PharmaSUG China 2021 - Paper 102X-A6J9C8F9E9

A case of mistaken ID: reimagining rescreenings and reenrollments

Matt Metherell, PHASTAR, London

ABSTRACT

When a subject is able to re-enroll across multiple studies within a single project, it can be difficult to keep track of this one subject while maintaining CDISC compliance and making it clear which observations were recorded pre- and post-rescreening. If we need to copy observations between studies as well, this can multiply the complications, and the guidance currently in place could be improved. This paper proposes the solution we arrived upon for a three-study project with multiple rescreenings within and between each study. We are proud to suggest a new variable: **SUBJIDC**, *Subject Identifier as Collected*.

INTRODUCTION

There are two main pieces of guidance for subjects with multiple enrollments or screenings, but arguably they contradict each other. The SDTM IG v3.3 is clear that for each subject "only one DM record should be submitted for the subject". Conversely, the FDA Technical Conformance Guide states that "the subject's SUBJID should be different for each unique screening or enrollment." These positions are incompatible because on the one hand, if we have one single row in DM, we can't capture all values of SUBJID in DM. Conversely, if we have one single value in DM, this should be a definitive value that can be merged on to any dataset by USUBJID, but as SUBJID can vary with the observation, this is not the case.

Therefore to square this circle we suggest a new variable: SUBJIDC, Subject Identifier as Collected. This will be applied to every observation in domains other than DM as a supplemental variable to show which subject identifier the observation was collected under. This leaves SUBJID in DM as a single, final, subject ID, which we also include as a supplemental variable, to allow for quick comparisons that we will detail shortly.

USUBJID VS SUBJID VS SUBJIDP VS SUBJIDC

Our approach begins by standardizing the value of USUBJID as the earliest study- and subject identifiers that a given subject received. If they can only pass screening once, the subject ID that successfully screens is kept as SUBJID in DM, and as a supplemental variable in all other datasets for easy reference. Otherwise, the 'final' assigned subject ID is kept. All other subject identifiers are stored in SUPPDM as Previous Subject Identifiers (SUBJIDP, SUBJIDP2 etc). We chose to go 'back in time' so that SUBJIDP was the most recent prior subject ID, SUBJIDP2 was the one before that and so on, but this was largely out of convenience given the way the data was captured on the CRF. Then the Subject identifier captured on a given CRF page or eDT is stored as SUBJIDC (Subject Identifier as Collected) in the relevant SUPP domain. This value of SUBJIDC must match either DM.SUBJID or one of the SUBJIDPx's in SUPPDM.

We can summarise these variables in Table 1, which we opted to include in the Reviewer's Guide:

Variable	Occurrence	In DM?	Other domains?	Derivation
USUBJID	1 per project	Yes	Yes	The earliest study and earliest subject identifier
SUBJID	1 per study	Yes	SUPPxx	The final subject identifier
SUBJIDC	1 per screening	No	SUPPxx	The subject identifier at the time of a given observation
SUBJIDP*y*	1 per screening	SUPPDM	No	Previous subject identifier. (extra variables can be added as appropriate)

Table 1. Summary of subject identifier variables

This allows us to quickly identify if a subject has rescreened, and if an observation is from their current screening:

SUBJID = SUBJIDC: Observation collected during or after the final screening period.

SUBJID ≠ SUBJIDC: Observation collected during a prior screening.

DM.SUBJIDP populated: Subject rescreened.

DM.SUBJIDP not populated: Subject has not rescreened.

Second portion of USUBJID=SUBJID: Subject has not rescreened.

Furthermore, when looking at SDTM data it can be difficult to trace back to observations in the raw data using USUBJID alone, as it may not contain the relevant subject ID if a subject has had multiple screenings. Therefore SUBJIDC becomes invaluable when troubleshooting, especially if USUBJID is too long to display in a proc compare.

STUDYID VS STUDYIDP VS STUDYIDC

If subjects can enroll for multiple studies, and perhaps some observations need to be copied between these studies, it can be useful to treat the study identifier in a similar way.

We standardize the value of STUDYID as the identifier of the study we are reporting. Therefore a subject may be associated with multiple STUDYIDs across a project. All study identifiers associated with previous subject identifiers are stored in SUPPDM as Previous Study Identifiers (STUDYIDP, STUDIDP2 etc). This means that STUDYIDP was the study in which SUBJIDP was the subject ID, STUDIDP2 was the study in which STUBJIDP2 was assigned and so on.

Then the Study identifier captured on a given CRF page or eDT is stored as STUDYIDC (Study Identifier as Collected) in SUPPxx, and will have to match either STUDYID or one of the previous study identifiers in SUPPDM.

We can summarise these variables in Table 2, which we again decided to include in the Reviewer's Guide:

Variable	Occurrence	In DM?	Other domains?	Derivation
STUDYID	1 per study	Yes	Yes	The final study identifier
STUDYIDC	1 per screening	No	SUPPxx	The study identifier for a given observation
STUDYIDP/ STUDIDP*y*	1 per screening	SUPPDM	No	Previous study identifier. (extra variables can be added as appropriate)

Table 2 Summary of study identifier variables

This again allows us to make some quickly deductions about an observation:

STUDYID = STUDYIDC: Observation collected during the current study.

STUDYID ≠ STUDYIDC: Observation collected during a different study.

DM.STUDYID = SUPPDM.STUDYIDP/STUDIDPy: Subject rescreened within current study.

DM.STUDYID ≠ SUPPDM.STUDYIDP/STUDIDPy: Subject rescreened from a different study.

SUPPDM.STUDYIDP null: Subject has not rescreened.

EXAMPLES

DEMOGRAPHICS

Let's take a look at how this would work in a fictional study in Table 3.

STUDYID	DOMAIN	USUBJID	SUBJID		STUDYIDP	SUBJIDP
STUDYA	DM	STUDYA-10001	10001			
STUDYA	DM	STUDYA-10002	10100		STUDYA	10002
STUDYA	DM	STUDYA-10003	10003			

Table 3 DM domain for STUDYA

In STUDYA, the subject on row 2 re-entered the study, gaining a new SUBJID. We know it was from within the same study as STUDYIDP = STUDYID.
In order to see how subjects move between studies, we need a second fictional study, such as in Table 4.

STUDYID	DOMAIN	USUBJID	SUBJID		STUDYIDP	SUBJIDP
STUDYB	DM	STUDYA-10003	20002		STUDYA	10003
STUDYB	DM	STUDYB-20001	20001			
STUDYB	DM	STUDYB-20003	20100		STUDYB	20003

Table 4 DM domain for STUDYB

In STUDYB, the first row is a subject from STUDYA, as the value of STUDYIDP informs us. The value of USUBJID must be consistent between studies for a given subject, and we can see that "STUDYA-10003" appears in both datasets above. When we order by USUBJID, which begins with STUDYID, then all subjects who entered from a previous study will be grouped together at the top of the dataset. In addition, we know that if STUDYID does not match the start of USUBJID, then we know they must have rescreened between studies.
We can also see that row 3 re-entered from within STUDYB, as STUDYID=STUDYIDP. We will use this STUDYB in all our subsequent examples.

ECG RESULTS

Sometimes we may use observations captured under previous user or study IDs. Using an example for the same subjects' observations in the EG domain, we would have something like Table 5 below:

STUDYID	DOMAIN	USUBJID	EGSEQ	EGTESTCD	EGSTRESC		SUBJID	STUDYIDC	SUBJIDC
STUDYB	EG	STUDYA-10003	1	INTP	NORMAL		20002	STUDYA	10003
STUDYB	EG	STUDYA-10003	2	INTP	NORMAL		20002	STUDYB	20002
STUDYB	EG	STUDYA-10003	3	INTP	NORMAL		20002	STUDYB	20002
STUDYB	EG	STUDYA-10003	4	INTP	NORMAL		20002	STUDYB	20002
STUDYB	EG	STUDYB-20001	1	INTP	NORMAL		20001	STUDYB	20001
STUDYB	EG	STUDYB-20001	2	INTP	NORMAL		20001	STUDYB	20001
STUDYB	EG	STUDYB-20001	3	INTP	NORMAL		20001	STUDYB	20001
STUDYB	EG	STUDYB-20003	1	INTP	NORMAL		20100	STUDYB	20003
STUDYB	EG	STUDYB-20003	2	INTP	NORMAL		20100	STUDYB	20100
STUDYB	EG	STUDYB-20003	3	INTP	NORMAL		20100	STUDYB	20100

Table 5 EG domain for STUDYB

This shows that SUBJIDC can change between observations, but SUBJID cannot, as each subject has one value per study. We can see that SUBJIDs 20002 and 20100 both have observations performed under a previous ID because SUBJID $\neq$ SUBJIDC (red highlighting). In the case of SUBJID 20002, this first observation came from a previous study, confirmed by STUDYID $\neq$ STUDYIDC (blue highlighting). Inferring this from a single observation, without having to refer back to DM or DS can really save a lot of time.

ALTERNATIVE SOLUTIONS

The CDISC working group on Multiple Subject Instances are suggesting a new domain: “Demographics as Collected”, abbreviated as DC. It has a similar structure to DM but with one row per screening/enrollment, and aims to prevent too much additional data being stored in SUPPDM. For example, if variables such as AGE, SITE, ARM etc vary between screenings, this could quickly become unwieldy if we were to multiply these variables for each previous screening. The group is hoping for inclusion in version 4 of the SDTM Implementation Guide.

DC AND DM WORKING TOGETHER

The proposed DC domain allows information collected at each screening to be grouped together on a single line. In the example in Table 6, RFICDTC is collected at every screening, whereas SEX is only collected at the first, and RFSTDTC is only collected at the last:

STUDYID	DOMAIN	USUBJID	SUBJID	DCSEQ	RFICDTC	RFSTDTC	SEX
STUDYB	DC	STUDYA-10003	10003	1	2020-01-01		M
STUDYB	DC	STUDYA-10003	20002	2	2021-01-01	2021-01-28	
STUDYB	DC	STUDYB-20001	20001	1	2021-02-01	2021-02-28	F
STUDYB	DC	STUDYB-20003	20003	1	2021-03-01		F
STUDYB	DC	STUDYB-20003	20100	2	2021-04-01	2021-04-28	

Table 6 Standard DC domain for STUDYB

Their proposed solution uses an incrementing SUBJID in DC, keeping the final one in DM as in Table 7:

STUDYID	DOMAIN	USUBJID	SUBJID		STUDYIDP	SUBJIDP
STUDYB	DM	STUDYA-10003	20002		STUDYA	10003
STUDYB	DM	STUDYB-20001	20001			
STUDYB	DM	STUDYB-20003	20100		STUDYB	20003

Table 7 Reminder of DM domain for STUDYB

However, DC could be complemented by our suggested SUPPDM variables (shown below in Table 8), as they provide the viewer with a quick orientation of observations. Our proposal keeps SUBJID constant in DC while SUBJIDC increments, meaning that SUBJID doesn’t mean something different depending on the context:

STUDYID	DOMAIN	USUBJID	SUBJID	DCSEQ	RFICDTC	RFSTDTC	SEX	SUBJIDC
STUDYB	DC	STUDYA-10003	20002	1	2020-01-01		M	10003
STUDYB	DC	STUDYA-10003	20002	2	2021-01-01	2021-01-28		20002
STUDYB	DC	STUDYB-20001	20001	1	2021-02-01	2021-02-28	F	20001
STUDYB	DC	STUDYB-20003	20100	1	2021-03-01		F	20003
STUDYB	DC	STUDYB-20003	20100	2	2021-04-01	2021-04-28		20100

Table 8 Augmented DC domain for STUDYB

Essentially, we treat SUBJIDC in the way that the FDA conformance guide suggested handling SUBJID (“the subject’s SUBJID should be different for each unique screening or enrollment”), while also satisfying the SDTM IG requirement that “only one DM record should be submitted for the subject”.

BEYOND SDTM

All of the above talks about SDTM datasets, but if SUBJIDPx and STUDYIDP/STUDIDPx were added to the core variables that typically already include STUDYID, USUBJID and SUBJID, then we can perform the same comparisons by keeping STUDYIDC and SUBJIDC in each relevant ADaM dataset. We can then easily include this information in listings.

AN EXAMPLE LISTING

In Output 1, we use SUBJID in a spanning line that summarises the subject, and then list SUBJIDC in the Subject ID column, so we can see which observations were captured under a previous ID (red text), and therefore a previous screening attempt.

Subject ID	Treatment	Visit Day/Date	ECG - Interpretation
20001, Age = 30 years, Sex=Female, Weight = 60 kg, BMI = 25.5 kg/m^2			
20001	Treatment A	-28/2021-02-01	Normal
		1/2021-02-28	Normal
		85/2021-05-23	Normal
20002, Age = 40 years, Sex=Male, Weight = 70 kg, BMI = 27.5 kg/m^2			
10003	Treatment A	-394/2020-01-01	Normal
20002	Treatment A	-28/2021-01-01	Normal
		1/2021-01-28	Normal
		85/2021-04-22	Normal
20100, Age = 50 years, Sex=Female, Weight = 80 kg, BMI = 30.0 kg/m^2			
20003	Treatment B	-28/2021-04-01	Normal
20100	Treatment B	1/2021-04-28	Normal
		85/2021-07-21	Normal

Output 1. ECG Listing for STUDYB

CONCLUSION

These additional variables allow us to handle a multitude of challenging scenarios with added clarity, and make the programming (and inevitable troubleshooting) much more straightforward.

RECOMMENDED READING

Kelly, Kristen and Hamidi, Mike. 2019. “Considerations When Representing Multiple Subject Enrollments in SDTM.” *PharmaSUG 2019*, DS-146. Available at https://www.lexjansen.com/pharmasug/2019/DS/PharmaSUG-2019-DS-146.pdf

Italian CDISC User Group Network Annual Meeting. 2020. “Updates on Handling Multiple Enrollments and Screenings Subjects in SDTM.” Accessed July 25, 2021. Available at https://wiki.cdisc.org/pages/viewpage.action?pageId=113577697&preview=/113577697/113577703/Updates on Handling Multiple Enrollments and Screenings Subjects in SDTM - ClinBuild - Eanna Kiely - V1.0.pdf

CONTACT INFORMATION

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

Matt Metherell
PHASTAR
matthew.metherell@phastar.com