PhUSE 2018

Paper SI05

Define'ing the Future

Johan Schoeman and Nicola Perry, GlaxoSmithKline, Stockley Park, UK

ABSTRACT

Define XML v2 guidelines were first published in March 2013 with FDA support starting in August 2013 while stopping support for the previous version of Define XML v1 for studies started after March 15th, 2018.

This paper looks to discuss the differences between ADaM Define v1 and v2 and how the changes impact the Analysis Data Reviewer's Guide (ADRG). We will also highlight how GSK implemented the changes within our internal systems and standards with the aim of increasing efficiency whilst maintaining quality and providing consistency. Finally, we are providing some hints and tips for effective production of Define xml v2 based on our experience and give you our proposal for a better CRT experience.

INTRODUCTION

One of our core principles at GSK is to focus on quality culture, governance and compliance. To support this principle, our in-house systems and documentation have been updated to produce the ADaM Define.xml document following the version 2 implementation guide.

The paper is broken down in three core areas. The first is where we provide insight of what the changes are between Define v1 and Define v2, how we at GSK aim to provide consistency across our studies with focus of quality and finally we provide some hints and tips with our experience of producing and reviewing CRT packages.

DISCLAIMER

The scope of this paper is to highlight the changes in Define v2 with respect to ADaM and how these have been implemented and as such any opinions, views or suggestions stated within this paper are those of the author.

It is understood that a review of this paper may invoke a reaction of 'this goes against the guidelines.' Hopefully it will also invoke a debate / discussion of some of our decisions.

DIFFERENCES BETWEEN ADAM DEFINE V1 AND ADAM DEFINE V2

There are five areas where there are changes between Define v1 and Define v2. They are Format, Origin, Controlled Terminology, Value Level Metadata (VLM) and what impact it has on Analysis Data Reviewer's Guide (ADRG).

We will be looking at each one individually and provide details of the changes.

Format

A new Stylesheet is being presented that incorporates the following information:

- Analysis Data Reviewer's Guide (ADRG) (link to external document)
- Analysis Results Metadata (ARM) including link to relevant display which is an external document
- Analysis Datasets
- Parameter Value Level Metadata (VLM)
- Controlled Terminology (CT)
- Analysis Derivations
- Comments

Analysis Results Metadata (ARM)

ARM information is now included within the Define.XML v2 as an appendix, rather than a separate document in Define.XML v1.

Analysis Datasets

This section is essentially two parts, one for the ADaM dataset definitions followed by the variable level metadata within each ADaM dataset.

The ADaM dataset definition section has been modified as follows:

- A new 'Documentation' column has been included where the user can optionally include ADaM dataset level information, including links to the ADRG.
- The order of datasets follows the dataset Class – ADSL, OCCURRENCE DATA STRUCTURE (e.g. ADAE), BASIC DATA STRUCTURE, and ADAM OTHER and then alphabetically within Class.

Dataset	Description	Class	Structure	Purpose	Keys	Location	Documentation
ADSL	Subject-Level Analysis	SUBJECT LEVEL ANALYSIS DATASET	One record per subject	Analysis	USUBJID, STUDYID	adsl.xpt	Includes screen failures

The format of the Define document is dependent on the stylesheet used and at GSK we use the stylesheet provided by CDISC. In Define v1 the 'Length/Display Format' column, although part of the define code, was not displayed in the variable level metadata. For Define v2, the CDISC stylesheet has been modified so this information is now displayed and the 'Comment' column in Define v1 has been changed to 'Source/Derivation/Comment'.

Variable	Label	Type	Length / Display Format	Controlled Terms or Format	Source/Derivation/Comment
STUDYID	Study Identifier	text	12		Predecessor: DM.STUDYID
USUBJID	Unique Subject Identifier	text	11		Predecessor: DM.USUBJID
SUBJID	Subject Identifier for the Study	text	4		Predecessor: DM.SUBJID
SITEID	Study Site Identifier	text	3		Predecessor: DM.SITEID
SITEGR1	Pooled Site Group 1	text	3		Derived: refer to SAP, Section 7.1 - if not pooled then SITEGR1=SITEID. If pooled, SITEGR1 will be 900

Value Level Metadata (VLM)

The format of this section has changed quite a bit and is more in line with the format of the VLM. The new layout has columns for 'Variable', 'Where', 'Type', 'Length or Display Format', 'Controlled Terms or Format' and 'Source/Derivation/Comment'.

Parameter Value List - ADTTE [AVAL]

Variable	Where	Type	Length or Display Format	Controlled Terms or Format	Source / Derivation / Comment
AVAL	PARAMCD = "IPDISC" (Study Treatment Discontinuation (days))	integer	8		Derived: AVAL=ADT - STARTDT + 1

Controlled Terminology

The Define v1 format used to display Code Value and Text even if the terminology was identical. In Define v2, the terminology is only displayed once under 'Permitted Value (Code)'. If variables included in an ADaM utilise standard CDISC terminology for SDTM or ADaM then the NCI codes are displayed as part of the controlled terminology.

APERIOD. [CL.APERIOD.]
APERIOD.

Permitted Value (Code)	Display Value (Decode)
1	Treatment Period 1
2	Treatment Period 2

APERIODC. [CL.APERIODC.]
APERIODC.

Permitted Value (Code)
Treatment Period 1
Treatment Period 2

DTYPE. [CL.DTYPE., *C81224*]
DTYPE.

Permitted Value (Code)	Display Value (Decode)
AVERAGE [*C81209*]	Average of Value Derivation Technique
IMPUTED [*]	Imputed from character value
LVPD [*C132342*]	Last Value Prior to Dosing Imputation Technique
WOCF [*C81199*]	Worst Observation Carried Forward Imputation Technique

* Extended Value

Computational Algorithms

This was the section in Define v1 that displayed lengthy or complex derivations for a particular variable with a link to the specific algorithm within the 'Comment' column for the variable it was explaining in the variable level metadata section.

In Define v2, computational algorithms have been replaced with a Computational Method element which is required for items (variables and VLMs) that have Origin="Derived". They are in separate elements in the background, but the stylesheet aggregates them into the 'Source/Derivation/Comment' column in the variable/value level metadata section. Define v2 includes a summary of all algorithms/derivations for variables entered with Origin='Derived' under the 'Analysis Derivations' section and a summary of all comments for variables entered with Origin='Assigned' under the 'Comments' section.

Analysis Derivations

Method	Type	Description
MT.ADAE.ADURC	Computation	ADURN converted to display in days, hours and minutes

Comments

CommentOID	Description
COM.ADAE.AEOUTN	Numeric version of AEOUT

Origin

The inclusion of Origin as part of the variable level and value level metadata is a new requirement for Define v2. For ADaM there are three possible values: Predecessor, Derived and Assigned. Origin can be null at the variable level metadata level only if a VLM is used, in which case Origin must be specified in VLM. The Origin value appears as part of the text in the 'Source/Derivation/Comment' column in the Analysis Datasets variable level metadata section.

SITEID	Study Site Identifier	text	3		Predecessor: DM.SITEID
SITEGR1	Pooled Site Group 1	text	3		Derived: refer to SAP, Section 7.1 - if not pooled then SITEGR1=SITEID. If pooled, SITEGR1 will be 900

The inclusion of Origin required an update to our in-house systems to include a field specifically for Origin and to assign a VLM name. In addition, all our standard ADaM metadata was modified to include a value of Origin, which provides guidance on how to choose the most appropriate Origin value.

Controlled Terminology (CT)

If any variables within an ADaM utilise CDISC controlled terminology whether from SDTM or ADaM then the NCI codes need to be included as part of the CT. This is one of the biggest changes that has been implemented for ADaM's in Define v2. The CT should also highlight if any values within the CT are extensible.

UNIT. [CL.UNIT., *C71620*]
UNIT.

Permitted Value (Code)
HRS [*]
YEARS [*C29848*]
mg [*C28253*]
ug [*C48152*]

* Extended Value

The inclusion of NCI codes required an update to our in-house systems to include a field specifically for the codes and a checkbox to highlight extended values. In addition, all our standard ADaM metadata was modified to include the relevant NCI codes.

Value Level Metadata (VLM)

Instead of having a 'Value' and 'Label' column as in Define v1, which usually showed PARAMCD and PARAM, Define v2 has a 'Where' column. In Define v1 each different parameter value had to be included in separate rows but in v2 the 'Where' column is designed to allow multiple parameter values if the 'Source/Derivation/Comment' is identical for each. It also has the ability to base rules off of variables other than PARAMCD, for example PARCAT1 EQ "HEMATOLOGY", which clears the way for much more flexibility for conditional definitions on all variables in an ADaM. The format of the VLM follows variable level metadata closely in that an Origin needs to be defined for each row, which means that values within a parameter can have different Origins.

When defining the 'Where' column, it is a requirement to use system independent operators in the xml code, such as NE, IN and EQ, so the VLM is displayed correctly in the Define document.

Example of VLM incorporating multiple parameters into one row:

Parameter Value List - ADQS [AVAL]

Variable	Where	Type	Length or Display Format	Controlled Terms or Format	Source / Derivation / Comment
AVAL	PARAMCD IN ("CGICE101" (Change in Severity of Eczema) , "CGICE102" (Change in Severity of Itch) , "CGICE103" (Rate Change in Severity of Eczema))	integer	8	QSAVAL1C [7 Terms]	Assigned: 1 WHERE QS.QSSTRESC EQ "VERY IMPROVED",2 WHERE QS.QSSTRESC EQ "MODERATELY IMPROVED",3 WHERE QS.QSSTRESC EQ "MINIMALLY IMPROVED",4 QS.QSSTRESC EQ "NO CHANGE",5 QS.QSSTRESC EQ "MINIMALLY WORSE", 6 QS.QSSTRESC EQ "MODERATELY WORSE", 7 QS.QSSTRESC EQ "VERY WORSE". (NOTE : Scores are Assigned as per CRF)

Example of VLM using conditional definitions

Variable	Where	Type	Length or Display Format	Controlled Terms or Format	Source / Derivation / Comment
AVALC	PARAMCD = "PHWRES" (PHASE: WITHDRAWAL REASON) and DSSCAT = "PHASE CONCLUSION" and DSDECOD NOT IN (" "UNKNOWN"", "COMPLETED")	text	200	NCOMPLT. [12 Terms]	Predecessor: DS.DSDECOD

The modifications to VLM format, including the addition of Origin, required an update to our in-house systems to accept a 'Where' clause and to include a field specifically for Origin. In addition, our systems now link to the CT for variables specified in the 'Where' column, such as PARAMCD, and adds the decode.

Impacts to the Analysis Data Reviewer's Guide (ADRG)

Due to the loss of the computational algorithms section with the Define document complex and/or lengthy derivations are displayed at the variable level metadata level within the Analysis Datasets section. GSK made the decision to utilize the ADRG as the place to include these derivations either as part of the Analysis Dataset description in ADRG Section 5.2.x or as an Appendix (Section 8). The relevant section of the ADRG becomes the entry in the Source/Derivation/Comment column in the Define document.

DEFINE'ING THE FUTURE AT GSK – CONSISTENCY AND QUALITY

When reporting a study, we have found there are many things that can impact the consistency and quality of the CRT package. The main ones we found were -

- Multiple programmers
- Multiple reviewers
- Change of study team members during a study
- Lack of adherence to in-house standards
- Lack of adherence to CDISC standards
- Lack of training
- Aggressive timelines

In order to minimize this impact, we have devised a number of methods to try and improve quality and consistency:

- Standard metadata for all common ADaM domains has been updated to include the NCI codes for controlled terminology and an Origin value. This can be accessed by both GSK programmers and CRO's whom are using our systems to produce eCRT packages.
- Include a review of the eCRT package by an experienced independent programmer. Our aim to review a package at the Dry Run (data look) stage.
- Training courses developed specific to Define v2 production, eCRT package development and effective review of a eCRT package.
- Our in-house system has been updated to include built in validation specific to Define v2.
- At GSK it is mandatory to run Pinnacle21 on ADaMs (AD-- P21 IDs) and Define.xml only (DD-- P21 IDs). However, if you are using the Enterprise version, it is not an issue,
- Focus on quality over time – The Define package is not left to be completed at the end of the study. It is worked on throughout the entire process.

HINTS AND TIPS

Here is a selection of hints and tips to help you with the production of an eCRT package including Define v2.

- Start creating the metadata for the CRT package as soon as final specifications are agreed and programming commences.
- Work on the ADRG alongside Define to ensure that information is consistent.
- Get the eCRT package reviewed by experienced independent programmer. They'll see the study in the same way as the external reviewer.
- Rerun and check the Pinnacle21 report through the course of the study on both SDTM and ADaM datasets.
- The definition of Origin can be very programmer dependent. If available, create or update standard ADaM metadata to include an Origin value to improve consistency.

  Here's our take on defining Origin…

  Assume you have a box containing a strawberry, a piece of cheese and a steak. The box represents a dataset and the food items are variables.

  You want to move the food items into a different box. If you do so without altering the food at all then this is an example of a PREDECESSOR.

  You want to move the food items into a different box but this time you want to separate them into different food categories with the strawberry slotted into 'Fruit', the cheese into 'Dairy' and the steak into 'Meat'. The categories represent a variable that would have an Origin of ASSIGNED.

  Lastly, you want to move the food items into a different box but this time you want to take the strawberry and cheese and make a strawberry cheesecake and take the cheese and the steak and make a cheeseburger. The cheesecake and cheeseburger are examples of variables that would have an Origin of DERIVED.

CONCLUSION

The processes put in place within GSK, including system updates, documentation updates and staff training, have enabled a smooth and efficient transition into Define v2. The main focus is to produce quality eCRT packages that are up to date with regulatory requirements. The requirements are ever evolving and the specification for Define v2.1 is out for review. Here is a list of possible updates coming soon:

- Identification of what version of CDISC/NCI Controlled Terminology is used.
- Identification of non-standard domains and variables.
- Enhanced Origin to include 'Source'. For ADaM datasets, if Origin Type is 'Derived' or 'Assigned' then both Type and Source attributes are required. The value of Source will be 'ADaM'.
- Using the Alias child element with Context =" SAS" for SAS names longer than 8 characters.
- Various attributes in Define-XML have a list of allowable values e.g. def:Origin/@Type. These enumerations are now part of the XML schema.

CONTACT INFORMATION

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

Johan Schoeman
GlaxoSmithKline
1-3 Iron Bridge Road
Stockley Park West
Uxbridge
Middlesex UB11 1BU
johan.x.schoeman@gsk.com

Nicola Perry
GlaxoSmithKline
1-3 Iron Bridge Road
Stockley Park West
Uxbridge
Middlesex UB11 1BU
Nicola.x.perry@gsk.com