PharmaSUG 2020 - Paper SS- 097

Data Review: What's Not Included in Pinnacle 21?

Jinit Mistry, Lyma Faroz, Hao Meng, Seattle Genetics, Inc., Bothell, WA

ABSTRACT

Many pharmaceutical and biotechnology companies outsource statistical programming activities and submission package preparation to CROs. Still, for all programming deliverables the sponsor remains responsible for quality, completeness, and compliance to published standards and regulatory guidance. This makes it critical for the sponsor to implement efficient vendor oversight that touches on enough detail to ensure quality of the product provided by the CRO.

Sponsors are widely using the Pinnacle 21 toolset to ensure SDTM and ADaM compliance with CDISC guidance. However, by itself this is not enough and additional review of how the CRO adopted CDISC implementation guides to assign or derive variables in alignment with study design, protocol, and SAP need to be conducted beyond Pinnacle 21 reports. For example, Pinnacle 21 checks whether variable values are present and runs several logic and interdependency checks, but it doesn't validate the correctness and accuracy of such values in relation to study documents and other specifications or constraints.

This paper will share various data validation checks that can be performed outside of Pinnacle 21 to significantly heighten the quality of any submission and help mitigate review questions and technical rejections.

INTRODUCTION

The clinical industry conducts a tremendous number of research studies to prevent diseases and/or cure patients in various therapeutic areas. Given the fast-paced nature of the pharma/biotech industry, it is very useful for sponsors to develop standards to gain efficiency for submissions to agency (e.g., FDA, PMDA, EMA, etc.) review. Ensuring compliance with CDISC standards is an industry adopted common practice. Sponsors benefit from standards development in terms of efficiency while regulatory agencies benefit through efficient data review by getting familiar more quickly with sponsor data via standards and supporting tools, so this is a win-win situation for both parties.

Many times, CROs also play a vital role to support sponsor companies in reaching their desired goal of study submissions. Often multiple companies partner up in the preparation of study data standards and many times data standardization activities are distributed to various programming teams according to domain expertise. For all the stakeholders such as CROs, sponsors, and regulatory agencies, study data quality is the most critical. Hence, programmers working on study data must ensure study data quality and compliance with CDISC rules are tested by Pinnacle 21 software. Pinnacle 21 is very helpful for CDISC data checks, however in our recent experience we noted a few additional checks could be run outside of Pinnacle 21 to ensure study data quality. Critical examples are shared in this paper to help others improve processes, uncover blind spots, and ensure study data quality.

BACKGROUND

FDA requires all electronic clinical data submissions to follow the CDISC standard for a faster and less resource-intensive review process. In order to automate the data review process, FDA has also built tools based on standards to aid in the analysis of data such as their clinical trial repository (JANUS). To help sponsors submit data of highest quality, FDA started publishing various validation rules and requirements which led to the release of documents such as the Study Data Technical Conformance Guide (TCG), the Data Standards Catalog, and validation rules for SDTM and SEND. The TCG has three sets of validation rules: Technical Rejection Criteria, Standards Conformance Rules, and FDA Business Rules. In Japan, the PMDA recently published similar expectations as well. The Pinnacle 21 tool has implemented a major chunk of these rules and continues to work towards implementing the rest of them as well for both FDA and PMDA compliance. However, there is a limit to what can be done programmatically. Hence, in this paper we will suggest more checks that can be executed alongside P21 checks towards electronic submissions of high-quality data to regulatory authorities such as FDA and PMDA.

TRIAL DESIGN DOMAINS

The Trial Design domains are important for quickly exposing trial-related information. All trial design domains are important since they contain key aspects of trial design, but the trial summary (TS) domain is arguably the most critical SDTM domain from a technical rejection perspective. This severity is further emphasized by looking at the Pinnacle 21 validation checks for FDA rejection criteria: 3 out of 4 rejection checks relate to the TS dataset. Therefore, there should be careful review of these trial design domains. Please review below Table 1 for details.

Rule ID	Publisher ID	Message	Description	Domains	FDA	PMDA 1511.6	PMDA 1810.3	FDA Severity
SD1115	eCTD 1734	Missing TS dataset	Trial Summary (TS) dataset must be included in every submission.	TS	X	X	X	Reject
SD2232	eCTD 1734, FDAB011, CG0287	Missing SSTDTC Trial Summary Parameter	'Study Start Date' (SSTDTC) record must be populated in Trial Summary (TS) domain. It is expected for SDTM IG v3.1.2 data and required for data in all more recent SDTM versions.	TS	X	X	X	Reject
SD2247	eCTD 1734, CG0285	Invalid TSVAL value for SSTDTC	TSVAL variable value must be in ISO 8601 format, when TSPARMCD='SSTDTC'.	TS	X	X	X	Reject
SD1020	eCTD 1736	Missing DM dataset	Demographics (DM) dataset must be included in every submission.	DM	X	X	X	Reject

Table 1. Pinnacle 21 Validation Rules for FDA Severity='Reject'

Another thing to note is that trial design domains are prepared from study documents, study data, and other published resources related to a study. This is a manual activity usually performed by study programmers in collaboration with cross-functional teams. P21 validation rules check for presence of key TSPARM(Trial Summary Parameter)/TSPARMCD(Trial Summary Parameter Short Name) values, validity of TSVAL (Parameter Value)/TSVALCD(Parameter Value Code), dependency checks, missing variable values in the TS dataset, etc.; but it does not verify study/protocol specific data. Below are a few cases where such study-specific data in TS can be verified programmatically.

TS CASES

Case 1.1: The TSVAL (Parameter Value) value is assigned incorrectly where TSPARMCD (Trial Summary Parameter Short Name) ='ACTSUB' (actual number of subjects).

P21 Rule ID 'SD2234' triggers if TSPARMCD='ACTSUB' (actual number of subjects) is missing in the TS dataset and Rule ID 'SD2249' triggers if TSVAL is not numeric in the same setting, however these cannot ensure the entered value is correct. This might sometimes happen when programmers are involved in the preparation of TS in multiple studies at the same time. In this scenario, the manually assigned value for the number of subjects can sometimes be incorrect in TS, leading to submission of an incorrect actual number of subjects to a regulatory. To prevent this, the actual number of subjects in this TS observation should be cross-checked against the SDTM DM (demographics) dataset.

Case 1.2: The TSVCDREF (Name of the Reference Terminology) value is truncated.

An example is where for a TS observation where TSPARMCD='AGEMAX' (TSPARM='Planned Maximum Age of Subjects'), variable TSVCDREF='ISO 2109' but according to the SDTM-IG it should be 'ISO 21090'.

P21 checks do not check for truncated values in TSVCDREF, hence programmers need to pay close attention.

Case 1.3: TSVALCD (Parameter Value Code) is not populated when TSVAL (Parameter Value) is present.

In one case, we noticed that on observations where TSPARMCD='ADAPT' (Adaptive Design), TSVAL='N', TSVCDREF='CDISC', and TSVCDVER='2015-12-18' had a missing TSVALCD but didn't trigger any

message/rule in P21. Per controlled terminology, TSVALCD should populate 'C49487'. The same scenario applies for TSPARMCD='ADDON' (Added on to Existing Treatments), TSPARMCD='RANDOM' (Trial is Randomized), TSPARMCD='FCNTRY' (Planned Country of Investigational Sites), etc.

Case 1.4: TSVCDVER (Version of the Reference Terminology) is incorrectly assigned for corresponding TSPARMCD (Trial Summary Parameter Short Name).

Here, TSPARMCD='OBJSEC' (add decode here please (-:) has assigned TSVCDVER='2016-03-25' but it should be based on protocol text and not controlled terminology. Hence, it should be missing. This check is not triggered in P21. There should be parameter-specific checks as some parameters are protocol-specific while others are standard-specific.

Case 1.5: Incorrectly assigned dose unit in TSVAL when TSPARMCD='DOSE'

Per the SDTM IG, dose unit should not be part of TSVAL when TSPARMCD='DOSE'. To specify dose units corresponding to TSPARMCD='DOSE' and TSGRPID records, TSPARMCD='DOSU' should be added in conjunction with TSGRPID to describe dose unit. For example, TSVAL='0.9 mg/kg' when TSPARMCD='DOSE' and TSGRPID=1 should be described as TSVAL='0.9', TSPARMCD='DOSE' and TSGRPID=1. Corresponding record can be added as TSVAL='mg/kg', TSPARMCD='DOSU' and TSGRPID=1. This check is not triggered in P21.

Case 1.6: TSVAL values when TSPARMCD='PCLAS' (pharmacological class) are not consistently assigned for multiple studies towards submission for same drug.

TSVAL values where TSPARMCD='PCLAS' (pharmacological class) should be consistently assigned for multiple studies using the same drug. sometimes it is seen different TSVAL assigned for different studies that had essentially the same drug towards submission for multiple studies. Programming was done by different companies for different studies in the same filing, and there can be different interpretations of pharmacological class based on different interpretations between company teams. Communication and discussion with cross-functional and partner teams are needed to assign consistent values for TSVAL where TSPARMCD='PCLAS'.

Case 1.7: TSPARMCD='TITLE' has TSVAL truncated.

This case is related to assigned data values and should be checked against the corresponding study protocol to ensure correctness and completeness of the assigned value.

Similarly, data truncation and assignment of data values must be cross-checked in other trial design domains such as TI, TV, TE, and TA in reference to the protocol, since no P21 check will trigger when it comes to study data values. P21 is useful for compliance with standards but not necessarily for checking of specific data values. The suggestion here is to be diligent while reviewing the trial design domains in a submission data package to avoid such scenarios.

OTHER SDTM CASES

In this section, more examples are shared to uncover the blind spots in other SDTM classes such as findings, events, interventions, and special-purpose domains. Note that most of these data are usually collected in CRF.

Case 2.1: Data should be cross-checked while dealing with external data and ambiguous/incorrect mapping/derivation instructions in a specification.

In oncology studies, tumor-specific datasets are PR, TU, TR, and RS. In critical studies, tumor evaluation is often done by two individuals: the investigator and an independent assessor. Independent assessor evaluations are typically collected by vendors and shared with sponsors in the form of source or SDTM-like structures. For such independent assessor records, often there would be two or more evaluators. An acceptance flag (-ACPTFL) is used to identify the most appropriate tumor result

If the acceptance flag is not a collected field, then there is programming logic involved in the assignment of TUACPTFL, TRACPTFL, and RSACPTFL. Since TUACPTFL, TRACPTFL, and RSACPTFL are permissible variables in the SDTM standard, there are no checks in P21 for such variables. In this situation, special attention is required for picking the appropriate records.

For example, consider a specification that mentions to combine data adjudication form records with demographics data. Consider programmatically, screening ID variable in the investigator data is used to merge it through a SUBJECT variable in the independent assessor data. Screening ID and SUBJECT represent different subjects, but in some cases, they may regrettably have the same value. This can be a special case depending on a company's data collection system. Here, the key thing to note is that this part can't be discovered via P21 checks, so the sponsor company should carefully review accepted tumor results, data, and specifications to identify any ambiguity. Another action that can be taken is to provide the programming derivation logic in the specification and have it checked by a SME to ensure the logic is correct. There should not be multiple interpretations while combining investigator data with independent assessor data. Note that in this example, specification instructions were ambiguous and led to a logical error in programming. A complete understanding of source data is needed while programming, which could be a time consuming and complex check.

We can share one more example where define.xml specifications are not accurately documented. At times there are gaps in programming specifications. For example, consider programming instruction mention 'Year of Birth' instead of 'Date/Time of Birth' in DM.BRTHDTC but as per dm dataset, it is complete date. In this example, spec instructions may be misleading and incorrect for reviewer even though programming is correct. We suggest study programmers to review such define.xml for correct interpretation for accuracy of derivations and programming.

Case 2.2: SDTM mapped CRF and define.xml should be reviewed for completeness of data with respect to study data collection.

Another important aspect while reviewing SDTM data is completeness of data mapping. Sources of data can be external files and CRF data. Many CRF-collected variables are database-specific and will not be mapped into SDTM, yet are important to cross-check other collected variables mapped to SDTM. One example is, usually identifier key variables are part of database design but not part of CRF. It is important to map such key variables in SDTM dataset variables so that relation of mapping records across multiple domains can be maintained. Example, TU, TR, RS and RELREC. Overall, all study-specific collected variables and identifier variables should be included in SDTM data unless an explanation is provided as to why certain collected data were excluded. Many SDTM variables are permissible variables in the SDTM IG such as LBTOXGR, AESEV, AESDTH, TUGRPID, etc. so P21 doesn't check for the presence of such variables. second example is, where collected values that should have been mapped into SDTM permissible variables were actually not included in dataset variable or improperly mapped into SUPP domains. These scenarios should be covered by reviewing the SDTM-annotated CRF, SDTM specifications, and SDTM datasets at an early stage. While preparing for submission down the line, define.xml can replace SDTM specifications with complete programming instructions.

Case 2.3: Dataset and specification are complete but Annotation missing in SDTM CRF variable for aCRF.

Another issue we observed was where a specification for CMENRF (end relative to reference period) in the concomitant medications (CM) domain referenced a dependency on a variable called CMONGO (CRF label='CM ongoing?') variable . Here, CMONGO reflected CRF-collected data but it wasn't mapped on the aCRF. This can be a gap in interpretation and can be avoided by reviewing the SDTM-annotated CRF, SDTM specifications/define.xml, and SDTM datasets. Here any variable reference from a raw (CRF) dataset should be annotated on the CRF in compliance with the SDTM IG.

Case 2.4: Special ASCII characters are not identified in P21.

In global clinical studies, it is common to have different data collection systems. Here it is noticed that based on different systems some special characters are collected in source data. These special ASCII characters are usually not identified by P21 checks. For example, the reported local lab unit given in display 1 has a special character. This special character is displayed differently in various systems and possibly impacts ADaM dataset programming and P21 checks. If user-defined codelists added such values at SDTM level then it won't trigger the message, but if such codelists do not contain the special symbol then P21 will trigger a codelist terminology warning. The suggestion here is to replace the special characters with generic characters so it won't create further issues.

Display 1. Lab Unit Collected a Special Character

Case 2.5: Reviewers Guide (RG)'s content should be reviewed thoroughly relative to transfer of SDTM datasets and issue summaries should be cross-checked against the P21 report.

It is observed few times that SDTM datasets transferred from CRO to sponsor are not marked (X) accurately in reviewer guide (RG). Example, PC dataset is part of submission data package but not marked (X) in RG. Here sponsor should ensure to put study specific datasets that are present in the study, in the reviewer's guide. Another important part is to document the unresolved P21 issues with proper details in the issue summary section of the reviewer's guide. Sometimes this issue summary is not updated in a cSDRG, so it is recommended to verify the RG relative to contents of the data package. This would cover this case.

Case 2.6: Manual review of define.xml

As part of submission preparation of define.xml, it's good practice to check the following:

- CRF page number links are directing as expected
- Hyperlinks are working as expected
- Origin is correctly mentioned
- Programming instructions are clear and complete with respect to data
- Links to critical study documents are provided and directing as expected
- Cross-check of high-level study information in define.xml based on protocol, for example: protocol name, study name, controlled terminology, CTCAE version, MedDRA version, WHODrug dictionary version, etc.
- Appropriate sort order variables are provided for each dataset
- Appropriate dataset attributes are provided

Case 2.7: eCTD guidelines are followed.

Sponsors should ensure the data set file names use all lowercase characters. Underscores ('_') and blank spaces should be avoided. If needed, dataset transport file names can have hyphens ('-').

ADAM DOMAINS

There are some requirements from the FDA for ADaM datasets, notably that an ADSL dataset must be submitted, ADaM datasets should be derived from SDTM and not from raw datasets, ADaMs should be analysis ready (one proc away), and derivations/variables in ADaM should be easily traceable to SDTMs. It is important to be aware of what exactly P21 covers and what it does not so that additional steps can be taken by the sponsor to adhere to good quality data. Apart from this, we're sharing a few examples related to ADaM data given below.

Case 3.1: ADaM datasets containing duplicate information should be avoided.

At times, while reviewing ADaM datasets, two different ADaM dataset names contain same information. For example, in some studies, ADDM (label ='Analysis level demographics') is developed in addition to ADSL (subject-level analysis data). ADSL serves the purpose of providing subject-level analysis data, which makes ADDM redundant. P21 couldn't do a dataset-level check for such a scenario. The genesis of this is understandable: ADaM is based on analysis needs and quite flexible. This kind of check can be done by sponsor to avoid such redundancies.

Case 3.2: ADaM definitions for derived data values (specifically efficacy and safety endpoint) should be checked thoroughly.

ADaM datasets have many derivations, some of them involve high complexity. ADaM specifications and derivations should be checked thoroughly against the SAP. P21 doesn't check for correctness of such

derived values; rather it checks for presence, standards compliance, and attributes of variables. we would like to share 2 issues identified in 'duration of response' derivation in ADTTE (label= 'Analysis level Time to event') dataset, that is one of efficacy endpoint derivation for oncology studies. One example is, ADaM spec and programming is aligned for DOR calculation which is (Date of event/assessment – Date of first dose of study medication) but according to SAP instructions it is missing "+1" to DOR result. Second example is, Duration of response should be derived for confirmed responders only but ADTTE has derived results for all the subjects that include confirmed responders and remaining safety population. There are logical issues observed while compared with SAP definition. One more case we would like to share is based on newly entered data. In this case subject has missed scan scenario, that is not present previously in data while programming initially ADTTE and occur in new data. SAP instructs to consider earlier date before missed scan rather than latest date for ADT (analysis date) assignment. This kind of critical issues can be caught by independent programming so it should be reviewed carefully. The prevention step is to carefully review data at regular intervals. The reason being many times new scenarios and data points come-up as study progress and programming should be updated accordingly.

Case 3.3: "ADaM-like" datasets that don't go through P21 checks should be checked. ADaM-like datasets that are used for analysis but don't necessarily follow any ADaM structure per the IG should be reviewed thoroughly for quality issues, as P21 checks won't trigger any validation rules.

Case 3.4: Derivation complexity should be documented and covered in define.xml.

Since ADaM data structures are quite flexible, it becomes a tedious task to document programming logic. Sometimes, instructions for combining datasets or filtering conditions are present in a program but not covered/explained in define.xml. The suggestion here is that in such cases, supporting documents should be prepared and provided in define.xml for complete traceability and interpretation.

Case 3.5: ADRG is not consistent with cSDRG for CDISC CT and MedDRA version.

We've seen cases where the cSDRG and ADRG mention different MedDRA versions for the same study; these should always be consistent. This could be due to different teams working on SDTM and ADaM independently. To avoid this, effective and efficient documentation and communication is required to ensure the recording of correct versions across multiple teams.

Derivation logic checks should cover all programming scenarios and align with protocol/SAP definitions. Variables should be analysis-ready with clear traceability to SDTM. Providing explanations in a reviewer's guide about ADaM-like datasets may help towards a smoother review process.

Also note that cases 2.5, 2.6, and 2.7 noted in the SDTM section of this paper are applicable for ADaM compliance as well.

P21 NOT RELEASED CHECKLIST

At the time of this writing, apart from the cases discussed in this paper, 16 checks are in development but not yet released by P21. Also, in correspondence to new guidelines from CDISC, further new rules will be released by P21. Since these checks would be useful to ensure data quality, we recommend to adjust the corresponding SDTM and ADaM programs or create logic checks to ensure those checks are still run while performing review as a sponsor company. These checks can be accessed through the hyperlink given in the References section.

CONCLUSION

As more standards get implemented by CDISC and FDA, there is a need for more validation checks and rules. Based on study submission timelines, we should be mindful of current validation rules available in P21 and ensure additional checks are implemented on our own end to cover what would otherwise be gaps in the review of data quality. The cases mentioned here are based on our experience and shared for awareness to help Industry in preparation of high-quality data submission packages.

REFERENCES

SDTM P21 published rules: https://www.pinnacle21.com/validation-rules/sdtm

ADaM P21 published rules: https://www.pinnacle21.com/validation-rules/adam

FDA study data standards resources: https://www.fda.gov/industry/fda-resources-data-standards/study-data-standards-resources

P21 not released checklist: https://www.pinnacle21.com/sites/default/files/blog/2019/06/adam-rules-cdisc-p21.xlsx

ACKNOWLEDGMENTS

We would like to acknowledge Shefalica Chand and Balavenkata Pitchuka for their input, constant support and inspiring us to contribute our learnings to PharmaSUG. We would like to thank Sreeram Kundoor, John Shaik and Michiel Hagendoorn for providing valuable comments on this paper.

CONTACT INFORMATION

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

Jinit Mistry

Seattle Genetics, Inc.

jmistry@seagen.com

Lyma Faroz

Seattle Genetics, Inc.

lfaroz@seagen.com

Hao Meng

Seattle Genetics, Inc.

hmeng@seagen.com