PharmaSUG 2017 - Paper DS15

Common Programming Errors in CDISC Data

Sergiy Sirichenko, Pinnacle 21

ABSTRACT

Data in standardized format is now a required part of regulatory submissions. CDISC standards have now achieved widespread adoption and have become a commodity skillset for thousands of clinical programmers. Nevertheless, there are still mapping and programming errors commonly observed in standardized data. These reduce the overall quality of submissions and should be avoided. Especially since majority of programming errors can be fixed even after the study data is locked, which is not the case with data management and data collection design issues.

This presentation will share our experience of the most common mapping and programming errors observed across hundreds of regulatory submissions. We will provide examples and recommendations on how to detect these issues, how to evaluate their impact on regulatory review, and how each can be corrected.

INTRODUCTION

Today, *Data Conformance* is a standard and required part of regulatory submissions. Regulatory agencies expect sponsors to validate their study data before submission and either correct or explain discrepancies in the Reviewer's Guide.

There are special Rejection rules for study data from both FDA and PMDA. Their violations can potentially result in a rejection of regulatory submission until the issues are fixed and data is resubmitted.

Sponsor is responsible for quality of submission data. Sponsor can outsource work, but cannot outsource their liability. Some programming errors may be critical and delay regulatory review.

Pinnacle 21 is the most commonly used tool by the industry including both regulatory Agencies.

A free open-source P21 Community version is a personal desktop application for SAS programmers to QC their work while preparing study data for regulatory submissions. A scope of P21 Community is limited to Regulatory Compliance according to officially published FDA and PMDA requirements.

PMDA has an explicit set of validation rules with a reference to P21 rule IDs.

FDA approach to manage validation rule is more complicated. They refer to three types of Study Data Validation Rules [1]:

- "*1. Standards Development Organizations (e.g., CDISC) provide rules that assess conformance to its published standards (See www.CDISC.org).*
- *2. FDA eCTD Technical Rejection Criteria for Study Data that assess conformance to the standards listed in the FDA Data Standards Catalog (See above).*
- *3. FDA Business and Validator rules to assess that the data support regulatory review and analysis*."

It means FDA will rely on CDISC to develop and publish validation rules for standards compliance while focusing on review-specific data requirements as well as additional business rules specific to their internal processes.

From practical approach for end users there is no major difference compare to PMDA case. Executable implementation of all FDA validation rules is still available through P21 software.

It's important to understand who is the *final user* for your study data. In our corner of the industry, the final users are FDA and PMDA Reviewers. Regulatory agency users are foremost interested in automating the

review process, which means that they expect “minimal standard compliance” and review specific data elements.

TYPES OF DATA QUALITY ASSESSMENTS

In terms of practical implementation, there are 3 general types of data quality assessments. It’s important to understand the difference between them.

P21 ERRORS

The original *Severity* in P21 Validator was a property of computational algorithm, rather than an estimation of *Risk* or *Impact* of reported data issue.

P21 *Errors* are *Checks* based on executable algorithms which produce issue reports with 100% confidence. For example,

- *Start Date is after End Date*
- *New terms in Non-Extensible CT* - Any new term is invalid in this case

P21 WARNINGS

P21 *Warnings* are *Reports* of potential data issues for manual review, which will decide if this is a real issue or not. For example,

- *New terms in Extensible CT* – New terms may be added. However, they should not overlap with existing standard terms. Such assessment cannot be automated and need manual review by a subject matter expert
- *Missing Units on Results* – Some assessments do not have units. Implementation of this business rule can filter out some common cases. However, at this point it cannot handle everything correctly. So, manual review is expected.

P21 DATA FITNESS REPORTS

P21 Data Fitness Reports are available only in Enterprise version. They represent additional diagnostics and a source of useful information. Here are some examples:

- *Death info reconciliation* – It’s a list of all subject death information in study data. Sometimes a subject death is not collected in expected way with missing *DEATH* records in DS, AE, DS domains. However, subject may have information about their actual death stored in very unexpected locations like CO, SUPPAE, DV, SUPPDV or any other domains. Consistency in subject *DEATH* info across DM, DS, AE and other domains is also imported. Due to lack of standardization for this type of information, data quality assessments performed by data fitness reports with flexible study-specific structure are superior to regular checks.
- *Quality of MedDRA Coding* – This report helps Reviewers with evaluation of submitted MedDRA coding by matching collected text for Adverse Events with MedDRA PT, LLT coding.
- *Content of SUPPQUAL domains* - This report provides a quick and convenient familiarization with non-standard study-specific data
- *Missing BASELINE* – There is *Missing Baseline Records* check which produces a list of subjects with this issue. An additional report includes all records for those subjects. It helps understand why the issue exist? For example, some subjects may not have assessments before dosing, some may not have Baseline Flag populated for existing records.

DIFFERENT SOURCES OF DATA ISSUES

The biggest challenge in Data Conformance is to understand a specific data issue, identify its source, determine risk, and find solution for its resolution.

There are three major places where data issues may originate:

- Study data collection design
- Data collection
- Mapping/Programming

STUDY DATA COLLECTION DESIGN

Study design and data collection usually represent actual data quality. Mapping/ Programming adds a value of standardization, but it may have negative impact if done incorrectly.

Usually issues related to Study Data Collection Design are the most complicated to fix, because it is too late. They may be costly.

For example:

- Missing Fatal AEs for subject Deaths *("we were not aware that it's important*")
- Missing Seriousness Criteria ("*this info is collected in different system*")
- Missing Study Termination Dates ("*not enough time to clean the latest data before data lock, so we'd rather not collect this info*")
- AE Action Taken info is collected as any actions rather that actions related to study drug. Such approach often results in missing info about expected Action Taken with Drug.

While Programmers cannot help when information was not collected, they should be involved in study data collection set-up to ensure correct implementation of standards and that good practices are followed in collecting data to support intended analysis.

DATA COLLECTION

Same is true about Data Collection process. Data Management team needs help from Programmers to clean data.

Standardized data allows automation including data cleaning. Today, some vendors can provide study data in SDTM format just one (1) week after the first subject first visit. It provides an opportunity to execute data management queries and clean collected data by standardized programming code.

In many cases, it's more efficient to implement edit checks outside of data collection system due to their complexity or high quantity, which reduces performance of EDC systems or other data collectors.

Data cleaning programming activities are expected to be included in study Data Management Plan, be part of Blind Data Reviews, etc.

Here are examples of data collection issues which may be easily identified by SAS programmers and fixed before study data lock:

- Start Date is after End Date
- Missing Original Units/Normal Ranges for some Lab results
- Missing Severity and Causality for Adverse Events
- Inconsistency in Death info (dates, missing records)
- Missing scheduled assessments
- Invalid data

In some cases, special programming is needed to handle data collection issues (e.g., a missing info for Required variables).

MAPPING/PROGRAMMING

There is no good explanation for mapping/programming errors. They all need to be fixed.

Programming cost is marginal compared to total cost of the study conduct. Programming errors may compromise study results.

The following sections provides examples of commonly observed programming errors in study data.

SDTM DATA

--STRF, --ENRF AND MISSING VALUE IN RFSTDTC, RFENDTC

According to SDTM IG --STRF and --ENRF variables "represent the timing of an observation relative to the sponsor-defined reference period" based on RFSTDTC and RFENDTC variables. If for some reason a subject does not have RFSTDTC info, then usage of --STRF variable is not applicable. Information, that some event is "*BEFORE*", make sense only when referenced to a specific timepoint represented by RFSTDTC value. Usually RFSTDTC and RFENDTC are not defined for Screen Failure subjects. RFENDTC is often missing for some subjects under study treatment in ongoing studies.

In such cases a programmer should utilize another set of variables --STRTPT and --STTPT.

--STTPT/--ENTPT variables define a timepoint like RFSTDTC, but it is not limited to date/time in ISO 8601 format and may have a descriptive value like "*Visit 1"* for Screen Failures. Interim data cut date may be used instead of RFENDTC for ongoing subjects.

INCONSISTENT --STRESU

Conversion of collected results into the same standard unit within each test assessment is required to perform valid and meaningful analysis. Definition of test assessment is not limited to --TESTCD value and may include other variables like Specimen Type (--SPEC), Method (--METHOD), or even Category (--CAT) and Subcategory (--SCAT) variables which are often used in addition.

It's quite simple programming task when utilizing standardized data collection or a central lab. There are some cases which introduce a challenge for programming.

There is no consistency in results format while using local labs. Each lab needs special handling to convert their results into standard units. It's quite tedious work. However, Programmer cannot skip this important step.

Here is an example of invalid approach when Sponsor decided not to fix "*Inconsistent LBSTRESU*" and provided an explanation that "*During Unplanned Visits lab tests assessments were done by local labs utilized different units for same tests. Conversion of results from local labs to standard units was not done because this data was not used in analysis. Sponsor considers this issue as not important and decided to not fix it."* The obvious problem in this example is that FDA and PMDA Reviewers may have different interpretation about importance of lab data during Unplanned visits. Usually study subjects had unscheduled visits due to some safety problems like Adverse Events. Therefore, analysis of data from Unplanned visits may be critical for evaluation of study drug safety.

There is a special case of technically inconsistent units due to violations in standard Control Terminology. While data is received from local labs or entered manually it may be the same actual units, but different presentations.

Here is an example for “Calcium” tests results submitted in one study:

LBTEST	LBCAT	LBSTRESU
Calcium	LOCAL LABS - SERUM	mg/dL
Calcium	LOCAL LABS - SERUM	mg/dl
Calcium	LOCAL LABS - SERUM	mg/ml
Calcium	LOCAL LABS - SERUM	mmol/
Calcium	LOCAL LABS - SERUM	MMOL/L
Calcium	LOCAL LABS - SERUM	mmol/L
Calcium	LOCAL LABS - SERUM	mmol/l
Calcium	LOCAL LABS - SERUM	mmols/L
Calcium	LOCAL LABS - SERUM	mmols/l
Calcium	LOCAL LABS - SERUM	mmom/L

Table 1. LBSTRESU for Serum Calcium results in one study

There is an inconsistency in Units for Standardized Result. However, some reported units are synonyms like “*MMOL/L”, “mmol/L”, “mmol/l”, “mmols/L”, mmols/l*”, and potentially *“mmol/”, “mmom/L*”.

IMPUTED STUDY DAYS

According to FDA TCG: “*No data should be imputed in SDTM datasets. Data should only be imputed in ADaM datasets”.* A value for Study Day and Epoch variables may be derived, but not imputed in SDTM data.

Nevertheless, some studies use imputation for Study Days when a date of observation is collected as a partial date. Usually an imputation is performed based on middle of unknown time period. Such errors should be avoided.

There are some exotic cases, when imputation of Study Days is done based on Subject Visit Number to correct data entry errors in Visit Dates. For example, for “Visit 1” the date was collected with an incorrect previous year. According to the study protocol, “Visit 1” is a date of randomization and the first study drug dose. In this case, a Programmer populated the actual value *DY=1 instead of *DY=-366 according to incorrectly collected Date. Nevertheless, it’s an obvious programming error.

MISSING AE TREATMENT EMERGENT FLAG IN SUPPAE

FDA asks sponsors to populate AE Treatment Emergent Flag (AETRTEM) in SUPPAE domain according to an example provided in SDTM IG 3.1.2 #8.4.3. This is an example of Review and Tool-Specific requirements. Most reporting tools utilize SDTM data instead of ADaM due to predictable data structure in Tabulation data.

Nevertheless, most Sponsors continue ignoring this FDA requirement and populate AE Treatment Emergent Flag only in ADaM data.

Such approach reduces the value of submission data and makes data analysis more complicated. When ignoring this FDA requirement Sponsors lose their opportunity to indirectly communication with Reviewers about their implementation of AE Treatment Emergent Flag.

Reviewers may have different definitions and algorithms for study specific AE Treatment Emergent Flag.

EPOCH variable may be helpful to select AETRTEM records. For example, if EPOCH for AE is “*SCREENING*”, then AE is expected to be non-treatment emergent. In some cases, EPOCH may be confusing. For example, EPOCH = “*FOLLOW-UP*”. If AE started on day 5 after the last dose it may be

considered as Treatment Emergent despite of *FOLLOW-UP.* Some Reviewers may have different interpretation of AETRTEM.

- It may be based on one (1) half-life time of study drug after the last dose
- Some Reviewers may use four (4) half-life periods
- Drug may be stored in tissue for longer period
- Some Reviewers may consider everything after the first dose as Treatment Emergent

INCORRECT RFPENDTC

As any diagnostics test, data validation produces not only correct True-Positive and True-Negative results, but incorrect False-Positive and False-Negative results.

False-Negative means that an issue exists, but is not reported.

Almost nobody complains about False-Negative validation messages mostly due to missing validation messages in contrast to False-Negative issues. It's challenging to perform diagnostics to actual issues which are not reported. However, an issue is still an issue regardless if it is reported or not.

It is important to remember that a Sponsor is responsible for regulatory submission including quality of study data. Sponsors can outsource work, but Sponsors may not outsource their liability to data vendors.

There are two major sources for False-Negative issues:

- Incorrect algorithm for existing checks
- New checks for implementation

For example, the check "Date is after RFPENDTC" was introduced in OpenCDISC v1.4.1 and removed in v2.0, which was limited to FDA official business rules. However, those checks are still a part of P21 Enterprise utilized by PMDA and FDA. Most studies prepared by P21 Community Validator have a problem with this business rule. An issue is not identified, not reported, and not explained by Sponsor in Reviewer's Guides.

MISSING VALUE IN REQUIRED VARIABLES

"*A missing value in Required variable*" is a severe violation of SDTM/SEND compliance which may result in failing automated processes like uploading study data into a data warehouse or running standardized analysis tools.

Typical examples of this rare, but still existing error are missing value in AETERM (Reported Term for the Adverse Event) or EXTRT (Name of Treatment).

In most cases this is a data collection problem. However, SAS Programmers should handle this error by special SDTM/SEND mapping.

A value for Required variables must be populated. SAS programmers should use some special custom terms like "*Unknown*", "*Not collected", "Missing value*", etc. Proper documentation in Reviewer's Guide is expected.

However, such approach may complicate assessment of data quality and still confuse analysis. Therefore, standardized industry-wide terms like "*NULL*" or (Null Flavor) Control Terminology may be a better solution.

MISSING EXPECTED VARIABLES AND REVIEW-SPECIFIC INFO

SDTM Expected variables must be present in domains. When information is not collected, all records may have a missing value for such expected variables.

There are also review-specific data elements requested by FDA like EPOCH, Study Day variables, AE Treatment-Emergent Flag, or Baseline records. Sponsor should populate all this information if possible.

There were some cases, when Sponsor tried to trick FDA validation process. For example, current implementation of validation check for EPOCH variable is limited to the presence of this variable and does not look for values populated across all records. So, a few Sponsors created a blank EPOCH variable instead of providing actual information requested by FDA.

Another bad practice is a modifying of collected data in to clean up validation results.

For example, subject Race and Ethnicity may not fit standard terms and are collected as "*MULTIPLE*", "*OTHER*" or "*UNKNOWN*". The expected approach is to store these terms in RACE variable and provide details in SUPPDM if needed like in cases of "*MULTIPLE*" or "*OTHER*".

P21 generates Warnings about non-standard CDISC terms. Expected approach is to review these validation issues and document that implementation is correct. However, some Sponsors instead populate a missing value in RACE variable because it does not produce the validation Warnings. Such approach reduces quality of study data. A missing value in RACE variable means that RACE information was not collected.

New validation checks for data completeness will be introduced soon.

An opposite approach "to please Validation checks" is populating a dummy value instead of the correct missing value.

There is a check about "*Missing value for LBSTRESU*". This business rule cannot be fully automated. Therefore, the existing check produces a report for manual review to decide if specific test result is expected to have units or not?

To avoid validation Warnings, some Programmers populate dummy values like "*NO UNITS*". This is incorrect implementation of CDISC standards.

INVALID SDTM MAPPING

Correct mapping of collected data to SDTM structure is important to ensure data standardization. There may be complicated cases with exotic data elements when mapping is not obvious.

However, there are many examples when standard data elements are stored in wrong locations.

SDTM standard has a special purpose domain Comments (CO). However, some Sponsors still prefer to use SUPPQUAL domains to keep comments. While sometimes such approach may seem more logical for a Programmer, it could be confusing for a Reviewer because information was put into a wrong location.

Subject DEATH dates are expected to be stored in DSSTDTC (Event Start Date) variable. However, in some studies this important information is incorrectly stored in DSDTC variable which represents Collection Date.

Subject BMI (Body Mass Index) is part of Vital Signs data with its dedicated location in VS domain. In some studies BMI is also stored in SUPPDM or SUPPLB datasets. Such data redundancy should be avoided.

Most studies have duplicate records due to data collection issues. An expected approach is to identify such problems early and try to clean data before database lock. If it's too late, then Sponsor should investigate reasons for duplicate records and provide explanations in Reviewer's Guide. Remember, that Warnings about duplicate records are a report for manual review. In some cases, reported duplicate records are not actual duplicates, but False-Positive messages due to non-perfect validation algorithms.

Some Sponsors introduce so called *Artificial Surrogate Key Variables* to hide reported duplicates. Typical examples are --SPID variable as a copy of --SEQ variable or --GRPID variable without any documentation on its origin, derivation or meaning. Correct approach is to explain existing non-fixed issue, rather than try to hide it.

DATA INCONSISTENCY

Data inconsistency is a severe violation of data integrity. These errors may make study data unusable.

For example, RELREC or SUPPQUAL domains have references to non-existing records. It's a severe violation of structural data integrity, which may prevent execution of review and analysis tools

SDTM Model has special paired variables with expected one-to-one consistency in their values. Examples include --TESTCD/--TEST, --PARMCD/--PARM, ARMCD/ARM, QNAM/QLABEL, VISITNUM/VISIT, --TPTNUM/--TPT, etc. These variable pairs can be used interchangeably in analysis. Thus, submission data should have consistency between their values.

Consistency across studies is another level of complexity in standards' implementation which is important for data integration.

All studies have some extensions to standard CDISC Control Terminology. Managing non-standard Control Terminology within a company may be very beneficial especially for data integration.

It's important to remember that integrated database should follow the same Control Terminology. If there is no consistency between original data of individual studies, then an additional mapping into the same Control Terminology is required. The same is true about the use of a single MedDRA version and other external dictionaries in integrated data.

Some Sponsors performed only initial puling of data from different studies and claimed such data as integrated. However, such approach can be utilized only if there is a high level of consistency in implementations across individual studies. Otherwise, puled data cannot be considered as integrated without additional mapping and re-coding.

SEND SPECIFIC

Creation of SEND data is different compared to SDTM. While SDTM mapping and conversion are usually done manually by programmers, SEND data is commonly populated by data collection systems. It's possible due to less variations in structure of pre-clinical data. Such automated approach reduces cost of data conversion and ensures consistency in deliveries. On the other side, it's more difficult to tune programming code embedded into COTS software systems compared to code controlled by internal team of SAS Programmers.

There are common SEND specific issues Programmers should be aware and avoid:

- All SEND variables are included into domains regardless if these variables were collected or derived in the study.
- Deficiencies in study data documentation due to "generic" define.xml file and Reviewer's Guide. For example, a generic and sometimes invalid explanations for data issues or generic Key Variables for domains with references to variables not utilized in the study.

CONCLUSION

Now that standardized study data is a required part of regulatory submissions, reviewers can take advantage of standardized review and analysis tools. Therefore, it's critical that Programmers avoid common mapping and programming errors, which can reduce the overall quality of submissions. Programmers can follow the examples and recommendations in this paper to detect, understand, and fix common issues to avoid impacting regulatory review process.

REFERENCES

1. FDA. "Study Data Technical Conformance Guide". March 2017. Available at https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

CONTACT INFORMATION

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

Name: Sergiy Sirichenko
Company: Pinnacle 21 LLC
Work Phone: 908-781-2342
E-mail: ssirichenko@pinnacle21.net