PhUSE 2012

Paper CD02

Mapping the Company's Legacy Data Model to SDTM

Nicolas Dupuis, Anja Feuerbacher, Bruce Rogers
F. Hoffmann-La Roche, Basel, Switzerland

ABSTRACT

In the pharmaceutical industry standards like CDISC SDTM have become increasingly important in the past years. With evolving standards companies have to adapt their data models to meet requirements of authorities. For projects where some studies use SDTM and some do not, project teams need to convert the legacy data.

This paper shows two cases where the company's legacy data model was mapped to SDTM and how a standardized tool was used to solve the task. Typical problems and possible solutions will be displayed. The different approaches for validation will be described as well as the effort for mapping.

INTRODUCTION

As many other companies, Roche has to adapt to evolving requirements of authorities. In the case of data standards, this means the implementation of CDISC SDTM. The defined global data standards are being used for all new studies, while older ongoing studies may still be in the legacy data model. Depending on the project size and future development of a product, project teams need to decide whether old studies are reported in the legacy data model, or if it makes sense to convert the data to the new model and create analyses in alignment with future developments. This paper describes a toolset, that has been developed to support the mapping from one data model to any other data model. The implementation and problems will be explained with two examples. It will be shown, that a company can benefit from the usage of a self-explaining metadata driven tool, that is flexible enough to be adapted to study specific needs.

DATA MODELS

GDM

Roche's legacy **Generic Data Model** (GDM) provides a standard data structure for all clinical trial data, split into up to 15 domain datasets including DEMO, EFEX, LABP etc. For any given study it is not required that all domains will exist, and additional variables and domains may be used as required for any study or project (collection of studies).

Content and terminology is not strictly controlled, so any data analysis or transformation logic may need to be varied across studies. For example, different lab parameter names may be used for the same measure in different studies, or a similarly-named parameter may contain dissimilar values, notably with respect to units and terminology.

A safety reporting system (MARS) and an efficacy reporting system (Share) have been used, based on these structures, for some years, and in the case of MARS expects a certain level of conformity, but also allows the user to encode some study-specific variations within the analysis system and programs.

These domains are not exactly equivalent to the CDISC SDTM definitions, although in some cases there is a strong overlap, but generally the domain mappings are of a many to many type, with several source domains per target domain, and source domains used in multiple target domains.

SDTMv

Starting in 2010, the **Global Data Standards** (GDS) Integration Initiative was launched, whose goal was to define standards for data collection (i.e., content of eCRF), data tabulation (i.e., mapping of eCRF data items to SDTM), and data analysis (i.e., imputation and derivation rules). The data collection and data tabulation working groups worked closely with each other to align the standard content and structure, if possible compliant with CDISC CDASH and SDTM.
The final structure adopted as a result is known as SDTMv and has been in use since early in 2011. New studies since late 2011 have only been made available in this data structure and are no longer supplied in the legacy GDM format.

The SDTMv data model is SDTM compliant but includes Roche-defined extensions. The Roche-defined extensions conform with any rules provided in the SDTM IG and include the following:

- sponsor-defined domains
- controlled terminology: sponsor-defined code lists and extension of SDTM code lists with sponsor-defined terms
- character variable lengths: defined by the sponsor, between 1-200 characters
- supplemental qualifiers: For internal use sponsor-defined variables that are required to further qualify the data are included in their respective parent domain data sets.

A standard tool is available to take any SDTMv data and de-couple the supplemental qualifiers to produce CDISC-compliant SDTM datasets.

In Q3 2011 a new reporting tool Standard Reporting and Analysis Modules (STREAM) was first released. This tool provides standard macros for generation of analysis datasets and reports. Analysis datasets are created on the assumption that input data comply with SDTMv.

DATA MODEL METADATA

The newly-defined SDTMv model has now been encoded within a single central metadata repository, known as the Global Data Standards Repository. At the time of development described in this paper, this was not yet available, although the definitions themselves were already defined and in use. This consisted of a set of metadata for each SDTMv domain, including lists of controlled terminology and descriptions of all the data elements therein contained.

TRANSFORMATION NEEDS

It was decided that the effort involved in attempting to remap all the existing legacy data would not be worthwhile, given that we have something in the order of 8000 sets of study data. The variability between the GDM instantiation in different studies is known to be quite large, so a general one-off transformation would not be achievable.

Nonetheless, it was quite clear that there would be a significant overlap period, where data from both models would need to be combined to perform pooled analyses, as well as maintaining the original source data from previously submitted studies. Therefore a validated means of transforming data between models, allowing individual study mappings to be performed, would be required.

RULE BASED MAPPING

The traditional approach to clinical data transformation would mean writing individual mapping programs for each required output dataset, which would be validated within a single study. These programs would then be copied to each study area and amended as required for the specific data contained. This method works very well if a small number of studies is envisaged, or if a high level of homogeneity exists between the study data models. In our case this homogeneity is very low, and the potential number of different studies high.

Instead, it was decided to develop a rule-based toolset, which would build towards a single central set of transformation rules shared between all studies. The rules would be stored as metadata alongside the domain definitions, allowing a full set of possible derivation rules to be created for any given source and target data model, which would allow any study to select the appropriate rules used in previous study mappings with identical requirement or creation of new rules where different derivations are required

These rules, once compiled into a complete set for a given study/target domain, would be parsed by a tool which would use them to generate and execute SAS code to perform the actual transformations, in a single process from the source datasets to the required target datasets.

MAPTRANS

MAPTRANS is the working title of a tool which was developed in 2011/2012 as a pilot and has now been used by several study teams. It performs the rule parsing and SAS code generation and execution process described above, working from a spreadsheet containing the complete rule set for a given study/target domain. At this stage the toolset to store and maintain rules in the central metadata repository has not yet been implemented.

The tool has been fully validated and made available as part of Roche's Across Project Program (ACP) package. It is made up of SAS macros, which perform the actual rule parsing, as well as defining a number of standard mapping rules (e.g. ISO date creation). Currently the Rules are stored as spreadsheets in the form of .csv files, which take the place of programs in the normal analysis programming area and are subject to study-level version control and validation just as any SAS program would be.

Metadata Definitions

The rules are stored in a CSV file contains the following metadata:

- the target model (TMODEL) and domain (TDOMAIN), e.g. SDTM DM
- the source model and domain, e.g. GDM DEMO
  - The first line of the mapping defines the primary source domain, e.g. SDTM.DM is mapped primarily from GDM.DEMO.
  - Other source domains are treated as secondary.
- the target SDTM variables and their attributes (Variable Name, Label, Type, Length)
- the source variables or formulae (SVARIABLE, Formula/Keys)
- an optional SAS format to be applied as part of the data derivation.
- the source variable (SVARIABLE) from the source domain. This may be from the primary source domain or a secondary domain.
  - Primary: Performs a simple assignment from source to target
  - Secondary: uses additional metadata to create SQL to join with a secondary source domain
- Formula/Keys - used to
  - assign a literal value (e.g. "YEARS")
  - derive a value using base SAS code or macro call (e.g. %mapdc2iso(birthdc))
  - specify the keys to be used for matching on a secondary domain mapping.
  - Specify conditional assignments by inserting a condition starting with the keyword ‘case’ (inserts an SQL 'case' statement)
  - other SQL derivations prefixed by ‘Select’, ‘Join’ or ‘SQL’.

Fig. 1: A simple mapping rule set

TMODEL	TDOMAIN	Variable Name	Label	Type	Length	FORMAT	SMODEL	SDOMAIN	SVARIABLE	Formula/Keys
SDTM	DM	STUDYID	Study Identifier	Char	200		GDM	DEMO	PROTO	
SDTM	DM	DOMAIN	Domain Abbreviation	Char	2					"DM"
SDTM	DM	USUBJID	Unique Subject Identifier	Char	50		GDM	DEMO	proto||crtn||PT	%GDM2SDTM_SUBJID
SDTM	DM	SUBJID	Subject Identifier for the Study	Char	50		GDM	DEMO	PT	
SDTM	DM	RFSTDTC	Subject Reference Start Date/Time	Char	19		GDM	DEMO	TRT1DC TRT1TC	%dc2iso(TRT1DC) || 'T' || trim(TRT1TC)
SDTM	DM	RFENDTC	Subject Reference End Date/Time	Char	19					""
SDTM	DM	SITEID	Study Site Identifier	Char	200		GDM	CENT	CTCNUM	proto crtn
SDTM	DM	BRTHDTC	Date/Time of Birth	Char	19		GDM	DEMO	BIRTHDT	
SDTM	DM	AGE	Age	Num	8		GDM	DEMO	AGE	
SDTM	DM	AGEU	Age Units	Char	200					"YEARS"
SDTM	DM	SEX	Sex	Char	1		GDM	DEMO	SEX	
SDTM	DM	RACE	Race	Char	200	$race	GDM	DEMO	RACE	
SDTM	DM	ETHNIC	Ethnicity	Char	200					"NOT REPORTED"
SDTM	DM	ARMCD	Planned Arm Code	Char	20		GDM	DEMO	RNDGRP	
SDTM	DM	ARM	Description of Planned Arm	Char	200		GDM	DEMO	RND	
SDTM	DM	COUNTRY	Country	Char	3		GDM	CENT	CTCNTRY	proto crtn
SDTM	DM	CRTN	Country	Num	8		GDM	DEMO	CRTN	

Additional rule parameters (not shown) include

- Temporary variables. Where a derivation is complex or used several times it may be useful to create a temporary value. The derived value may then be used by following derivations. Variables created with the prefix _t will not be retained on the output dataset.
- PreProcess flag. Ensures that the derivation indicated is performed as a preprocessing step before the main transformations are carried out. This rule must be specified as a standalone SAS Macro call. This may be used, for example, where there is no unique subject id on the source data model, so it must be generated but consistent across domains.
- Repeat block. A set of variables may be repeated, with a flag R1 for the first set, R2 for the second, etc. and different derivation rules for each set. to indicated that these mappings will produce multiple occurrences of the output record, with all non-flagged variables simply repeated with the same values for each set of the flagged variables. Used particularly for Parameter/Value sets.

Rule dependencies and sequencing

Due to the nature of the rule parsing process, a certain level of dependency between mapping rules is inevitable. In general, the order in which the rules are specified will be the order in which they are processed, but the use of SQL statements does not follow this rule, to the sequence of execution of the different mapping types is as follows:

1. Preprocessing.
2. SQL rules, including both those using a secondary source domain and those with Case, Join, SQL, or Select keywords as the start of the Formula.
3. All other non-repeating derivations
4. Those blocks of variables marked as Repeating.

Additionally, certain constraints between various mapping types may also exist. In particular, Any variables in Repeat blocks may not use any of the SQL operators, as they are parsed and eventually executed within a SAS data step. If any of the SQL operations are required, the use of a temporary variable to perform the derivation and then to be assigned directly to the 'Repeat' variable, is indicated.

IMPLEMENTATION

CASE 1: DSMB ACTIVITIES

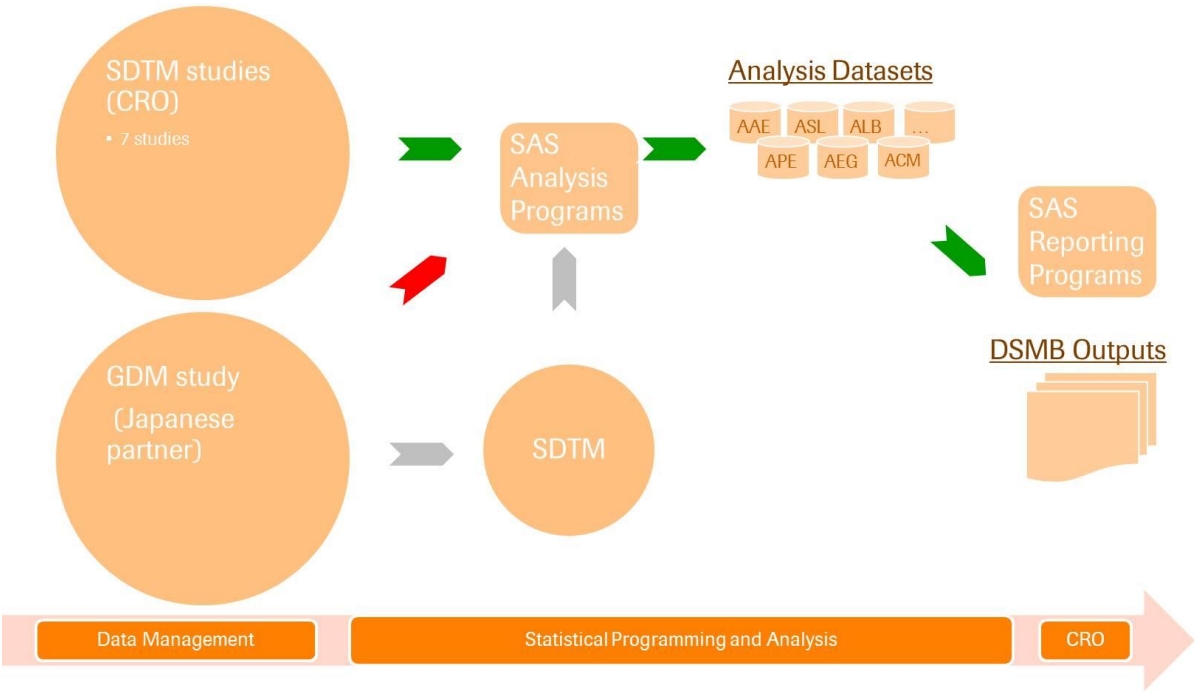
The programing department was asked to pool different studies from a CNS project. To do this, the project team had to create value added analysis datasets (VADs) for seven studies, all conducted in SDTM. An external CRO was then creating outputs for the DSMB (Data Safety Monitoring Board).

After a while another study in the same project, but this study was still conducted in GDM by the Japanese partner in Chugai. The project team had to add this study to the DSMB pool. One possibility was to write dedicated VAD programs, but the team thought it could become painful to maintain two sets of programs. The alternative was to map the GDM data into SDTM. When the team learnt about Maptrans, it was decided to map the data.

Since the scope of the DSMB package was limited (mostly safety), so was the scope of the mapping itself. The only target was to be able to run the VAD programs and nothing more. Therefore, the team had to create less than ten SDTM domains, and within these domains not all available information from GDM was necessary.

Additionally, the VAD programs were among the first in Roche using SDTM, when SDTMv was not yet available. As a result, the team tried to stick a closely as possible to the final result (the SDTM from the other studies) and not to the Roche standard, which was introduced later.

Fig. 2: Pooling studies for DSMB activities



CASE 2: FINAL ANALYSIS WITH STANDARD REPORTING TOOL

As mentioned earlier, past studies were reported with the safety reporting tool MARS and the efficacy reporting tool Share. For this purpose project teams created VADs based on GDM data and project specific derivation rules. Different sites used different approaches to achieve their goals. Statistical programmers created non-safety analysis datasets, used the efficacy reporting tool to create standard efficacy tables and graphs, and developed programs for non-standard tables, listings and graphs. Specialised Safety Data Analysts used the safety reporting tool to create safety analysis datasets and resulting tables and listings.

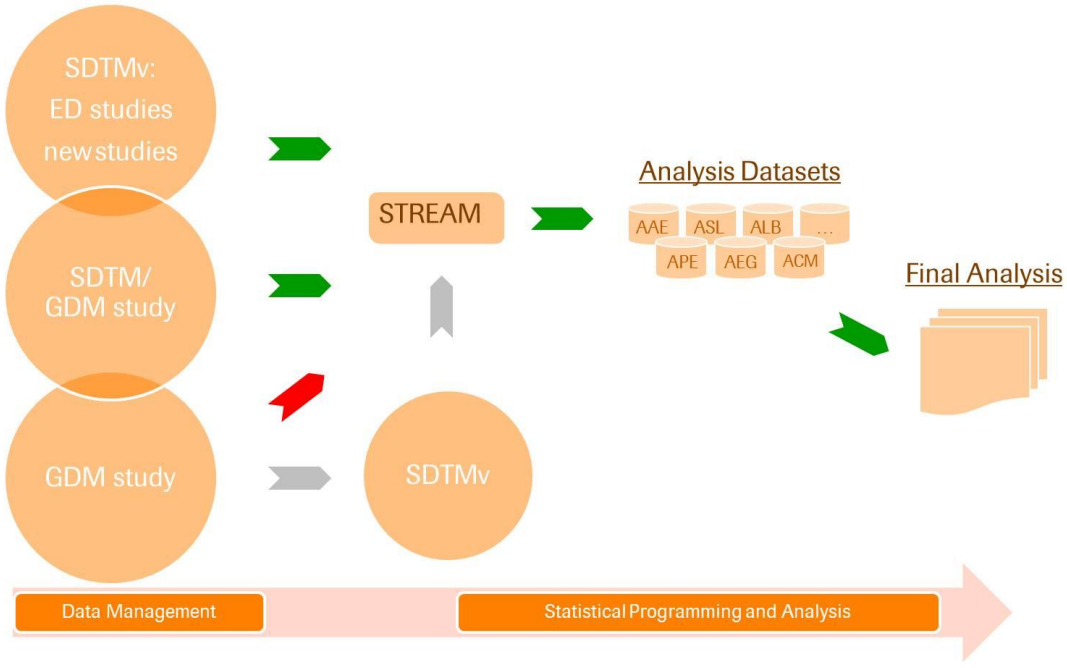
In the example project most finished Phase I and II studies are available in SDTM or SDTMv format, and new studies will be provided in SDTMv format. One ongoing study is available in SDTM format, but is converted to GDM format for DSMB activities. Another study with database lock and analysis in Q3 2012 is set up in GDM format.
For analysis the ongoing SDTM/GDM study could either use legacy reporting tools based on GDM, or the new reporting tool STREAM based on the available SDTM datasets. The GDM study in its original format could only be analysed with legacy reporting tools.

However, the project team decided to convert the GDM study to SDTMv format for two reasons:

- With few exceptions all studies in the project are available in SDTM or SDTMv format. In the future it will probably become necessary to combine the data for pooled analyses.
- Three out of four project programmers were involved in specifications and UAT of STREAM and had gained valuable experience in understanding and using this tool. None of the project programmers had experience with the legacy safety reporting tool. It was estimated that the effort for mapping GDM to SDTMv would be similar to training the old system to all team members.

By mapping the GDM data to SDTMv the project team was able to create VADs and reports with the standard reporting tool in alignment with future analyses in the same project.

Fig. 3: Converting a study for future alignment with other studies



CHALLENGES

IMPLEMENTING THE DATA STANDARDS

Many domains did not map directly, or entirely, from a single source to a single target. One SDTMv domain could be applicable for information from more than one GDM source, and one GDM source could contribute to more than one SDTMv domain. For example, information for subject disposition (DS) was extracted from five GDM datasets and one external file (.csv format). On the other hand the GDM domain containing demographics contributed to several SDTMv domains. In case 2, adverse events of special interest were captured in one standard GDM domain and two additional domains. The standard data could be mapped directly to the SDTMv AE domain. Non-standard information on detailed symptoms was mapped to the clinical events (CE) domain, while information on diagnosis confirmation, possible etiologies and potential triggering events were identified as findings about the adverse events, and thus mapped to the FA domain.

Another challenge was the controlled terminology. As the GDM only provides a standard structure and does not describe standard content, variables in GDM domains usually contain the values as provided in the CRF. For mapping, this meant that values had to be recoded when they deviated from the respective SDTM code lists. In many cases, the GDM values could be recoded directly to SDTM terminology, e.g. values “YES” and “NO” would become “Y” and “N”. In other cases, however, the recoding was not obvious, and pragmatic decisions had to be taken on a case by case basis in conjunction with statisticians and clinical scientists. Additionally, expectations of STREAM had to be met with respect to certain terminology used in standard programs.

The example below shows how reasons for discontinuation were mapped to SDTM controlled terminology.

CRF term	DSTERM	DSDECOD
adverse event or intercurrent illness	Adverse Event/Intercurrent Illness	ADVERSE EVENT
death	Death	DEATH
insufficient therapeutic response	Lack of Efficacy	LACK OF EFFICACY
failure to return	Lost to Follow-up	LOST TO FOLLOW-UP
violation of selection criteria at entry	Violation of Selection Criteria at Entry	PROTOCOL VIOLATION
other protocol violation	Other Violation	PROTOCOL VIOLATION
refused treatment/did not cooperate	Non-compliance	NON-COMPLIANCE
withdrew consent	Withdrawal by Subject	WITHDRAWAL BY SUBJECT
administrative/other	Other	OTHER

Also, the team working on the DSMB case with the Japanese study only had a CRF in Japanese. Only the main annotations were in English. Therefore, to ensure correct mapping the team had to use Google’s translation tools!

Fig. 4: Fitting a banana skin on an apple using a Japanese manual

USING THE MAPPING TOOL

At the time of implementing the mapping tool for the example projects, the tool had some constraints in the support of multiple inputs and repeating variable blocks. In the used version, multiple inputs were handled as one primary and multiple secondary domains, that required unique keys for correct merging. Besides, the functionality of repeating variable blocks for populating multiple parameters within one domain, did not allow for more than nine repeat blocks.

In case 2 these constraints led to problems, that would have required major changes of the mapping tool. For the disposition domain (DS) and the findings about domain (FA), multiple inputs were used to create many parameters. Each source dataset could be used as primary domain, while the secondary datasets could not always be merged to the first one by a unique key. The definition of the metadata file would have been quite complicated to achieve correct merge steps. In the DS domain 12 parameters had to be created, and additional information on screen failures had to be included from a .csv file. Due to time pressure it was decided, not to adapt the mapping tool in a hurry, but to create the DS and FA domains without the mapping tool.

QC APPROACH

The project teams in the described examples followed different validation strategies because of the history of Maptrans and due to case specific details:

- Case 1: The team chose to independently create the .csv files and calling programs twice. The .csv files were compared using a simple diff tool. The resulting SDTM domains were also compared. Finally, to make sure the scope of the mapping was correct, the domains were used to run the VAD programs created for the DSMB package.
- Case 2: The team double programmed from scratch the mapping and then compared the resulting SDTMv domains, because at that time the mapping tool was not fully validated within this context. Additionally, certain expectations of the reporting tool generated issues that had to be resolved in the mapping.

CONCLUSION AND NEXT STEPS

The use of this rule-based approach has been shown to assist the transformation process considerably, providing a much simpler single set of rules, which can be modified individually without the need to revalidate an entire program each time. Real benefit was shown, particularly in the validation phase, where the simplicity of the rule set gave a much better level of transparency and simplicity for a QA programmer.

No standard transformations were available, in both cases virtually all mappings needed to be tackled from scratch. For future studies, project teams will have a good base set of mappings for most SDTM(v) domains. Developing these for a given new study should then involve minimal effort.

Real advantages have already been gained for

- Review, change and validation
- Transparency and ease of re-use
- Defined framework for mapping effort

A reduction of documentation has been achieved, because specifications can be entered directly into .csv files and SAS formats. The specifications and calling programs are self-documenting, so that no other program code needs to be used. The human- and machine-readable .csv files are both specification and mapping documentation at the same time.

RELEASE +1

During the course of the pilot studies, several features were added and enhanced, but other improvements were judged to be better to wait and implement in a subsequent release. These include features to address certain problematic areas, such as preprocessing to check for non-unique keys in secondary domain mappings, as well as ‘bug-fix’ improvements like the ability to include more than 10 repeat variable blocks within a single domain mapping.
It was also recognised that some of the columns in the mapping spreadsheet were redundant and others would be better split, but these changes too were considered better to wait until a second release.

RELEASE +2

The next step in this development will be to replace excel as a means of entering and maintaining the mapping rules, together with the ability to store the rules as part of our central metadata store. This would allow any new study to specify their source and target data models, which would then use existing metadata to compile a set of 'standard' or 'suggested' mapping rules for the transformation programmer to start from. In this way it is expected that the lead time for a new study transformation may be as little as a few hours from first data delivery to initial creation of target data domains.

ACKNOWLEDGMENTS

We would like to thank the SDTM experts group and the Data Standards Office for their valuable support and open discussions.

CONTACT INFORMATION

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

Nicolas Dupuis
F. Hoffmann – La Roche
Bldg. 670/406
CH-4070 Basel, Switzerland
Work Phone: +41 61 68-89164
e-mail: nicolas.dupuis@roche.com

Anja Feuerbacher
F. Hoffmann-La Roche Ltd.
Bldg. 670/R. 406
CH-4070 Basel, Switzerland
Work Phone: +41 61 68-80216
e-mail: anja.feuerbacher@roche.com

Bruce Rogers
F. Hoffmann-La Roche Ltd.
Bldg. 670/R.
CH-4070 Basel, Switzerland
Work Phone: +41 61 68-89145
e-mail: bruce.rogers@roche.com