



March 8th–11th
Renaissance Orlando at
SeaWorld Florida

US 2020

The Clinical Data
Science Conference

Handling Duplicate Records in SDTM for Effective Resolution and Explanation in Data Submission

Paper DS03

Varun Debbeti, Cytel, Waltham, USA



Table of Contents

- INTRODUCTION
- STRUCTURE OF A DATASET
- CLASSIFICATION OF DUPLICATE RECORDS
- SUMMARY/CONCLUSION
- REFERENCES
- QUESTIONS

INTRODUCTION

- The purpose of SDTMs in submission is to create an interchange standard in the industry for an effective submission process
- FDA reviewers are trained in the principles of standardized datasets and the use of standard tools to work with data effectively with less preparation time
- Duplicate records in data is one such deviation of standard
- Duplicate records represent contradictory information and make difficult to summarize and interpret results
- This Paper discusses the standard structure of domains, and the concept of natural and surrogate keys to classify duplicate records for easier understanding and prescribes actions to resolve issues with duplicate records

STRUCTURE OF A DATASET – SDTM DOMAIN

- ❑ A unique record in **SDTM domain** has unique values in at least one of selected variables known as **key variables**

Example:

STUDYID	USUBJID	AEDECOD	AESTDTC	AEENDTC
AAA0000	001-001	HEADACHE	2012-01-01	2012-01-01
AAA0000	001-001	HEADACHE	2012-01-02	
AAA0000	001-001	FEVER	2020-01-02	

Dataset	Description	Class	Structure	Purpose	Keys*	Location
DM	Demographics	Special Purpose Domains	One record per subject	Tabulation	STUDYID, USUBJID	dm.xpt
CO	Comments	Special Purpose Domains	One record per comment per subject	Tabulation	STUDYID, USUBJID, COSEQ	co.xpt
CM	Concomitant Medications	Interventions	One record per medication occurrence or dosing interval per subject	Tabulation	STUDYID, USUBJID, CMTRT, CMSTDTC	cm.xpt
AE	Adverse Events	Events	One record per adverse event per subject	Tabulation	STUDYID, USUBJID, AEDECOD, AESTDTC	ae.xpt
LB	Laboratory Test Results	Findings	One record per analyte per planned time point reference per visit per subject	Tabulation	STUDYID, USUBJID, LBTESTCD, LBSPEC, VISITNUM, LBTPTREF, LBTPTNUM	lb.xpt

A table from SDTM IG 3.2 showing key variables list for each domains

STRUCTURE OF DATASET – FDA Expectations

FDA Business Rules

FDA released Final version final version of Business rules in December 2017

Rule ID FDAB021 describes FDAs expectation on duplicate records

P21 Rules

Pinnacle 21 has two corresponding rules describing the structure of domains to handle duplicate records

version 1.2, finalized December 2017

FDA Business Rule ID	FDA Business Rule	FDA Validator Rule	Domains	SDTM 3.1.2	SDTM 3.1.3	SDTM 3.2	SEND 3.0
FDAB021	Duplicate records are not permitted (as constrained by the unique key in the underlying standard).	The structure of Findings class domains should be one record per Finding Result per subject. No Finding Result with the same Test Short Name (--TESTCD) for the same Subject (USUBJID) and the same Collection Date (--DTC) are expected.	FINDINGS	X	X	X	X
		The structure of Events class domains should be one records per Event per subject. No Events with the same Collected Term (--TERM), Decoded Term (--DECOD), Category (--CAT), Subcategory (--SCAT), Severity (--SEV), and Toxicity Grade (--TOXGR) values for the same Subject (USUBJID) and the same Start Date (--STDTC) are expected.	EVENTS	X	X	X	

Pinnacle 21	Publisher ID	Message	Description	Category	Severity
SD1117	FDAC212	Duplicate records	The structure of Findings class domains should be one record per Finding Result per subject. No Finding Result with the same Test Short Name (--TESTCD) for the same Subject (USUBJID) and the same Collection Date (--DTC) are expected.	Consistency	Warning
SD1201	FDAC213	Duplicate records in domain	The structure of Events class domains should be one records per Event per subject. No Events with the same Collected Term (--TERM), Decoded Term (--DECOD), Category (--CAT), Subcategory (--SCAT), Severity (--SEV), and Toxicity Grade (--TOXGR) values for the same Subject (USUBJID) and the same Start Date (--STDTC) are expected.	Consistency	Warning

STRUCTURE OF DATASET – Key Variables

Standard Keys

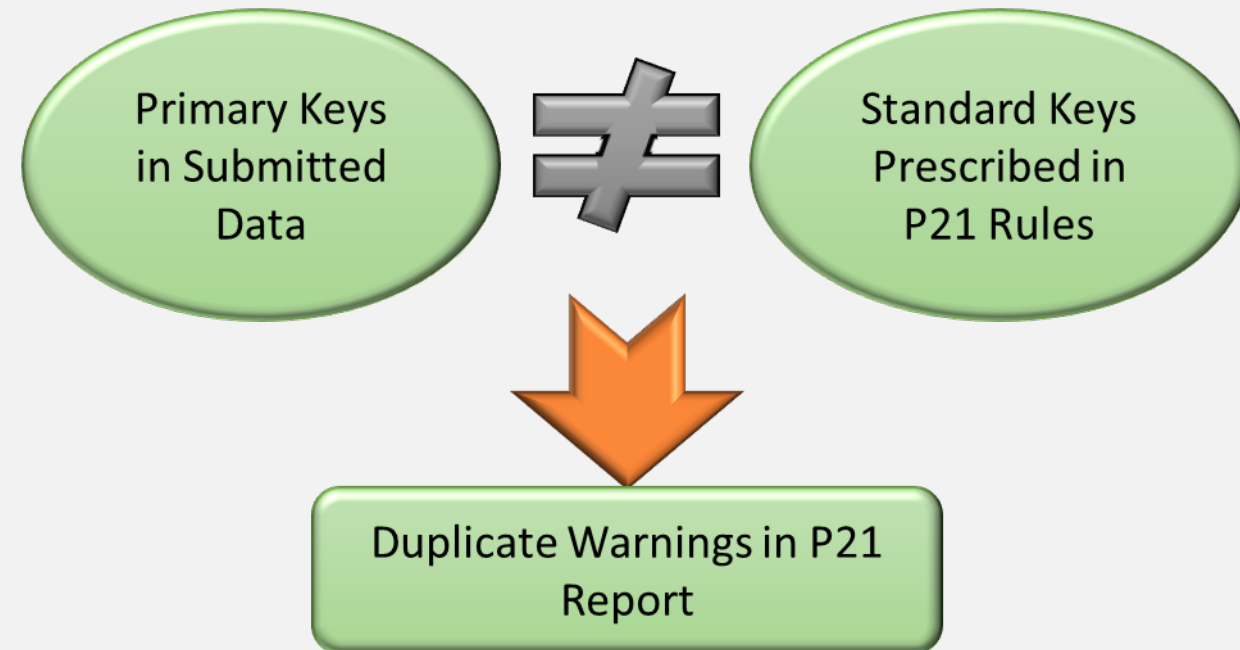
- Key variables listed in P21 rules to define each class of SDTM domain

Primary Keys

- Key variables listed in define.xml submitted for reviewers to understand the structure of submitted data.

They are further divided in to

1. *Natural Key: A natural key is a piece of data that uniquely identify that entity and distinguish it from any other row in the table. Natural keys already exist in the data as opposed to surrogate keys. However, as the natural keys are associated with business requirements, they tend to change when the requirements change. Example: In clinical trials data, addition of method or position or location becomes new key, which did not exist in previous studies.*
2. *Surrogate Key: A surrogate key is a single-part, artificially established identifier for a record. Surrogate key assignment is a special case of derived data, one where a portion of the primary key is derived. A surrogate key does not change with business needs. In addition, the key depends on only one field, so it's compact. A common way of deriving surrogate key values is to assign integer values sequentially. The --SEQ variable in the SDTM datasets is an example of a surrogate key for most datasets;*



CLASSIFICATION OF DUPLICATE RECORDS

Based on usage of Natural and Surrogate keys to distinguish records, this paper classifies duplicate records to 5 types

1. True double entries per key variables and improper use of surrogate keys
2. True double entries per key variables and improper use of additional natural keys
3. Duplicate records due to insufficient natural keys
4. Incorrect primary keys/incorrect mapping of data
5. Duplicate records due to missing values in keys

1. True double entries per key variables and improper use of surrogate keys

The data has two or more records with most of the variables having same information.

The sponsor chose to add surrogate keys to the key list to make duplicate records appear unique, but still the data is incorrect, and lacks accuracy.

Standard key variables to identify unique records in LB: STUDYID, USUBJID, LBTESTCD, LBSPEC, VISITNUM, LBDTC

Sponsor chosen key variables: STUDYID, USUBJID, LBREFID, LBTESTCD, VISITNUM, LBDTC

USUBJID	LBSEQ	LBTEST	LBORRES	LBORRESU	LBDTC
005-005	1	Glucose	155	mg/dL	2015-05-12T09:40
005-005	2	Glucose	155	mg/dL	2015-05-12T09:40

Example 1: Actual duplicate information. Sponsor chose to use LBSEQ, a surrogate key to make it appear unique

USUBJID	LBREFID	LBTESTCD	LBORRES	LBORRESU	LBNRIND	LBFAST	VISIT	LBDTC
001-001	1	GLUC	95	mg/dL	NORMAL	N	UNSCHEDULED	2011-03-06T10:10
001-001	2	GLUC	185	mg/dL	HIGH	N	UNSCHEDULED	2011-03-06T10:10
001-001	3	GLUC	135	mg/dL	HIGH	N	UNSCHEDULED	2011-03-06T10:10
001-001	4	GLUC	210	mg/dL	HIGH	N	UNSCHEDULED	2011-03-06T10:10
001-001	5	GLUC	67	mg/dL	NORMAL	N	UNSCHEDULED	2011-03-06T10:10
001-001	6	GLUC	85	mg/dL	NORMAL	N	UNSCHEDULED	2011-03-06T10:10

Example 2: Different results collected at same date/time. Data is highly suspicious, but sponsor chose to explain uniqueness using LBREFID, a surrogate variable

Resolution/Explanation: Data should be cleaned before submitting to FDA as there is no meaningful explanation to the duplicity

2. True double entries per key variables and improper use of additional natural keys

The data has two or more records with most of the variables having same information.

The sponsor chose to add natural keys to the list to make duplicate records appear unique, however, the reason for adding natural keys is not explained in a meaningful way.

Standard key variables to identify unique records in LB: STUDYID, USUBJID, LBTESTCD, LBSPEC, VISITNUM, LBDTC

Sponsor chosen key variables: STUDYID, USUBJID, LBTESTCD, LBORRES, LBSTAT, VISITNUM, LBDTC

USUBJID	LBTEST	LBORRES	LBORRESU	VISIT	LBDTC
005-005	Hemoglobin	155	mg/dL	VISIT3	2015-05-12T09:40
005-005	Hemoglobin	147	mg/dL	VISIT3	2015-05-12T09:40

Example 1: Two different results on same time point

USUBJID	LBTEST	LBORRES	LBORRESU	LBSTAT	VISIT	LBDTC
005-005	Albumin	4.2	g/dL		UNSCHEDULED	2015-05-12
005-005	Albumin			NOT DONE	UNSCHEDULED	2015-05-12
005-005	Sodium	134	mmol/L		UNSCHEDULED	2015-05-12
005-005	Sodium			NOT DONE	UNSCHEDULED	2015-05-12

Example 2: Two different results (one NOT DONE, and one actual result) on same time point

Resolution/Explanation: The proper resolution would be to remove duplicate data by data management team. If the database is locked, the proper explanation would be to reference any comments entered by the laboratory staff, as to why the second test was performed at same timepoint. (e.g., instrument calibration not performed, etc.)

3. Duplicate records due to insufficient natural keys

In this category, sponsor collected duplicate records per business meaning

USUBJID	LBREFID	LBTEST	LBORRES	LBORRESU	LBNAM	LBDTC
001-001	12345	Albumin	4.2	g/dL	Central Lab	2011-03-06T10:10
001-001	12346	Albumin	4.2	g/dL	Local Lab	2011-03-06T10:10

Example 1: Results from different labs

Standard key variables to identify unique records in LB: STUDYID, USUBJID, LBTESTCD, LBSPEC, VISITNUM, LBDTC

Sponsor chosen key variables: STUDYID, USUBJID, LBTESTCD, LBNAM, VISITNUM, LBDTC

Resolution/Explanation: It is a genuine case of duplicity, which can be explained by adding LBNAM to the primary key list. As the addition of LBNAM is driven by the business requirement, and the protocol design, both records should be accepted to be present in the data.

3. Duplicate records due to insufficient natural keys (continued..)

In this category, sponsor collected duplicate records per business meaning

USUBJID	TRSEQ	TRGRP	TRLNKID	TRTESTCD	TRTEST	TRORRES	TRMETHOD	TREVAL	TRDTC
002-002	2	TARG	T01	LDIAM	Longest Diameter	11	PET/CT SCAN	INVESTIGATOR	2015-10-22
002-002	3	TARG	T02	LDIAM	Longest Diameter	18	PET/CT SCAN	INVESTIGATOR	2015-10-22
002-002	4	TARG	T03	LDIAM	Longest Diameter	36	PET/CT SCAN	INVESTIGATOR	2015-10-22
002-002	5	TARG	T04	LDIAM	Longest Diameter	81	PET/CT SCAN	INVESTIGATOR	2015-10-22
002-002	6	TARG	T05	LDIAM	Longest Diameter	17	PET/CT SCAN	INVESTIGATOR	2015-10-22
002-002	7	TARG	T06	LDIAM	Longest Diameter	19	PET/CT SCAN	INVESTIGATOR	2015-10-22

Example 2: Results from a split tumor

Standard key variables to identify unique records in TR: STUDYID, USUBJID, TRTESTCD, TRMETHOD, TREVAL, VISITNUM, TRDTC

Sponsor chosen key variables: STUDYID, USUBJID, TRLNKID, TRTESTCD, TRMETHOD, TREVAL, VISITNUM, TRDTC

Domain	Record	Count	Variables	Values	Pinnacle 21	Publisher ID	Message	Category	Severity
TR	7		TRMETHOD, VISITNUM, TRDTC, TRTESTCD, USUBJID, TREVAL	PET/CT SCAN, -28, 2015-10-22, LDIAM, 002-002, INVESTIGATOR	SD1117	FDAC212	Duplicate records	Consistency	Warning
TR	8		TRMETHOD, VISITNUM, TRDTC, TRTESTCD, USUBJID, TREVAL	PET/CT SCAN, -28, 2015-10-22, LDIAM, 002-002, INVESTIGATOR	SD1117	FDAC212	Duplicate records	Consistency	Warning
TR	9		TRMETHOD, VISITNUM, TRDTC, TRTESTCD, USUBJID, TREVAL	PET/CT SCAN, -28, 2015-10-22, LDIAM, 002-002, INVESTIGATOR	SD1117	FDAC212	Duplicate records	Consistency	Warning
TR	10		TRMETHOD, VISITNUM, TRDTC, TRTESTCD, USUBJID, TREVAL	PET/CT SCAN, -29, 2015-10-22, LDIAM, 002-002, INVESTIGATOR	SD1117	FDAC212	Duplicate records	Consistency	Warning
TR	11		TRMETHOD, VISITNUM, TRDTC, TRTESTCD, USUBJID, TREVAL	PET/CT SCAN, -28, 2015-10-22, LDIAM, 002-002, INVESTIGATOR	SD1117	FDAC212	Duplicate records	Consistency	Warning

Figure 4. Pinnacle 21 report generated on data shown in Table 7 above.

Resolution/Explanation: It is a genuine case of duplicity, which can be explained by adding TRLNKID to the primary key list. TRLNKID is a needed additional natural key variable to make the records unique as they all represent separate split tumors associated with a parent tumor. TRLNKID, a natural key variable could be added to the key list, and all records should be accepted to be present in the data

4. Incorrect primary keys/incorrect mapping of data

Incorrect primary keys: Pinnacle 21 assumes FA as a Findings domain, and applies the same key variables of a findings domain to identify unique records. Whereas, FA is a FINDINGS ABOUT domain, with FAOBJ variable being an important variable

Example: SDTM IG example above has more than one record per subject per FATESTCD per FACAT, per VISITNUM. When the sponsor mapped data similar to the example, the Pinnacle 21 throws Duplicate Warnings in the P21 report.

Standard key variables by IG to identify unique records in FA: STUDYID, USUBJID, FATESTCD, FAOBJ, VISITNUM, FATPTREF, FATPTNUM.

Sponsor chosen key variables: STUDYID, USUBJID, FATESTCD, FAOBJ, FACAT, FADTC.

Standard key variables used by P21 to identify unique records in FA: STUDYID, USUBJID, FATESTCD, FACAT, FASCAT, VISITNUM, FADTC

RESOLUTION/EXPLANATION: As it is following the structure prescribed by IG, it can be explained the same in reviewers guide.

fa.xpt

Row	STUDYID	DOMAIN	USUBJID	FASEQ	FASPID	FATESTCD	FATEST	FAOBJ	FACAT	FAORRES	FADTC	FATPT
1	ABC	FA	ABC-123	1	90567	SEV	Severity/Intensity	Migraine	MIGRAINE SYMPTOMS	SEVERE	2007-05-16	5M PRE-DOSE
2	ABC	FA	ABC-123	2	90567	OCCUR	Occurrence	Sensitivity To Light	MIGRAINE SYMPTOMS	Y	2007-05-16	5M PRE-DOSE
3	ABC	FA	ABC-123	3	90567	OCCUR	Occurrence	Sensitivity To Sound	MIGRAINE SYMPTOMS	N	2007-05-16	5M PRE-DOSE
4	ABC	FA	ABC-123	4	90567	OCCUR	Occurrence	Nausea	MIGRAINE SYMPTOMS	Y	2007-05-16	5M PRE-DOSE
5	ABC	FA	ABC-123	6	90567	OCCUR	Occurrence	Aura	MIGRAINE SYMPTOMS	Y	2007-05-16	5M PRE-DOSE
6	ABC	FA	ABC-123	7	90567	SEV	Severity/Intensity	Migraine	MIGRAINE SYMPTOMS	MODERATE	2007-05-16	30M POST-DOSE
7	ABC	FA	ABC-123	8	90567	OCCUR	Occurrence	Sensitivity To Light	MIGRAINE SYMPTOMS	Y	2007-05-16	30M POST-DOSE
8	ABC	FA	ABC-123	9	90567	OCCUR	Occurrence	Sensitivity To Sound	MIGRAINE SYMPTOMS	N	2007-05-16	30M POST-DOSE
9	ABC	FA	ABC-123	10	90567	OCCUR	Occurrence	Nausea	MIGRAINE SYMPTOMS	N	2007-05-16	30M POST-DOSE
10	ABC	FA	ABC-123	12	90567	OCCUR	Occurrence	Aura	MIGRAINE SYMPTOMS	Y	2007-05-16	30M POST-DOSE
11	ABC	FA	ABC-123	13	90567	SEV	Severity/Intensity	Migraine	MIGRAINE SYMPTOMS	MILD	2007-05-16	90M POST-DOSE
12	ABC	FA	ABC-123	14	90567	OCCUR	Occurrence	Sensitivity To Light	MIGRAINE SYMPTOMS	N	2007-05-16	90M POST-DOSE
13	ABC	FA	ABC-123	15	90567	OCCUR	Occurrence	Sensitivity To Sound	MIGRAINE SYMPTOMS	N	2007-05-16	90M POST-DOSE
14	ABC	FA	ABC-123	16	90567	OCCUR	Occurrence	Nausea	MIGRAINE SYMPTOMS	N	2007-05-16	90M POST-DOSE
15	ABC	FA	ABC-123	18	90567	OCCUR	Occurrence	Aura	MIGRAINE SYMPTOMS	N	2007-05-16	90M POST-DOSE

USUBJID	FASEQ	FATESTCD	FATEST	FAOBJ	FACAT	FAORRES	FADTC
002-002	2	OCCUR	Occurrence	BCL-2 alterations/overexpression	DIAGNOSIS HISTORY	Y	2014-06-18
002-002	4	OCCUR	Occurrence	BCL-6 alterations/overexpression	DIAGNOSIS HISTORY	Y	2014-06-18
002-002	5	OCCUR	Occurrence	C-MYC alterations/overexpression	DIAGNOSIS HISTORY	N	2014-06-18

Domain	Record	Count	Variables	Values	Pinnacle 21	Publisher	Message	Category	Severity
FA		4	FATESTCD, VISITNUM, FACAT, FADTC, FASCAT, USUBJID	OCCUR, -28, DIAGNOSIS HISTORY, 2014-06-18, null, 002-002	SD1117	FDAC212	Duplicate records	Consistency	Warning
FA		5	FATESTCD, VISITNUM, FACAT, FADTC, FASCAT, USUBJID	OCCUR, -28, DIAGNOSIS HISTORY, 2014-06-18, null, 002-002	SD1117	FDAC212	Duplicate records	Consistency	Warning

4. Incorrect primary keys/incorrect mapping of data (continued....)

Incorrect Mapping of data: A sponsor chose to map data to a custom Findings domain, instead it would have been a better fit in FA or SR domain.

In this example, sponsor mapped Sub-Location as separate TESTCD records in the ZA custom finding domain although the Wheal Diameter and Sublocation are related to the same objective “Histamine Control 10 mg/mL”.

Also, the objective of test is mapped to ZACAT. Which causes duplicity as per standard key variables of a findings domain.

Standard key variables of a Findings domain: STUDYID, USUBJID, --TESTCD, --CAT, --VISIT, --DTC

Suggested Mapping structure , to avoid duplicity is: STUDYID USUBJID SRTESTCD SROBJ SRLOC QNAM.SRSUBLOC

RESOLUTION/EXPLANATION: The solution is to choose and remap the above data into a domain to make data structure standard by choosing SR and SUPPSR which are perfect for the Findings About an Event.

USUBJID	ZASEQ	ZATESTCD	ZATEST	ZACAT	ZAORRES	ZALOC	VISIT
002-002	1	WHEALDIA	Wheal Diameter	Histamine Control 10 mg/mL	5	BACK	VISIT1
002-002	2	WHEALDIA	Wheal Diameter	Histamine Control 10 mg/mL	4	BACK	VISIT1
002-002	3	WHEALDIA	Wheal Diameter	Histamine Control 10 mg/mL	5	BACK	VISIT1
002-002	4	WHEALDIA	Wheal Diameter	Histamine Control 10 mg/mL	5	BACK	VISIT1
002-002	5	SUBLOC	Sub-Location	Histamine Control 10 mg/mL	QUADRANT1	BACK	VISIT1
002-002	6	SUBLOC	Sub-Location	Histamine Control 10 mg/mL	QUADRANT2	BACK	VISIT1
002-002	7	SUBLOC	Sub-Location	Histamine Control 10 mg/mL	QUADRANT3	BACK	VISIT1
002-002	8	SUBLOC	Sub-Location	Histamine Control 10 mg/mL	QUADRANT4	BACK	VISIT1

Table 9. ZA domain mapped WHEALDIA and SUBLOC as separate ZATESTCD records

USUBJID	SRSEQ	SRTESTCD	SRTEST	SROBJ	SRORRES	SRLOC	VISIT
002-002	1	WHEALDIA	Wheal Diameter	Histamine Control 10 mg/mL	5	BACK	VISIT1
002-002	2	WHEALDIA	Wheal Diameter	Histamine Control 10 mg/mL	4	BACK	VISIT1
002-002	3	WHEALDIA	Wheal Diameter	Histamine Control 10 mg/mL	5	BACK	VISIT1
002-002	4	WHEALDIA	Wheal Diameter	Histamine Control 10 mg/mL	5	BACK	VISIT1

Table 10. SR domain remapped from the ZA domain in Table 9

RDOMAIN	USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL	QORIG
SR	002-002	SRSEQ	1	SRSUBLOC	Sub-Location	QUADRANT1	CRF
SR	002-002	SRSEQ	2	SRSUBLOC	Sub-Location	QUADRANT2	CRF
SR	002-002	SRSEQ	3	SRSUBLOC	Sub-Location	QUADRANT3	CRF
SR	002-002	SRSEQ	4	SRSUBLOC	Sub-Location	QUADRANT4	CRF

Table 11. SUPPSR domain remapped from the ZA domain in Table 9

5. Duplicate records due to missing values in keys

The data appears to be duplicate per standard key variables due to the missing values in one or more of the standard key variables

Standard key variables to identify unique records in DA: STUDYID, USUBJID, DATESTCD, DADTC

Sponsor chosen key variables: STUDYID, USUBJID, DATESTCD, DADTC

USUBJID	DATESTCD	DATEST	DAORRES	DAORRESU	EPOCH	DADTC
001-001	DISPAMT	Dispensed Amount	68	mL	TREATMENT	
001-001	DISPAMT	Dispensed Amount	68	mL	RETREATMENT	

Resolution/Explanation: Based on the unique Epoch values (TREATMENT, RETREATMENT) associated for each record, it is appropriate to assume the protocol design required the drug to be dispensed at two different occasions during the study, i.e, during two Epochs separately. Data management should fix this by populating dates accordingly. If the database is locked, it is perfectly acceptable to explain in reviewers guide saying the dates were not available, but still the data is not duplicate.

SUMMARY/CONCLUSION

In summary, when duplicate records are identified in the data, sponsors should do

1. Make records unique to identify by adding necessary key variables

2. It should be explained meaningfully how the added variables are natural keys and are already present in the data and are necessary to have multiple records based on the protocol design of study. Below are conclusions for each of 5 categories described above.

CATEGORY	CONCLUSION
1. True double entry per key variables and Improper use of surrogate keys	These are true data entry errors, should be fixed before database lock. There is no meaningful explanation of the presence of duplicates in this category.
2. True double entry per key variables and improper use of additional natural keys	These are duplicate data entries, made to appear unique by adding natural keys, but not properly explained. If no meaningful explanation associated with added natural keys, sponsor should not expect the data to be accepted by reviewers.
3. Insufficient natural Keys	These duplicates are not true duplicates. Their collection is necessary as per protocol design. So, adding natural keys in define.xml will make records unique in a meaningful way.
4. Incorrect primary Keys/Incorrect mapping of data:	Incorrect mapping, should be fixed.
5. Missing values in keys	Missing data causes duplicity. They could be fixed by Data management or explained in reviewers guide.

REFERENCES

- [1] Duplicate records- it may be a good time to contact your data management team, PharmaSUG 2016 – TT18
- [2] Doi, Mary. “How Good is Your SDTM Data? Perspectives from JumpStart”. PhUSE CSS. March 2016. Available at <http://www.phusewiki.org/docs/CSS%202016%20Presentations/SDTM%20Mary%20Doi.pptx>
- [3] CDISC SDTM Implementation Guide. Available at <http://www.cdisc.org/sdtm>
- [4] Best Practice for Explaining Validation Results in the Study Data Reviewer’s guide, PhUSE EU Connect 2018 – DS06
- [5] Confusing Data Validation Rules Explained, PhUSE US Connect 2018 – RG03
- [6] FDA Business rules/Validator rules.
<https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm>
- [7] How to Prepare High-quality Metadata for Submission, PhUSE US Connect 2018 – SI12

QUESTIONS

