Paper 180-2023

Demystifying the define.xml: Overcoming the challenges of CRT Package.

Yoganand Budumuru

ABSTRACT

New regulatory guidelines for electronic submissions have significantly changed the statistical programming life and sometimes it is very tedious and inefficient particularly when we are creating Case Report Tabulation (CRT) package. Regulatory submissions are the most crucial part in clinical trials data analysis and accurate CRT package is needed during submissions for accelerating research investments to help bringing the advantage of new drugs or treatments to patients. Submitting poor quality CRT package would risk the outcome of the trial and hamper the years of research and development including cost escalation as Sponsors are investing huge efforts and billions of dollars in new technologies and solutions to optimize the processes to improve the quality of electronic submissions. Creating Case Report Tabulation (CRT) can come with several challenges. Ensuring consistency and harmonization across different data sources is crucial for accurate CRTs. Creating a comprehensive and well-structured CRT requires clear and concise reporting. It can be challenging to present complex medical information in a standardized format that effectively communicates the key findings, treatment outcomes, and adverse events. Ensuring consistency in the presentation style and adherence to reporting guidelines is essential. Generating CRTs involves significant time and resources. Researchers and healthcare professionals may face constraints due to limited availability, competing priorities, or restricted budgets. These limitations can affect the depth and quality of the case report, leading to potential gaps or limitations in the CRT. Overcoming these challenges requires meticulous planning and addressing common errors and warnings can help address these challenges and improve the quality of CRTs. This paper will describe the preparation of various steps including cleaned specs and addressing some common errors in developing specification and provide recommendations for best practices in planning and preparing CRT packages for regulatory submissions that helps speedy review by regulatory agencies for better outcome sooner.

INTRODUCTION AND BACKGROUND

According to CDISC “Define-XML transmits metadata that describes any tabular dataset structure. When used with the CDISC Foundational Standards, it provides the metadata for human and animal model datasets using the SDTM and/or SEND standards and analysis datasets using

ADaM". The Food and Drug Administration (FDA) of the United States and the Pharmaceuticals and Medical Devices Agency (PMDA) of Japan require Define-XML for each study in each electronic submission to specify the datasets, variables, controlled terms, and other metadata that were used.

The Clinical trial data information, including Dataset-Level, Variable-Level, Value-Level information, Code-Lists, Derivations, Comments, Supplemental Documents, and Analysis Results Metadata (ARM), are represented in the Define-XML document (Fu, Qiuping, 2023).

The annotated case report form (CRF), SAS datasets, metadata, and source programs that make up a piece of the NDA package submitted to the FDA are all included in the Case Report Tabulation (CRT). It is employed by the FDA during application review. The Define document, which provides metadata outlining the datasets, variables, and values, is reviewed first. To make it simple to navigate, everything is connected via both internal and external hyperlinks, bookmarks, and destinations (Connolly, Christine et al.).

Creating a high-quality submission package for clinical trials and regulatory submissions are challenging and indeed can be a complex process. There are several challenges involved in this process and some of the challenges will be discussed in this paper though it mainly focusses on spec cleaning and preparing the final define ready specifications for bringing the high-quality submission package for the approval by the regulatory agencies such as FDA or PMDA.

Derivations of derived variables:

Derivations of variables play key role in preparing both SDTM and ADaM specifications. To get the accurate specifications ready for creating CRT package we must keep in mind the following instructions or suggestions that are essential for creating high quality submission package.

First and foremost is about Any crossed-out text and we must make sure that any crossed-out text should be removed. Similarly, if there are any notes in the derivation of any variable those are not necessary for creating define.xml and we must remove those notes.

Another important issue most used is keeping SAS code in the derivation column of the specifications. any such SAS code be avoided and use simple English for derivation/description of all variables. Use of simple English statement to provide clear descriptions (avoid

programming assumptions) instead of SAS code for the derivations of derived variables in the define.xml.

- Avoid '=' in the beginning of the sentence in Description/Derivation column. Avoid any other special character at the beginning of the sentence as well (-, ‘, “, #, etc.)
- Any '=', 'Eq', 'ne', or '^=' should be expressed in words. The exact phrase should read clearly in English, depending on the context/location. For example, 'equal to', 'equals', or 'Set to' may replace an '=' or 'Eq'. And, 'does not equal', 'is not equal to', or 'are not equal', may replace 'ne' or '^='. Please check readability after replacing any symbol.
- Use space where required for example ‘missing(ASTDT)’ needs to be ‘missing (ASTDT)’.

It is very important to use proper punctuations in the derivation of all the derived variables. Some examples for use of proper punctuations in the derivation of any derived variable are listed below.

- There should be punctuation after any complete thought. In a list of steps, consider adding semi-colon to the end and a blank space after semi-colon. Otherwise, add a period (full stop) and capitalize the first letter of the next word, if needed. End the derivation with a period.
- Avoid unnecessary extra spaces before or after commas/quotes.
- Avoid use of “ to open quotes and ' to close or vice versa, they should be consistent.
- Use Quotes when needed for example ‘(g, ug)’ should be ‘('g', 'ug')’.

Another important suggestion is about sentence case and the Derivation is always in sentence case. For example,

- Use uppercase letter at the start of the sentence for example 'If' should be used instead of 'IF' or 'iF’.
- Clear capitalization important, for example 'Else If' needs to be 'Else if'.

Any variable name in the derivation should be in UPPERCASE. Dataset name should be included, also in UPPERCASE. If the variable comes from a pooled ISS dataset (rather than

study level dataset), include ‘ISS ’ before the dataset name.

When referring to a study, use the naming and capitalization from the SAP for consistency in documentation.

Here is an ISS study example where pooling 10 different studies for an integrated study and those names of those 10 different individual studies are listed in the below table and naming and presentation of these studies while developing the specs particularly the derivations of different variables of a pooled study.

Individual Study
PALBO2
MYOTIC
ATLAS
Study 110
ARCNEW
PALBO
HAWAI
CONDRO
EGUIN
Japan Study 1

References to ‘Myotic’ study should be updated to ‘MYOTIC’. References to ‘JPN’, ‘JAPAN’, or ‘Japan 1’ as a study should be updated to ‘Japan Study 1’. (‘JPN’ might appear as a COUNTRY value and it that case should remain. Use caution with find/replace for this particular case.).

In the excel spreadsheet cell, normally ALT+ Enter shortcut will be used for returns or go next line. It’s suggested that before finalizing specifications for creating define.xml All ALT+ Enter (returns) needs to be removed. After removing, check if adding a space is required (it

usually will be, but do not add a second space if there is already a space). Remove any bullet symbols or indentations as the formatting will not make sense without the returns.

Another important suggestion is about the maximum characters used for derivation of any variable. It is suggested that derivation of any variable be restricted to maximum of 1000 characters or if any derivation more than 1000 characters long will have to be dropped from the specifications and instead it should go into SDRG or ADRG. If the derivation is very close and a simple edit can reduce the length to 1000 characters without changing the meaning, make that edit. Otherwise, flag it in the Guide column of the spec and try to link to specific SDRG or ADRG.

Another suggestion is about mentioning any references. There can be no references to files or locations that are not part of the submission package (e.g., directory locations, extra tabs/sections in specifications, raw data sets/variable names or the SAP). Where there are references to extra tabs or to external spreadsheet (i.e., xls or xlsx file name), update the derivation reference to refer instead to ADRG. Information from the tab or spreadsheet will be included in the ADRG. (Flag in the Guide column that a link to ADRG will be needed.) Note: References to external xpt files which will appear in the define can remain as is.

Som other suggestions on spec cleaning worth mentioning below.

- Check if any variable is not present in ADaM dataset but used in the definition for a derived variable.
- Make sure using actual variable name instead of using VAR in the description for the variable.
- Need to be consistent with use of quotes across the different variables and domains.
- Use 'SoC' instead of 'SOC' for Standard of Care. Reserve ‘SOC’ for System Organ Class.
- Perform spell check (may use F7 to run spell check).

More spec checks:

- Check keys for submission-ready datasets are unique and that **XXSEQ** are not used in the final sort. (If XXSEQ is the only possible unique sort due to duplicates in study level data, then leave it at the end.)

- Check the control terms or formats for the variable are displayed correctly (for example 'SBTSTDTC', 'AESTDTC', 'AEENDTC' should be displayed correctly as 'ISO 8601 Format' instead of ISO8601).
- Do not include formats such as Date9.
- Dataset Name and Tab name needs to be consistent.
- The same naming convention for columns across specs and tabs should be used for a study.
- Ensure Origin is always populated. All origins like for example **'Assigned'** which are not 'Derived' should not have a comment present.

ADaM Define Rules:

1. Check contents of codelist within the Define.XML against the variable contents within the submission-ready dataset and make sure no 'Mismatches between codelist values in the define.xlsx and what exists in the actual data'.
2. Use sentence case to display the contents of 'Structure' for example for ADSL dataset, please display as '**One record per subject'.**
3. Ensure attributes (excluding length) of XPT files being submitted are the same as source datasets. *We are converting the ADaM SAS datasets (source datasets) into SAS transport file (*.XPT) which in turn is used for creating the define.xml. Sometimes it is observed that there is mismatch in the variable attributes between them which should be avoided by keeping only the relevant ADaM datasets in the designated folder.*
4. Check max lengths and casing:
    - Ensure that all variables have a label and make sure that there are no variable names > 8 or variable labels are greater than 40 characters.
    - File name of any external docs needs to be in lowercase (8 chars).

	Max length (character)	Letter case
ADaM variable	8	UPPERCASE
ADaM label	40	Propcase

ADaM dataset	N/A	UPPERCASE
External dataset	8	lowercase

5. The P21 reports should be created before define.xml (and need to be stored in …\Define\ADaM\OutPut folder).
6. Include all the external files not in eCRF used to create the analysis datasets under ‘Supplemental documents’.
7. Any external file used for the derivation; all those should have hyperlink in define.xml.
8. All the values under the 'Description' column should be hyperlinked and when clicked it should open the related xpt file of that domain for example adsl.xpt.
9. The label of a variable needs to be consistent in spec, define and ADRG.
10. The variable order in the spec should be also same as displayed in define.xml.
11. Any variable mentioned in ADRG as core variable should be present in all datasets.
12. Both ‘Tagged PDF’= and ‘Fast web view’= should have ‘Yes’.
13. Check if we need the footnote: ‘Copyright © 2022 IQVIA. All rights reserved. The contents of this document are confidential and proprietary to IQVIA Holdings Inc. and its subsidiaries. Unauthorized use, disclosure or reproduction is strictly prohibited.’
14. ADaM CRT package would normally include the following:
    - ADaM define.xml and define.html (with ARM (analysis results metadata) or without ARM since it is required for ADaM).
    - ADRG
        - PDF version is required.
        - Sometime docx version is also required based on specific customers.
        - We need to confirm with customer if PMDA-specific ADRG is required.
    - SAS transport files in XPT format (for both ADaM and external datasets)
        - If any datasets are over size limit of 5GB, the split datasets should also be delivered.
    - Pinnacle 21 reports (3 reports: datasets only, Define.xml only, Datasets and Define.xml combined)
        - SAS programs (in ASCII text format if required for submission). We need to also confirm SAS program requirement prior to the delivery.

SOME OTHER IMPORTANT POINTS TO CONSIDER:

Faroz, Lyma (2021) in his article mentioned some important points that needs to be considered for effective and high-quality submission package.

- For the define.xml to open in a browser with the expected display format, the define.xml file and its associated style sheet must be in the same location.
- Check define.xml's 'Comment' or 'Method' column for any truncations, especially for variables with lengthy derivation logic. Check if all the hyperlinks work and open as desired.
- The reviewer's guide is a useful area to record significant research-related information as well as structural or logical decisions made during the study that relate to the data set and should be communicated to the reviewers of a submission.
- Verify that every hyperlink functions properly throughout the whole document in its final place, not just where you first authored the reviewer's guide.
- Any unsolved errors, warnings, or notices should be properly explained in the "Issue Summary" section so that reviewers may see why the issue wasn't fixed prior to submission.
- The reviewer's guide should be named 'csdrg.pdf and adrg.pdf respectively for SDTM and ADAM.

CONCLUSION

As difficult a process as it may appear, early and thorough planning is crucial to producing an FDA-acceptable electronic submission package. It is crucial to become familiar with and knowledgeable about the FDA's guidelines. Sponsors can produce a top-notch package for regulatory bodies with the aid of effective preparation, understanding, and numerous tools and technologies made available to industry by companies like Pinnacle 21 and others. The two critical recommendations are spending time to clean spec as described in this paper and make sure the documentation should happen simultaneously in parallel much earlier point in the cycle, rather than having it occur months or years later.

ACKNOWLDGEMENT

The author would like to thank colleagues of the AZ ISS team at IQVIA especially Reena Khurana, Lauren Murray and MyLinh Nguyen for their inputs and suggestions for improving the draft.

REFERENCES

CDISC.org. Define.xml, **https://www.cdisc.org/standards/data-exchange/define-xml**

Fu, Qiuping, 2023 Generate perfect define xml v2.1 with Analysis Results Metadata using python. PharmaSUG China 2023 - Paper PO-131.

Faroz, Lyma. First Time Creating a Submission Package? Don't Worry, We Got You Covered! PharmaSUG 2021 - Paper EP-070.

Connolly, Christine et al. Creating the Case Report Tabulation (CRT) for an NDA submission at the absolute last moment – NOT.

CONTACT INFORMATION

Yoganand Budumuru
yogsmail@gmail.com