PhUSE 2018

Paper SI25

A proprietary, CDASH/SDTM-hybrid data model to expedite clinical data review

Lieke Gijsbers, OCS Life Sciences, 's-Hertogenbosch, the Netherlands
Tom Van der Spiegel, Janssen R&D, Beerse, Belgium

ABSTRACT

In 2016 Janssen identified the need to expedite clinical data review. A proof of concept demonstrated the value of pursuing a new and proprietary data model for data review, serving as single source of truth. The Data Review Model (DRM) that was introduced is strongly based on CDISC SDTM and CDASH. DRM provides full traceability and describes both clinical and operational (system) data consistently across studies. On the longer term, Janssen plans to implement a metadata-driven environment, including data conversion from source data into DRM.

In 2017, OCS Life Sciences and Janssen piloted DRM by implementing a mapping framework that supports both documentation and execution of source to target data mapping. This paper will describe how multiple trials were mapped to support the pilot phase of DRM, to learn, refine and document the value of DRM prior to moving to production implementation.

INTRODUCTION

Some history, framing the picture. At Janssen, Data Management (DM) activities are outsourced to DM CROs. These DM CROs are contracted to deliver SDTM datasets to Janssen during trial conduct and to prepare the SDTM Submission Package after Database Lock. Janssen DM performs ongoing Quality Control on these SDTM deliverables.

In 2016, spending time evaluating the current data flow, Janssen identified the need to expedite clinical data review. The idea of a new data model was introduced, serving as single source of the truth to all consumers of clinical data. What followed was a proof of concept (POC), showcasing 5 newly designed domains. The Data Review Model, in its very early stages, was born. As part of the POC we tried out several use cases:

- How can we most logically cluster/group information in DRM? Avoiding the use of SUPPQUAL and Findings About datatypes.
- How to represent relationships in DRM, without the need for RELREC as we know it from SDTM?
- Can we add value by adding new data (variables or datasets) in DRM not possible to include in SDTM?
- How will DRM help when mapping a new exploratory data stream?
- Will DRM allow an easy transformation to SDTM?

As outcome of the POC we noted the following conclusion:

- Having a Janssen owned/controlled proprietary data model and related controlled terminology offers these advantages:
    - A less complex data model compared to the CDISC SDTM model, focuses on data review activities and not on data submission.
    - A Janssen controlled DRM, both structure and content, enables the operations to be less vulnerable to the changing CDISC SDTM versions. The DRM controlled terminology maps to the CDISC/NCI CT to achieve stability and isn't changing with each CDISC/NCI quarterly update.
    - DRM stores additional 'value added' content, like operational tracking data (e.g. AEYN), data points documenting the traceability to the source, additional cleaning identifiers, conventional results, additional coding data points, etc.
    - DRM is not a data submission model and is less strict on implementation at trial level. Complex data streams can be delivered in DRM in stages. First a 'quick and dirty' DRM dataset suited for immediate consumption that is later harmonized with all other datasets.

- We also noted several other learnings:
    - Enabling early access to the data in the DRM model requires a high-degree of re-use, from standard or previous trials.
    - Data harmonization will require a controlled environment (i.e. process, tool and resources) to enable a consistent application of CT in DRM and SDTM.
    - Getting from DRM to SDTM was relatively easy since DRM adheres to many CDASH and SDTM design principles and business rules.

Following the positive outcome of the POC, we moved the project into the next phase, the pilot phase. In 2017-2018 we conducted 13 pilot studies, spread over 3 waves, starting small with only a handful of DRM domains focused on Data Management pilot teams, and gradually when moving to the next wave we increased the scope by adding new domains and including other functional groups to the pilot teams.

With this paper we try to highlight some of the key principles of the Janssen Data Review Model, give you some insights into the conversion framework used during the DRM pilot phase, as well as describe some learnings, next steps and future perspectives.

DATA REVIEW MODEL

The Data Review Model describes both clinical and operational (system) data and is strongly based on CDISC CDASH (standard for data collection) and SDTM (standard for data tabulation). In general, the data in DRM are presented in a structure that is similar to SDTM, i.e. in a vertical structure with one record per finding, event or intervention for each time point. Hence, DRM adheres to the fundamentals of SDTM:

- Built around the concept of observations collected about subjects who participated in a clinical study.
- Each observation can be described by a series of variables, corresponding to a row in a dataset or table.
- Observations are reported in a series of domains, usually corresponding to data that were collected together. Each domain is represented by a single dataset.
- Dataset and variable names are standardized according to DRM naming conventions.

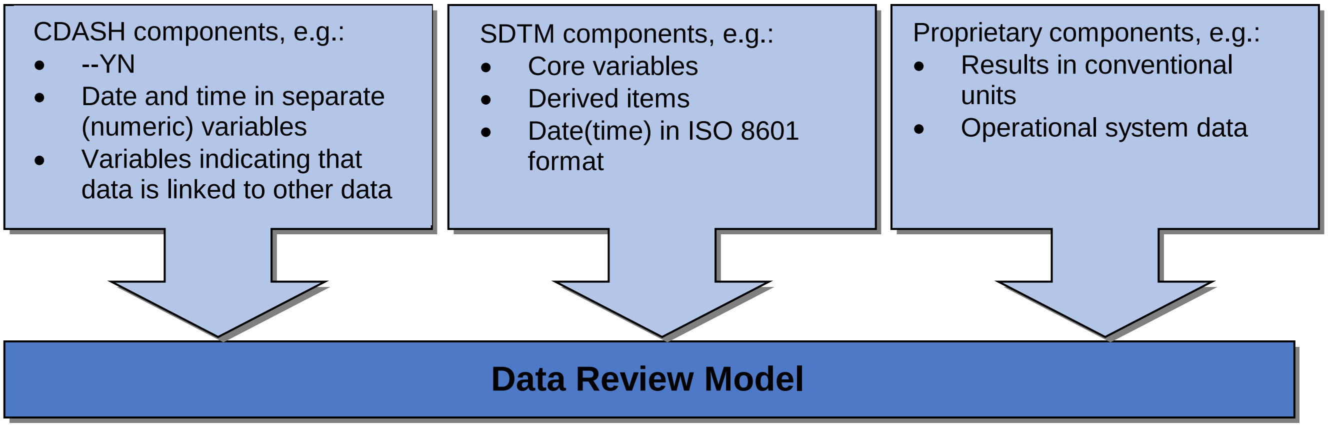

CDASH components, e.g.:
• --YN
• Date and time in separate (numeric) variables
• Variables indicating that data is linked to other data
SDTM components, e.g.:
• Core variables
• Derived items
• Date(time) in ISO 8601 format
Proprietary components, e.g.:
• Results in conventional units
• Operational system data
Data Review Model

Figure 1. Data Review Model components

Some key principles of DRM (see also Figure 1):

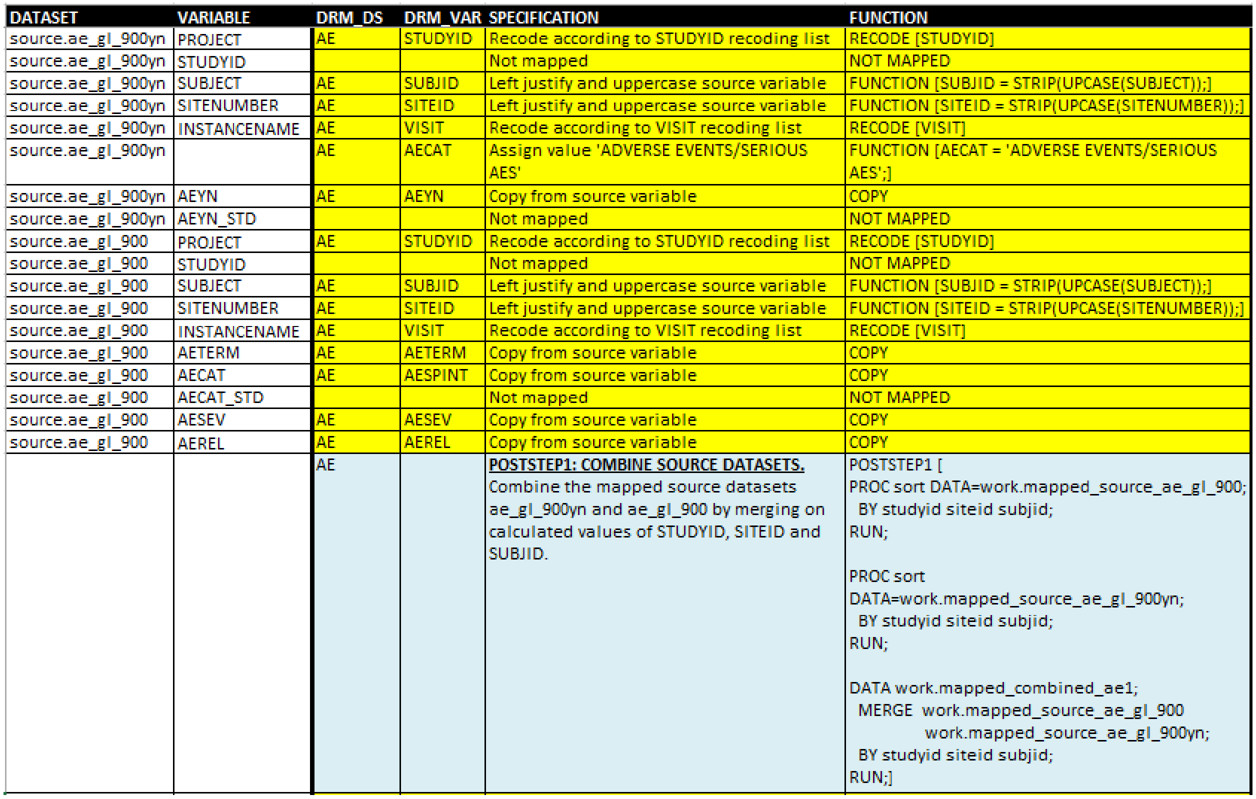
- In general, DRM contains all core SDTM variables
- DRM includes the CDASH recommended field --YN (with the two dashes (--) replaced by two-character domain code) which captures answers on questions like “Were any adverse events experienced?”, “Were any medications taken?”, “Were examination performed?” or “Was sample taken?”. The variables can be used in the verification that no data are unintentionally missing and hence support data review and monitoring.
- DRM contains SDTM like derived items as study day (variable --DY), results in standardized units, lab normal range indicator and dictionary-derived variables (e.g. AEHLGT, AEHLGTCD). Besides these derived items, DRM also contains proprietary derived items, e.g. results in conventions units.
- In DRM three variables are available representing date and/or time. Based on CDASH, date (variable --DAT) and time (variable --TIM) are represented in two (numeric) variables, and based on SDTM, date and time are represented together in ISO 8601 format (variable --DTC). Having the date and time in separate numeric fields makes DRM user-friendly, having the date time in ISO 8601 format, makes the conversion to SDTM easy.
- DRM does not have any supplemental qualifier domains, but data are stored in additional, proprietary, variables in the parent domain.

- In SDTM collected relationships between data (e.g. for example that medication was taken for an adverse event) are represented in RELREC, a separate dataset. In DRM, just like in CDASH, links among records are explicit available in the parent domain.
- Besides clinical data DRM also describes operational (system) data, such as the source of the raw data, the CRF page name and page repeat number, a unique record identifier, and a date indicating when the record was initially created and last updated. The purpose of these variables is to facilitate full traceability to the collected source data.

DRM MAPPING

MAPPING FRAMEWORK

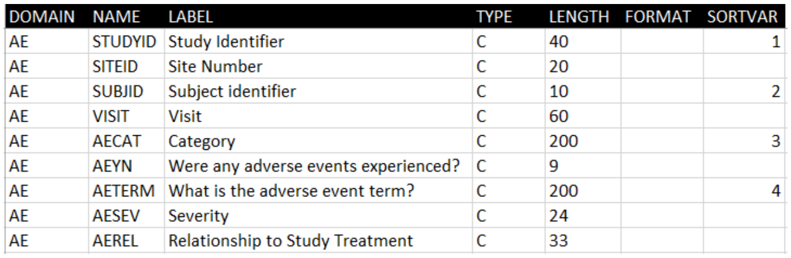
The conversion of source data into DRM was performed using the mapping framework which was developed by OCS Life Sciences and implemented in Janssen's SAS® Life Science Analytics Framework (LSAF). The mapping framework describes and executes source-to-target mappings in a structured way and because of its modular structure, the framework can be implemented with minimum effort. A full description of the mapping framework is available in the paper of Bas van Bakel, 2016 [1]. In short, there is one single Microsoft Excel spreadsheet (i.e. mapping specification document) that contains all source-to-target specifications and, directly next to them, the translation of these specifications into code or pseudo-code. An example of this mapping specification document is shown in Figure 2. The white columns contain the source datasets and variables. The yellow columns contain the target DRM datasets and variables and the specifications and (pseudo-)code to convert the sources to DRM. In the post steps 'mapped' source datasets can be combined and further processed if needed (see the blue cells in Figure 2). Macros are available to translate the (pseudo-)code available in the column 'FUNCTION' into SAS code, execute that SAS code in a specific order and output the DRM dataset with attributes (type, length, label) and the order of variables and records as defined in the 'target metadata'. An example of this 'target metadata' is available in Figure 3.



DATASET	VARIABLE	DRM_DS	DRM_VAR	SPECIFICATION	FUNCTION
source.ae_gl_900yn	PROJECT	AE	STUDYID	Recode according to STUDYID recoding list	RECODE [STUDYID]
source.ae_gl_900yn	STUDYID			Not mapped	NOT MAPPED
source.ae_gl_900yn	SUBJECT	AE	SUBJID	Left justify and uppercase source variable	FUNCTION [SUBJID = STRIP(UPCASE(SUBJECT));]
source.ae_gl_900yn	SITENUMBER	AE	SITEID	Left justify and uppercase source variable	FUNCTION [SITEID = STRIP(UPCASE(SITENUMBER));]
source.ae_gl_900yn	INSTANCENAME	AE	VISIT	Recode according to VISIT recoding list	RECODE [VISIT]
source.ae_gl_900yn		AE	AECAT	Assign value 'ADVERSE EVENTS/SERIOUS AES'	FUNCTION [AECAT = 'ADVERSE EVENTS/SERIOUS AES';]
source.ae_gl_900yn	AEYN	AE	AEYN	Copy from source variable	COPY
source.ae_gl_900yn	AEYN_STD			Not mapped	NOT MAPPED
source.ae_gl_900	PROJECT	AE	STUDYID	Recode according to STUDYID recoding list	RECODE [STUDYID]
source.ae_gl_900	STUDYID			Not mapped	NOT MAPPED
source.ae_gl_900	SUBJECT	AE	SUBJID	Left justify and uppercase source variable	FUNCTION [SUBJID = STRIP(UPCASE(SUBJECT));]
source.ae_gl_900	SITENUMBER	AE	SITEID	Left justify and uppercase source variable	FUNCTION [SITEID = STRIP(UPCASE(SITENUMBER));]
source.ae_gl_900	INSTANCENAME	AE	VISIT	Recode according to VISIT recoding list	RECODE [VISIT]
source.ae_gl_900	AETERM	AE	AETERM	Copy from source variable	COPY
source.ae_gl_900	AECAT	AE	AESPINT	Copy from source variable	COPY
source.ae_gl_900	AECAT_STD			Not mapped	NOT MAPPED
source.ae_gl_900	AESEV	AE	AESEV	Copy from source variable	COPY
source.ae_gl_900	AEREL	AE	AEREL	Copy from source variable	COPY

		AE		**POSTSTEP1: COMBINE SOURCE DATASETS.**
				Combine the mapped source datasets ae_gl_900yn and ae_gl_900 by merging on calculated values of STUDYID, SITEID and SUBJID.
					POSTSTEP1 [
					PROC sort DATA=work.mapped_source_ae_gl_900;
					BY studyid siteid subjid;
					RUN;
					
					PROC sort
					DATA=work.mapped_source_ae_gl_900yn;
					BY studyid siteid subjid;
					RUN;
					
					DATA work.mapped_combined_ae1;
					MERGE work.mapped_source_ae_gl_900
					work.mapped_source_ae_gl_900yn;
					BY studyid siteid subjid;
					RUN;]

Figure 2. Mapping specification document containing all source variables, all target DRM variables, and the mapping specifications and (pseudo-)code needed to generate the DRM variables from the sources.



DOMAIN	NAME	LABEL	TYPE	LENGTH	FORMAT	SORTVAR
AE	STUDYID	Study Identifier	C	40		1
AE	SITEID	Site Number	C	20		
AE	SUBJID	Subject identifier	C	10		2
AE	VISIT	Visit	C	60		
AE	AECAT	Category	C	200		3
AE	AEYN	Were any adverse events experienced?	C	9		
AE	AETERM	What is the adverse event term?	C	200		4
AE	AESEV	Severity	C	24		
AE	AEREL	Relationship to Study Treatment	C	33		

Figure 3. Target metadata

DRM IMPLEMENTATION PROCESS

Janssen developed the Data Review Model and defined the ‘target’ metadata and the specifications for the mapping of source data to DRM domains. Based on the specifications, OCS Life Sciences populated in close collaboration with Janssen the mapping specification document for the 17 domains in scope, which were spread over 3 waves. Since Janssen works in a standardized environment, first a standard (global) mapping library was created containing re-usable mappings. These standard mappings were subsequently implemented in 13 trials, again spread over 3 waves. On a trial level the mapping specification document was adjusted according to trial specific input data and trial specific mapping specifications. A SAS program facilitated this process by comparing the available trial source data with the standard (global) source data. It checks:

- Which source datasets and variables are available at a global level and not at a trial level? Mapping specifications that are related to the global datasets and variables that are not available on a trial level were removed from the trial mapping specification document and if needed post steps were updated.
- Which source datasets and variables are available on a trial level but not on a global level? For new source datasets and variables mapping specifications and corresponding (pseudo-)code were added to the trial mapping specification document and if needed post steps were updated.
- Differ common variables on global and trial level in attributes (e.g. type, length)? Mapping specifications and the (pseudo-)code were adjusted in the trial mapping specification document in case the source variables differed in type; in the ‘target’ metadata the length of the DRM variable was adjusted in case values exceeded the length in the ‘initial’ metadata.

After the creation of the trial mapping specification document, the ‘target’ metadata was adjusted on a trial level. A SAS program facilitated this process by removing the metadata of DRM variables that were removed from the trial mapping specification document (i.e. if the variable was not available in the DRM_VAR column in the mapping specification document, the metadata was removed from the trial ‘target’ metadata). For DRM variables that were added or adjusted on a trial level, Janssen provided the required attributes, and these were manually incorporated in the ‘target metadata’.

Upon this, initial trial DRM datasets were created and reviewed by Janssen. After review and approval, the DRM datasets were created and monitored on a daily basis during trial execution by scheduled jobs in LSAF. Despite an initial successful creation of the DRM datasets, data conversion failures could still occur for instance because source variables were added or removed, values exceeded the length in the ‘target’ metadata and were truncated, or the automatic upload of the source data failed and hence no source data was available and thus DRM datasets could not be created. The DRM creation process was therefore monitored on a daily basis, which involved the following checks:

- Whether a new log has been created. If not, then the job did not initiate correctly.
- Whether the log contained errors and/or warnings.
- Whether truncation of values has occurred. Values were truncated in case the value exceeded the length specified in the ‘target’ metadata.
- Whether the DRM datasets were updated.
- Whether the DRM datasets were populated.

In case any of the listed checks failed, an email with a notification, which is a LSAF specific service, was sent to the assigned person. No email was sent if all checks passed successfully.

In case truncation of values occurred, an Excel file was additionally created, which listed per truncated value: the dataset and variable name, observation number, the value before truncation, the length of the value before truncation, and the length of the value after truncation (i.e. the length specified in the ‘target metadata’).

If needed, the mapping specification document or ‘target’ metadata were updated according to the outcome of the daily monitoring.

This process was completed for all 13 trials in scope of the pilot project and the created DRM datasets were used in the review of the data and the evaluation of DRM.

SDTM MAPPING

For one trial, DRM data were converted into SDTM data as part of the pilot project. Since DRM is based on CDASH and SDTM, data could rather easily and smoothly be converted into SDTM datasets. In general, a selection of the records had to be made (e.g. where --TERM or --TRT is not missing) and most variables could directly be copied to SDTM. Only the creation of domains DM (in DRM these data are stored in a vertical structure) and RELREC (not available in DRM), and the derivation of the baseline flags (not available in DRM) took some effort, but not more than in ordinary source-to-SDTM conversions.

NEXT STEPS AND FUTURE PERSPECTIVES

Upon completion of the DRM pilot phase, an abundance of constructive feedback, lessons learned, and questions were received. Most feedback related to DRM design, requesting for updates to the domain models or asking for the introduction of new requirements. The pilot phase also highlighted the need for more documentation and introduced some questions on the ‘to be’ process. Some technical challenges were faced throughout the pilots, where the incoming data structure sometimes did not meet the expected input format or where we observed failing scheduled ‘data uploader’ jobs.

Overall, the pilots did successfully demonstrate the business value of the Data Review Model and it was decided to plan for a staged roll-out to facilitate early access to data for:

- Early medical review, in the Early Development/Clinical Pharmacology space, to enable Safety Review meetings
- Responding quickly to key critical data required at the beginning of trial, e.g. for central monitoring
- Ensuring full data traceability to the source and high data availability

To achieve the above use cases, the DRM project team is diligently working on:

- Refining the DRM domain models, business rules and implementation guidelines based on the lessons learned from the pilot teams.
- Janssen and OCS are collaborating to install the above-mentioned Mapping Framework ready for use in a production setting, introducing some enhanced functionalities.
- Preparing for a library of mapping rules for Janssen’s Data Capture standards (EDC + external transfers) to DRM.
- Process design, training framework, communication plan, etc.

In parallel, Janssen is working with OCS to introduce several new utilities to help find efficiencies in the DRM conversion process. Few examples,

- Janssen would like to introduce a metadata driven way to compose a draft mapping spreadsheet at trial level, by using the incoming source datasets and standards mapping spreadsheet as input. This will greatly reduce the manual actions to be taken during trial set up and only limit manual intervention to ‘true’ trial specific additions/changes.
- A fail-safe mechanism will be introduced to check the incoming source data prior to passing it on to the data conversion service. This to prevent the conversion from failing, or from producing unexpected output.

As first real live use case, Janssen is targeting roll-out of DRM to those trials in need for early access to data, for decision making and patient review. The normal data flow, engaging DM CROs to prepare the SDTM data packages, will continue in parallel.

CONCLUSION

To improve the data flow, following a successful proof of concept and pilot phase, Janssen is introducing a new data model to help expedite access to data and facilitate data review operations. This Data Review Model provides a general framework for describing clinical trial and operational data in a rather simple, well-structured and uniform way. It provides clear traceability to the collected source data, it positively impacts the review of data and it allows an easy and controlled transformation to SDTM.

ACKNOWLEDGMENTS

We thank Bas van Bakel, Jules van der Zalm, Inge Van den Broek, and Stefaan Vervloet for their contribution to this paper.

REFERENCES

1. Bas van Bakel, OCS Consulting, 's-Hertogenbosch, the Netherlands, DIY: Create your own SDTM mapping framework, PhUSE 2016, Paper CD03 (https://ocs-consulting.nl/wp-content/uploads/Bas-van-Bakel-Create-your-own-SDTM-mapping-framework-2.pdf)

CONTACT INFORMATION

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

Author Name	Lieke Gijsbers	Tom Van der Spiegel
Company	OCS Life Sciences	Janssen R&D
Address	Ruwekampweg 2G	Turnhoutseweg 30
City / Postcode	‘s-Hertogenbosch / 5222 AT	Beerse / 2340
Work Phone:	0031 (0)73 523 6000	0032 (0)14 60 2994
Email:	sasquestions@ocs-consulting.com	tvdspieg@its.jnj.com
Web:	www.ocs-lifesciences.com	https://www.janssen.com