	JOHNSON, M

Row	BRTHDTC	AGE	AGEU	SEX	RACE	ETHNIC	ARMCD	ARM	ACTARMCD	ACTARM	COUNTRY	DMDTC	DMDY
1	1955-03-22	50	YEARS	M	ASIAN	NOT HISPANIC OR LATINO	Category1	ABC245-category1-IP	Category1	ABC245-category1-IP	USA	2021-01-30	
2	1955-03-22	50	YEARS	M	ASIAN	NOT HISPANIC OR LATINO					USA	2020-02-11	

Display 3. Main Domain DM

In DM domain, ARM and ACTARM variables are kept blank for subject ABC12301002 because the subject may be randomized to different ARMs in different study. Nevertheless, this subject's randomization details can be found in custom domain. Also, for the subject ABC12301002, because of re-enrollment we have multiple exposure periods. Hence RFXSTDTC represents the date/time of the first study exposure across all participations, and RFXENDTC represents the date/time of the last study exposure across all participations.

There is an argument that states instead of creating another custom domain DC, one can add all possible subject IDs in SUPPDM. Though this approach reduces the time and effort, but we might lose the valuable information such as different RFSTDTC, RFENDTC, RFICDTC. Also technically, as per SDTM IG 3.3 it is not allowed to use "MULTIPLE" to indicate different SUBJID within one USUBJID, but rather is only used to indicate content captured for given subject (e.g., a subject report more than a single race). Refer to Display 5 DM domain.

Display 4 DC domain

Row	STUDYID	DOMAIN	USUBJID	SUBJID
2	ABC123	XY	ABC12301002	1002
3	ABC123	XY	ABC12301002	1003
4	ABC123	XY	ABC12301002	1004

Display 5 DM domain

Row	STUDYID	DOMAIN	USUBJID	SUBJID
2	ABC123	XY	ABC12301002	1002

Row	STUDYID	DOMAIN	USUBJID	SUBJID
2	ABC123	XY	ABC12301002	MULTIPLE

Row	STUDYID	DOMAIN	USUBJID	SUBJID	...	...	..	RACE
2	ABC123	XY	ABC12301002	MULTIPLE				MULTIPLE

USE OF DC AND DM IN POOLED STUDY FOR SDTMS

In general cases where we do not have custom domain for DM, we use DM data sets to derive the variables like xxDY. However, in pooled study, wherein we have multiple subject IDs for one subject we use the DC to get the required details. All other domains except DM can have all the SUBJID within the given USUBJID.

Display 6

ROW	STUDYID	DOMAIN	ARMCD	ARM	TAETORD	ETCD	ELEMENT	TABRANCH	TATRANS	EPOCH
1	ABC123	TA	Category1	ABC245-category1-IP	1	SCRN	Screening			SCREENING EPOCH
2	ABC123	TA	Category1	ABC245-category1-IP	2	C1	Category 1			TREATMENT EPOCH
3	ABC123	TA	Category1	ABC245-category1-IP	3	FUP	Follow up			FOLLOW-UP EPOCH
1	ABC123	TA	Category2	ABC245-category2-IP	1	SCRN	Screening			SCREENING EPOCH
2	ABC123	TA	Category2	ABC245-category2-IP	2	C2	Category 2			TREATMENT EPOCH
3	ABC123	TA	Category2	ABC245-category2-IP	3	FUP	Follow up			FOLLOW-UP EPOCH
1	ABC123	TA	Category3	ABC245-category3-IP	1	SCRN	Screening			SCREENING EPOCH
2	ABC123	TA	Category3	ABC245-category3-IP	2	C3	Category 3			TREATMENT EPOCH
3	ABC123	TA	Category3	ABC245-category3-IP	3	FUP	Follow up			FOLLOW-UP EPOCH

Display 6 TA domain

In pooled study, TA dataset will hold the information from all the studies.

Display 7

Row	STUDYID	DOMAIN	USUBJID	SUBJID	SESEQ	ETCD	ELEMENT	SESTDTC	SEENDTC	TAETORD	EPOCH
1	ABC123	SE	ABC12301001	1001	1	SCRN	Screening	2021-01-30	2021-02-25T11:50	1	SCREENING EPOCH
2	ABC123	SE	ABC12301001	1001	2	C1	Category 1	2021-02-25T11:50	2021-03-31	2	TREATMENT EPOCH
3	ABC123	SE	ABC12301001	1001	3	FUP	Follow up	2021-03-31		3	FOLLOW-UP EPOCH
4	ABC123	SE	ABC12301002	1002	1	SCRN	Screening	2020-02-11	2020-02-27T11:50	1	SCREENING EPOCH
5	ABC123	SE	ABC12301002	1002	2	C1	Category 1	2020-02-27T11:50	2020-03-25	2	TREATMENT EPOCH
6	ABC123	SE	ABC12301002	1002	3	FUP	Follow up	2020-03-25		3	FOLLOW-UP EPOCH
7	ABC123	SE	ABC12301002	1003	4	SCRN	Screening	2020-08-15	2020-09-11T11:50	1	SCREENING EPOCH
8	ABC123	SE	ABC12301002	1003	5	C1	Category 1	2020-09-11T11:50	2020-09-14	2	TREATMENT EPOCH
9	ABC123	SE	ABC12301002	1003	6	FUP	Follow up	2020-09-14		3	FOLLOW-UP EPOCH
10	ABC123	SE	ABC12301002	1004	7	SCRN	Screening	2021-01-15	2021-02-11T11:50	1	SCREENING EPOCH
11	ABC123	SE	ABC12301002	1004	8	C2	Category 2	2021-02-11T11:50	2021-03-02	2	TREATMENT EPOCH
12	ABC123	SE	ABC12301002	1004	9	FUP	Follow up	2021-03-02		3	FOLLOW-UP EPOCH

Display 7 SE domain

In this scenario, to derive SE data set, DC needs to be used otherwise EPOCH will not get populated correctly. Note that SESEQ will be unique per USUBJID, while TAETORD and EPOCH will be unique within SUBJID. Each participation is treated as a distinct and independent interaction with the planned study elements. SUBJID in conjunction with the timing of the elements (SESTDTC/SEENDTC) allows for the participations to be distinguished.

Rows 4,7 and 10: Represents the screening element for USUBJID ABC12301002, with SUBJID 1002, 1003 and 1004, respectively.

Rows 5,8, and 11: Represents the category 1 and category 2 elements for USUBJID ABC12301002, with SUBJID 1002, 1003 and 1004, respectively.

Rows 6, 9 and 12: Represents the follow-up element for USUBJID ABC12301002, with SUBJID 1002, 1003 and 1004, respectively.

USE OF DC AND DM IN POOLED STUDY FOR ADAM

ADAM.ADSL should not have more than one record per subject. Hence in ADaM as well same approach of DM and DC can be adopted.

The custom ADSL data set structure will be similar to DC which includes all the subject identifiers (SUBJID)

which can be used for analysis.

VALIDATION USING PINNACLE 21

In case of multiple enrollments, Pinnacle 21 may throw an error like 'different SUBJID within same USUBJID' or 'multiple EPOCH within same USUBJID', to which we can add the necessary explanation in reviewer's guide (RG) at the time of final submission.

CONCLUSION

This additional custom data set allows us to handle an array of challenging scenarios with added clarity and makes the programming (and inevitable troubleshooting) much more straightforward. Only rule of this custom domain is:

- SUBJID becomes the primary identifier for PARTICIPATIONS.
- USUBJID is the primary identifier for a PERSON.

REFERENCES

1. Italian CDISC User Group Network Annual Meeting. 2020. "*Updates on Handling Multiple Enrollments and Screenings Subjects in SDTM*." Accessed July 25, 2021.

   7th CDISC Italian User Network Day (Virtual) - Italian User Network - Wiki

2. Kelly, Kristen and Hamidi, Mike. 2019. "*Considerations When Representing Multiple Subject Enrollments in SDTM*." PharmaSUG 2019, DS-146.

   https://www.lexjansen.com/pharmasug/2019/DS/PharmaSUG-2019-DS-146.pdf

3. Matt Metherell 2021 "*A case of mistaken ID: reimagining rescreenings and reenrollments*." PharmaSUG China 2021 - Paper 102X-A6J9C8F9E9.

   A case of mistaken ID: reimagining rescreenings and reenrollments (pharmasug.org)

ACKNOWLEDGMENTS

The author would like to extend her sincere thanks to Ephicacy Lifescience Analytics for giving her an opportunity to write this paper. Any brand and product names are trademarks of their respective companies.

CONTACT INFORMATION

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

Vibhavari Honrao, Manager - Biostatistics & Programming
Ephicacy Lifescience Analytics
Bengaluru, India
E-mail: vibhavari.honrao@ephicacy.com
www.ephicacy.com