Paper DS03

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

Varun Debbeti, Cytel, Waltham, USA

ABSTRACT

One of the challenges of SAS® programmers is handling duplicate records in the raw/SDTM data. These are observed due to unclean data or programming inefficiency or due to the necessity to collect multiple results for special analytical or statistical purpose in the trial. In some cases, these are much harder to manage and can complicate analysis by producing unexpected outcomes and programming delays. This paper presents the background to identify different types of duplicate records by using the concept of Natural Keys and Surrogate Keys described in the SDTM IG 3.2. It also discusses FDA business rules, FDA Validator rules, corresponding Pinnacle21® (short form P21) Validator rules and warnings in P21 report that covers the duplicate records. The background is used to classify the most common types of duplicate records in to 5 categories. This paper then provides examples and prescribes standard set of actions to resolve the issue in each category.

INTRODUCTION

The Purpose of SDTMs in submission to create an interchange standard in the industry and for effective submission process. But, deviating the standards can cause delays in programming due to unexpected outcomes. Sometimes, it can be challenging for the reviewers as well. As per SDTM IG "*The availability of standard submission data will provide many benefits to regulatory reviewers. Reviewers can be trained in the principles of standardized datasets and the use of standard software tools, and thus be able to work with the data more effectively with less preparation time*." Having duplicate records in data is one such deviation of standard. Duplicate records represent contradictory information and make difficult to summarize and interpret results [2, 7]. This paper discusses the concept of natural keys and surrogate keys to identify unique records in SDTM data and classify duplicate records in to 5 categories. It also provides examples and prescribes standard set of actions to resolve these issues.

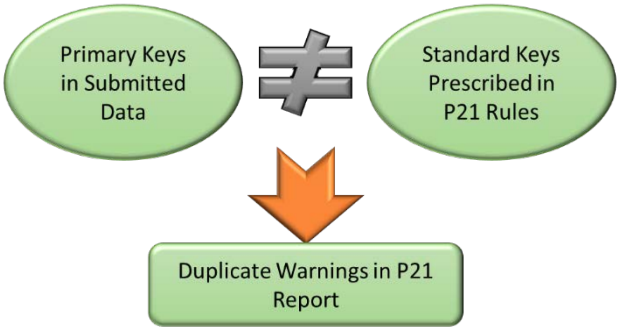

Primary Keys in Submitted Data
≠
Standard Keys Prescribed in P21 Rules
Duplicate Warnings in P21 Report

Figure 1. A drawing showing duplicate warnings in P21 report are generated when Primary keys in submitted data are not equal to the standard keys prescribed in P21/FDA rules.

STRUCTURE OF DATASET

What is a unique record in a general dataset vs an SDTM domain?

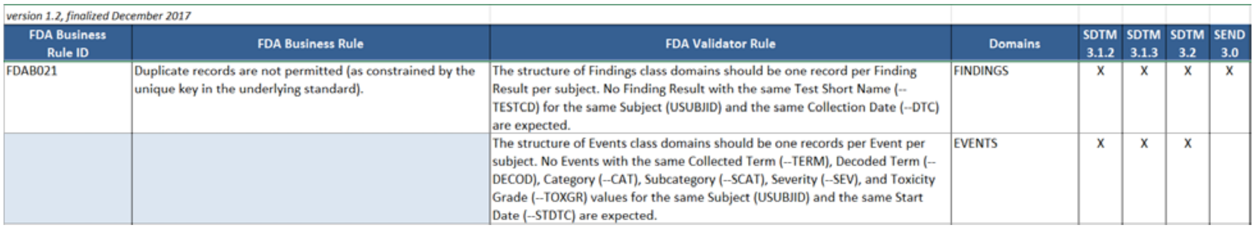
- In a general SAS dataset, a unique record has unique values in all variables of the dataset.
- However, in a SDTM domain, a unique record has unique values in selected variables, also known as Primary/Standard keys based on structure of a domain as per the SDTM model.

The structure of an SDTM domain is defined by using Key variables. SDTM IG 3.2 [3] says "*The define.xml that accompanies a submission should also describe each dataset that is included in the submission and describe the natural key structure of each dataset.*" For example, CM domain structure one record per medication per dose interval per subject is defined by key variables STUDYID, USUBJID, CMTRT, CMSTDTC.

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

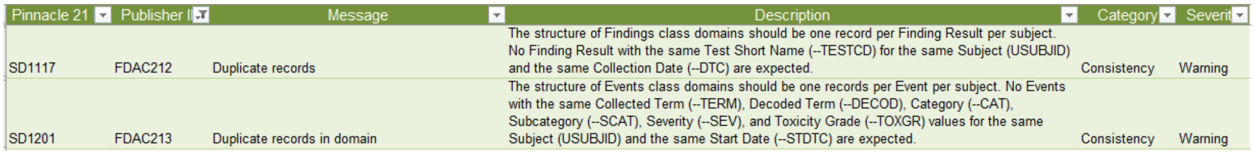
Figure 2. A table showing data-level metadata section of a standard define.xml listing DM, CO, SE domains etc.

FDA published (shown below in Fig. 2) business rules [6] in 2017 which precedes FDA Validator rules to provide the FDA's expectations on the uniqueness of data. One of the rules is FDAB021 regarding duplicate records in data. There are 2 corresponding FDA validator rules (Publisher ID is FDAC212 and FDAC213) describing the structure of different classes of domains. These validator rules are incorporated in to Pinnacle21 tools in the form of Pinnacle 21 rules SD1117 and SD1201. A standard P21 report run on SDTM domains will show warnings SD1117 and SD1201 when duplicate records are identified in the dataset.



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 II	Message	Description	Category	Severit
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

Figure 3. FDA Business rules, FDA validator rules (Above image), P21 rules (below image) that covers duplicate records.

KEY VARIABLES

To understand what type of duplicate records can occur one should know the variety of key variables. For the illustration purposes, this paper classifies key variables in to 1) standard keys 2) Primary keys.

- What are standard keys: Standard keys are the key variables listed in P21 rules to define each class of SDTM domain.
- What are primary keys: Primary keys are key variables listed in define.xml submitted for reviewers to understand the structure of submitted data. These keys should define uniqueness for records within a dataset and may define a record sort order. The naming of these keys should be consistent with the description of the structure in the Structure column. Primary keys are further divided in to two types per IG.
    - ❖ 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.
    - ❖ 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

As shown in Figure 1, when primary keys (as listed in define.xml) of the submitted SDTM data do NOT match with the standard keys prescribed in the Pinnacle 21 validator based on the FDA validator rules, the tool will show warnings in the validation report. A P21 warning for a duplicate record is an alert to the user for a potential duplicity of record w.r.t standard keys mentioned in the SDTM model. So, sponsors should not interpret this warning in a very strict meaning of the word duplicate. It is possible the duplicate records would have legitimate reasons. This paper considers some common situations in data that could lead to duplicate warnings, and then classifies them to 5 categories below. Some categories of warnings below would require fix in programming or requires a fix in database by data-management, and if the database is already locked, they could be explained in define.xml submitted as part of the data package. Note: Due to limitations in width of the page, only selective key variables columns are shown in each example below to make a meaningful example.

1. True double entry per key variables and Improper use of surrogate keys

In this category, the data has two or more records with most of the variables having same information [1]. And 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.

Example 1: Actual duplicate information, except for the sponsor assigned sequence variable (--SEQ) which cannot be considered as a natural key in LB domain. This is still a common case.

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

Table 1. Example of exact duplicate records

Resolution/Explanation: The obvious solution is for the data-management to remove duplicate in the database before the database is locked. In some cases, this may happen due to programming issue in SDTM program as well.

Example 2: Records that are only differentiated by a sponsor-defined variable.

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

Table 2. Example of exact duplicate records differentiated with sponsor-defined surrogate variable

Sponsors tend to interpret this rule with a very strict meaning of the word ‘duplicate’. Therefore, it is common to see incorrect explanations in the cSDRG for this rule. A typical explanation from a sponsor for this validation rule might be something like:

Check ID	Diagnostic Message	Severity	Dataset	Count (Issue Rate)	Explanation
SD117	Duplicate records	Warning	LB	320 (30.2%)	The keys defined by the P21 check are not sufficient to identify a unique record for patient. LBREFID should be added to the keys list.

Table 3. Incorrect explanation for SD1117 and improper use of surrogate keys

Above is an actual explanation from a sponsor, and the actual keys listed in define.xml for this domain, Laboratory Test Results (LB) were: STUDYID, USUBJID, LBREFID, LBTESTCD, VISITNUM, LBDTC. The --REFID variable contains a sequence number assigned by the sponsor and is unique per record within a subject. It acts as a surrogate for the variables that create the natural keys in the data so using this variable as a key provides no useful information regarding why there is duplicity as per standard keys in the structure of a domain.

It is understood that there may be other variables that contribute to the keys of the sponsor’s dataset, however many times these are sponsor-defined variables, such as --SPID (Sponsor-Defined Identifier) and --REFID. Stating that a variable such as --SPID is needed to uniquely identify a record is typically not useful information. According to SDTM IG 3.2, “*Although --SPID and --REFID are Identifier variables, usually not considered to be grouping variables, they may have meaning across domains*”. So, --SPID and –REFID are mostly useful to have meaning across domains when they are related. But, should not be used to make duplicate records to appear as Unique records for identification.

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

In this category, the data has two or more records with most of the variables having same information. And 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.

Example 1: Records with the same timing information for a test, but the results are different. Sponsor explained in reviewers guide as “Standard Key variables STUDYID, USUBJID, LBTESTCD, VISITNUM, LBDTC are not enough to identify unique record. LBORRES should be added to the list”.

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

Table 4. Example of exact duplicate records with different results 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. Reasons may include “Incorrect reading first time”, “instrument calibration was not accurate in first round” etc. Then, map the comments (or any additional information) to an SDTM variable to be listed in the primary key variables list to identify unique records.

Example 2: Records with the same timing information for a test, but one of the records is NOT DONE. In this example, sponsor chose to add LBSTAT to the key variables to make records appear unique.

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

Table 5. Example of duplicates due to NOT DONE records and actual results

Resolution/Explanation: Similar to above example, the first choice should be to remove duplicate records by data management. The second choice is to find the reason for why the test was not able to be done and map the reason to one of the key variables to make the records appear unique and have a reason for the duplicity.

3. Insufficient natural Keys

In this category, sponsor collected duplicate data per key variables, but due to a business meaning.

Example 1: The standard natural keys described for PE domain (as a part of Findings domain) in P21 rules is below.

STUDYID, USUBJID, VISITNUM, PETESTCD.

But, a sponsor collects data in such a way that the location (PELOC) and method (PEMETHOD) variables need to be included in the natural key to identify a unique row, but they do not collect a visit variable; instead they use the visit date (PEDTC) to sequence the data. Thus, the new primary keys in define should become following

STUDYID, USUBJID, PEDTC PETESTCD, PELOC, PEMETHOD

Another sponsor might have used ultrasound as a method of measurement and might have collected additional information such as the makes and models of ultrasound equipment employed. The sponsor considers the make and model information to be essential data that contributes to the uniqueness of the test result, and so creates Supplemental Qualifier variables for make (QNAM=PEMAKE) and model (QNAM=PEMODEL). The primary key then becomes as below. The point to be noted here is that the domain is not limited by its physical structure.

STUDYID, USUBJID, PEDTC, PETESTCD, PELOC, PEMETHOD, QNAM.PEMAKE, QNAM.PEMODEL

Example 2: Example of results from different labs. In this example, the same lab test was conducted at two different laboratories. Although, the values are all same with in the standard key variables STUDYID, USUBJID, LBTESTCD, VISITNUM, LBDTC, 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.

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

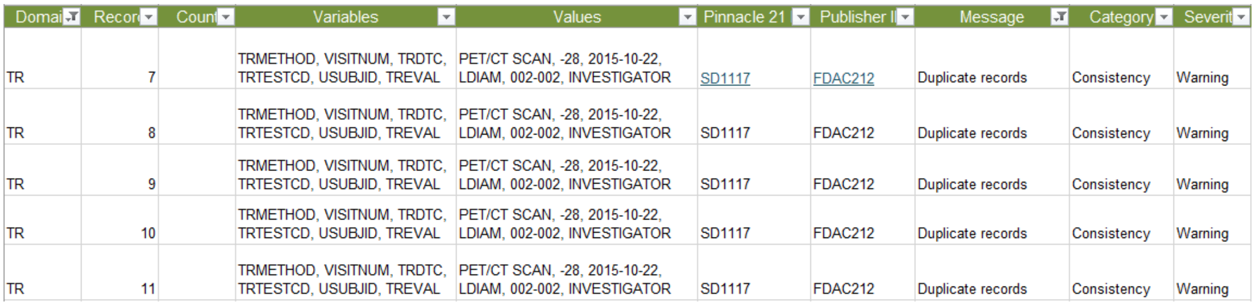
Table 6. Example of duplicate records per key variables with results from different labs

Example 3: Example of results of a split tumor with different linkID. As a typical Findings domain, P21 identifies unique records in TR domain by STUDYID, USUBJID, TRTESTCD, TRMETHOD, TREVAL, VISITNUM, TRDTC.

Similar to the above example, the data below shows TRLNKID as 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.

USUBJID	TRSEQ	TRGRPID	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

Table 7. Example of duplicate records per key variables with results from a split tumor



Domai	Recor	Count	Variables	Values	Pinnacle 21	Publisher I	Message	Category	Severit
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, -28, 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

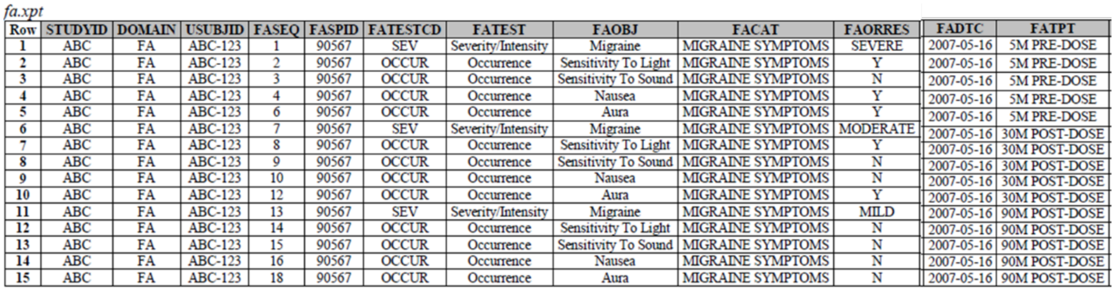
Figure 4. Pinnacle 21 report generated on data shown in Table 7 above.

Resolution: All above are completely valid cases and correct implementation. We use this example to emphasize a need for good documentation. Add additional natural keys required by the business meaning to define.xml and explain in reviewers guide. Because these additional natural keys are required to describe the structure of the data collected as per the protocol schedule of assessments.

4.Incorrect primary Keys/Incorrect mapping of data

Example1: The P21 identifies FA domain as a Findings domain, and uses standard key variables of a typical findings domain to identify unique records. The standard keys used are STUDYID USUBJID FATESTCD FACAT VISITNUM FADTC. However, the standard keys listed in SDTM IG 3.2 are below.

The keys in SDTM IG (v 3.2) are STUDYID, USUBJID, FATESTCD, FAOBJ, VISITNUM, FATPTREF, FATPTNUM.



fa.xpt

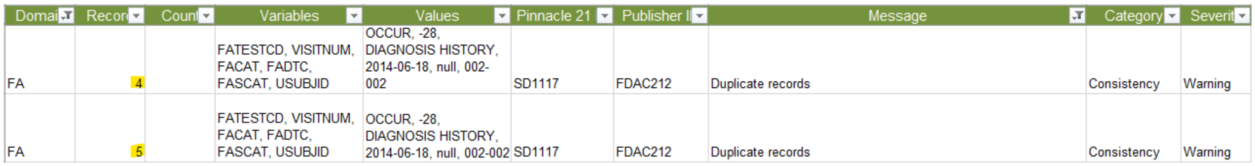
Row	STUDYID	DOMAIN	USUBJID	FASEQ	FASPID	FATESTCD	FATEST	FAOBJ	FACAT	FAORRES	FADTC	FATPT
1	ABC	FA	ABC-123	1	90567	SEV	Severity/Intensity	Migraine	MIGRAINE SYMPTOMS	SEVERE	2007-05-16	5M PRE-DOSE
2	ABC	FA	ABC-123	2	90567	OCCUR	Occurrence	Sensitivity To Light	MIGRAINE SYMPTOMS	Y	2007-05-16	5M PRE-DOSE
3	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

Figure 5. An example of a typical FA domain mapping in SDTM IG v3.2 showing FAOBJ as one of natural key

Table 8 below shows data using FAOBJ as a key variable to identify unique records. All three records in data below have same test collected at same timepoint, but the objective of each test is different. However, It is clear that the P21 does not consider FAOBJ to find the uniqueness of records in FA. As a result, the P21 report generated on FA domain shows warnings like below in Figure 4 and identifies records with FASEQ = 4 and 5 as duplicate records.

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

Table 8. FA domain using standard keys prescribed in SDTM IG v3.2 including FAOBJ in the list.



Domai	Recor	Count	Variables	Values	Pinnacle 21	Publisher I	Message	Category	Severit
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

Figure 6. Pinnacle 21 report generated on data shown in Table 8 above.

Resolution/Explanation: The obvious resolution in this example is to explain that the FAOBJ determines the uniqueness of records and should be added to the primary key variable list. As it is a legitimate natural key variable, it should be accepted by reviewers, and both all three records can be kept in the data.

Example 2: 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 example below, the domain (in table 9) name ZA starts with letter "Z" indicating it is a findings domain. The data has duplicate records based on standard key variables of a Findings domain USUBJID, --TESTCD, --CAT, --VISIT, --DTC (although date variable is not shown below, assume the date is same for all the records shown below). Records with ZASEQ = 1, 2, 3 and 4 has same ZATESTCD = "WHEALDIA", same ZACAT = "Histamine Control 10 mg/mL" and same time location and timepoint. As a result, the P21 will identify these as duplicates. The Sub-Location information for each record (Where ZASEQ = 1, 2, 3 and 4) which would make these records unique if mapped as suppqual variable (and reattached to parent domain), are instead chosen to be mapped as separate ZATESTCD records where ZASEQ = 5, 6, 7, 8. As a result the records where ZASEQ = 2, 3, 4 and 6, 7, 8 will be considered duplicates by P21.

za.xpt

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

Resolution: The solution is to choose and remap the above data into a domain to make data structure standard. For this type of data, the best fit domain to be mapped is SR domain and SUPPSR supplemental qualifier domain. As SR is under Findings About classification, values in ZACAT from above example are mapped to SROBJ to make a more meaningful sense of the Standard key variable of a domain in Findings About class. The records where ZATESTCD = "SUBLOC" are now mapped to SUPPRS domain. As explained in Category 3 above, a domain is not limited by its physical structure, the QNAM in SUPPRS can be added to the primary key variables list. The new key variables to define the structure would be as below. The new key variables describe the structure perfectly, and all records below are unique without losing any key information in the data.

STUDYID USUBJID SRTESTCD SROBJ SRLOC QNAM.SRSUBLOC

Sr.xpt

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

Suppsr.xpt

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. Missing values in keys

In this category, the data appears to be duplicate per standard key variables due to the missing values in one or more of the standard key variables.

Example: For DA domain, standard keys are STUDYID USUBJID DATESTCD and DADTC in DA domain. As the data below has missing values of DADTC, multiple records will have same DADTC (i.e., missing) as per Pinnacle tool and are identified as duplicates.

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

Table 12. Example of duplicate records with missing values in key variables

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.

6. Subjects Rescreened into the study

Some studies require patients to be re-screened due to amendments in the protocol. Other cases like, in extension studies, investigators allow patients from primary study to enroll in the extension study. When the main study and extension studies data are pooled for safety and efficacy analyses, the same subject records may be found duplicate in the subject level data. This is a whole separate issue and is out of scope of this paper. Previously, many articles discussed about this issue. See recommended readings section.

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, and 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

RECOMMENDED READING

[1] Prepare for Re-entry: Challenges and Solutions for Handling Re-screened Subjects in SDTM, PharmaSUG 2016 – DS09

CONTACT INFORMATION

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

Varun Debbeti
Cytel
1050 Winter Street STE 2700
Waltham, MA 02451
Work Phone: 6175287176
Email: Varun.debbeti@cytel.com
Web: www.cytel.com