PharmaSUG 2017 - Paper DS06

Harmonizing CDISC Data Standards across Companies: A Practical Overview with Examples

Keith Shusterman, Chiltern; Prathima Surabhi, AstraZeneca;
Binoy Varghese, Medimmune

ABSTRACT

Whether due to the fact that standardized data are more useful, or that submission of CDISC data is now required, many companies have worked hard to establish local CDISC interpretation guides. Those guides support consistent application of CDISC SDTM and ADaM standards within a company. When companies merge or collaborate, company-specific data standards need to be harmonized to ensure a smooth hand-off of data for the purposes of analysis and reporting. We will share a general overview of the issues encountered in harmonization and present strategies on how to effectively plan and harmonize company-specific CDISC implementation standards. We will walk through a few representative examples from a recent harmonization effort.

INTRODUCTION

Companies spend a lot of time and effort on integrating CDISC standards into their processes, and local interpretations are often closely tied to collection forms, table shells, and other key process steps. Existing standards should be respected as much as possible, as even small changes can have a large downstream effect. As with many things in clinical research, it is often useful to start with the end in mind when harmonizing data standards across companies.

DETERMINING THE SCOPE OF HARMONIZATION

The length of the association between the two companies can impact how the harmonization effort is planned. If the association is short term, then harmonization only needs to cover the areas where the study data may be combined into a single submission. In fact, it may suffice to simply select the company's standards that best suit the project. In a long term association, such as an acquisition or a merger, the goal is to eventually have a single set of standards. Fortunately, not everything needs to be done all at once. Prioritize what is needed and what can be most easily implemented.

A reasonable place to start an inter-organization harmonization effort is with data standards. The assumption is that each company's local standards will usually include complete variable level content, and that content is by and large fully supported by the applicable CDISC models. Sponsor-specific additions to CDISC controlled terminology, however, will need to be closely monitored to ensure a single comprehensive set of values. The ADaM standard is much more flexible at the variable level. ADaM variables are broadly specified, but it is very common to use custom variables that follow ADaM conventions. This is where the bulk of variable level harmonization needs to occur. It is possible for both companies to handle items like population flags, grouping variables, and sequence variables in a very different manner while still complying with CDISC standards. The end goal is to use the ADaM datasets to generate tables, listings, and figures. The shells for those outputs can serve as a very useful guide for determining the purpose of the custom ADaM variables across both organizations and how they should be harmonized.

After variable level standards are harmonized, the next step is typically to harmonize value level standards. This is more often done at the SDTM level, as the ADaM datasets are largely driven by the SDTM input. This often includes xxTESTCD and xxTEST in findings domains and QNAM and QLABEL values in supplemental qualifier domains. At the value level, SDTM content is often naturally consistent due to extensive CDISC controlled terminology. If the same information is collected by both companies, then the goal is to have the same information represented the same way in the same SDTM domain. However, differences in case report forms can have a very strong impact on the SDTM datasets. It may be necessary to transform case report form values into CDISC controlled terminology. Additionally, there are many opportunities for values without controlled terminology to have similar but not exactly matching values in each company's standards. For example, certain key details may need to be standardized within a submission, such as ARM and ARMCD in the DM domain and pharmacological class in the TS domain. However, this level of harmonization is typically at the therapeutic area level as opposed to the corporate standards level. For example, a company may have naming conventions for ARM and ARMCD, but they are unlikely to specify arms by study at the corporate level.

HARMONIZATION OF SDTM DATA

In order to have a single, unified standard, variable-level metadata must be established up front. If variables with different names contain the same information, one name must be chosen. If there is variation in the variable labels, a single label must be chosen as well. The same goes for data type (numeric, character, etc.), permissibility, and any formats that are applied. Metadata harmonization is usually not needed for SDTM standards, as all variable-level information in SDTM is strictly controlled by CDISC. However, one company's standards may include SDTM permissible variables that another may not. For example, the SDTM variable VISITDY, which captures the study day of a visit, is not required under any circumstances. One company may include VISITDY in its SDTM domains for analysis or traceability purposes, where the other company may prefer to derive the study day of the visit in the ADaM datasets, or they may deem the visit study day unnecessary to capture altogether. In that situation, the two companies should discuss the purpose of the permissible variable and come to a decision on whether the value of including the variable is worth the extra effort involved.

While the SDTM model is very strict in regards to variable-level metadata, there could be complications that require harmonization if the companies' standards are modeled to different versions of SDTM. For example, domains that were not present in SDTM 3.1.3 were added to SDTM 3.2. One of these domains is Subject Status (SS). It is entirely possible that the company using SDTM 3.1.3 has studies where subject status is captured, but their standards may not match the CDISC standards outlined in SDTM 3.2. The company using SDTM 3.1.3 may have the information captured in a custom domain, or even in an events domain like Disposition (DS). The solution here depends on which SDTM version is being used for the current project. The company using SDTM 3.1.3 may want to adopt the other company's standards if they will be using STDM 3.2 as part of the joint effort. If SDTM 3.1.3 will continue to be used, then any number of compromises can be made in such a situation.

HARMONIZATION OF ADaM DATA

Variable level metadata harmonization is much more critical in ADaM datasets, where the CDISC model is open to general variable conventions as opposed to set variable names and labels in certain situations. Occurrence flag variables are an example where the ADaM model has the generic variable AOCCzzFL, with label "1st Occurrence of…" that can be filled in any way the sponsor chooses as long as the total label is forty characters or less. If both companies use these custom occurrence flags, then it is almost guaranteed that their labels will not match, and a standard sequence of these AOCCzzFL variables will need to be agreed upon. The table shells may provide context here on what occurrence flag variables are needed.

Like SDTM, ADaM has the issue of determining which permissible variables should be included. This is actually an even greater challenge in ADaM than SDTM, because most SDTM variables are generally considered to be permissible variables in ADaM datasets. The ADaM model allows for an analysis dataset to derive values based on the SDTM collected results without having the SDTM result variable present in the analysis dataset, but some companies prefer to include those SDTM variables in their analysis datasets for traceability purposes. A complete standard would include conventions for which SDTM variables to include.

HARMONIZATION OF TERMINOLOGY

The CDISC standards come with a wealth of controlled terminology in order to ensure consistency of value-level information. While both companies are ideally basing their own standards in CDISC, there are many reasons why two companies' standards may not match in terminology. First, while CDISC controlled terminology is vast, it is not all-encompassing. Many companies use questionnaires not yet included in CDISC's QS terminology list. Many studies have obscure lab tests that may not be included in the CDISC LBTEST codelist. It is possible for both companies to collect the same information that slips through the cracks in controlled terminology in different ways. This is often driven by the case report forms, so the case report forms for the studies within the scope of the collaboration can assist in harmonizing terminology in these cases. Additionally, CDISC controlled terminology is updated frequently. The end goal is to have a single controlled terminology version specified, and to have that terminology version guide the terminology harmonization.

If tabulation data is within the scope of harmonization, then most terminology harmonization would occur at the SDTM level. ADaM-specific controlled terminology is extremely limited, and any value-level harmonization that occurs in the SDTM domains will carry over into the ADaM datasets. However, certain ADaM variables may still require harmonization, such as PARAMCD and PARAM. While SDTM findings domains contain variables that capture the units of measurement, the ADaM BDS structure does not. As a result, a common method for defining PARAM is to concatenate the associated SDTM xxTEST value with the xxSTRESU value. If the xxTEST and xxSTRESU values were harmonized in the SDTM, then this simple concatenation will often produce unique parameter names. This just one approach, but having both companies agree on such a convention will greatly simplify the task of harmonizing parameters.

If only the ADaM model is within the scope of harmonization, then all terminology harmonization will occur in the ADaM datasets. The process for terminology harmonization in analysis datasets is very similar to the SDTM process, but the case report forms no longer need to be considered. Instead, additional SDTM variables will likely be included in the ADaM datasets for traceability. If more than one study has an SDTM variable with varying terminology, then the SDTM variable can be included in the ADaM dataset, and a new analysis version of the variable with harmonized terminology can be derived.

PROCESS OVERVIEW

.

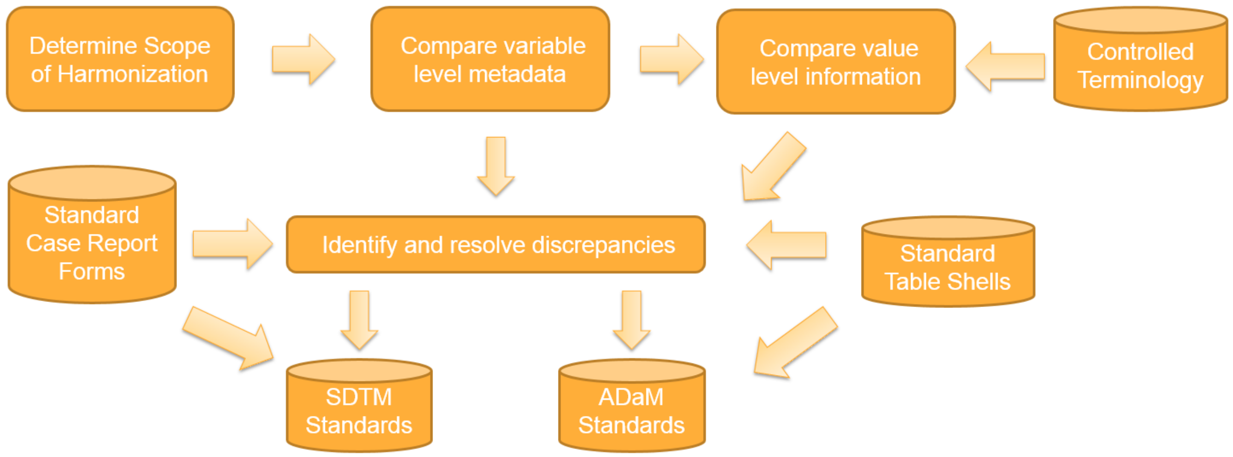

Determine Scope of Harmonization
Compare variable level metadata
Compare value level information
Controlled Terminology
Standard Case Report Forms
Identify and resolve discrepancies
Standard Table Shells
SDTM Standards
ADaM Standards

Figure 1: Illustration of the harmonization process

IMPLEMENTATION OF DATA STANDARDS HARMONIZATION

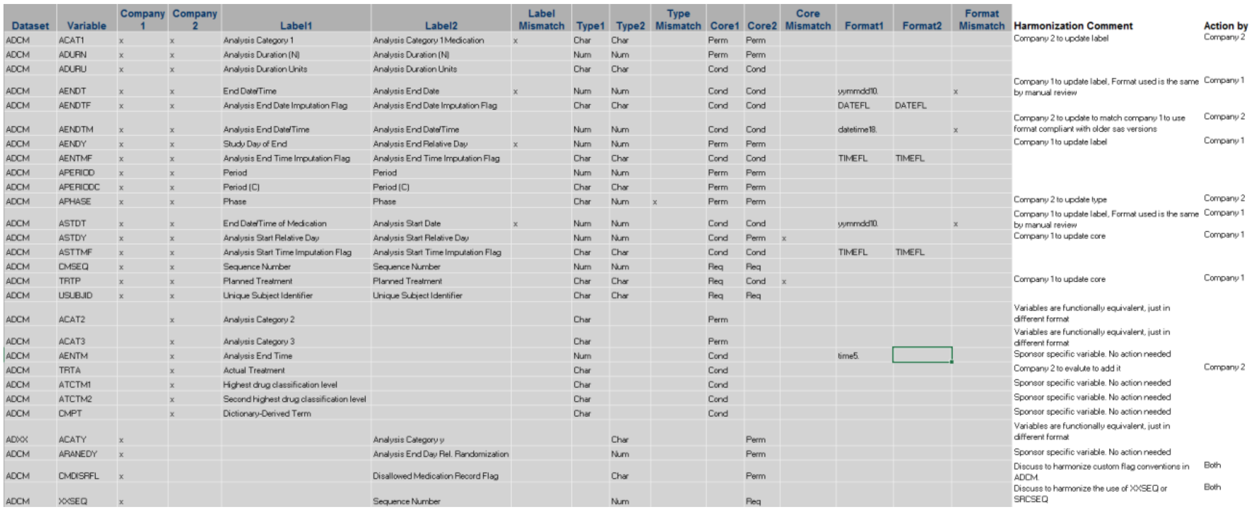
Once a road map is prepared and scope determined, the next steps are planning and implementation of harmonization. Identifying common information between these standards will serve as a good starting point. This assumes that both companies will develop their flavor of implementation guides in compliance with guidelines set by CDISC. In some cases, metadata composition and version of CDISC guidelines adopted during development may differ. Such a difference will introduce an additional step to preprocess the metadata to achieve a certain degree of consistency between both standards to facilitate comparison. The best case scenario is when both standards are platform independent and identical in structure.

Companies may elect to store standards metadata in various formats like excel files, custom metadata repositories or proprietary analysis and reporting systems. The development approach may also differ influencing the final outcome. For example, some companies take source to target SDTM approach while others create target SDTM metadata and then devise a mapping process to fit in the raw data. Supplemental variables may be a separate domain or part of the parent domain. All CDISC permissible variables may be included or limited to the ones frequently used. Extensible controlled terminology may also exhibit such differences based on the development methodology.

Below is a high-level outline of various steps involved in the harmonization process.

- Extract all the metadata components in a machine-readable format. For example, excel files can be converted into SAS datasets that allows for comparison through programming.
- Compare variable level metadata like name, label, role, core, type, derivation, format and codelist for common variables. Flag records to indicate the differences.
- Harmonize value-level information if in scope.
- Save this report in a user friendly portable format such as excel.
- Manually review the report to document reasons for differences. This especially applies to sponsor defined ADaM variables and supplemental SDTM domains.

Figure below shows an example output from ADCM comparison. The gray columns are auto generated based on programmatic comparison. Other columns are compiled during manual review to document findings and decisions.



Dataset	Variable	Company 1	Company 2	Label1	Label2	Label Mismatch	Type1	Type2	Type Mismatch	Core1	Core2	Core Mismatch	Format1	Format2	Format Mismatch	Harmonization Comment	Action by
ADCM	ACAT1	x	x	Analysis Category 1	Analysis Category 1 Medication	x	Char	Char		Perm	Perm					Company 2 to update label	Company 2
ADCM	ADURN	x	x	Analysis Duration (N)	Analysis Duration (N)		Num	Num		Perm	Perm						
ADCM	ADURU	x	x	Analysis Duration Units	Analysis Duration Units		Char	Char		Cond	Cond						
ADCM	AENDT	x	x	End Date/Time	Analysis End Date	x	Num	Num		Cond	Cond		yymmdd10.		x	Company 1 to update label, Format used is the same by manual review	Company 1
ADCM	AENDTF	x	x	Analysis End Date Imputation Flag	Analysis End Date Imputation Flag		Char	Char		Cond	Cond		DATEFL	DATEFL			
ADCM	AENDTM	x	x	Analysis End Date/Time	Analysis End Date/Time		Num	Num		Cond	Cond		datetime18.		x	Company 2 to update to match company 1 to use format compliant with older sas versions	Company 2
ADCM	AENDY	x	x	Study Day of End	Analysis End Relative Day	x	Num	Num		Perm	Perm					Company 1 to update label	Company 1
ADCM	AENTMF	x	x	Analysis End Time Imputation Flag	Analysis End Time Imputation Flag		Char	Char		Cond	Cond		TIMEFL	TIMEFL			
ADCM	APERIOD	x	x	Period	Period		Num	Num		Perm	Perm						
ADCM	APERIODC	x	x	Period (C)	Period (C)		Char	Char		Perm	Perm						
ADCM	APHASE	x	x	Phase	Phase		Char	Num	x	Perm	Perm					Company 2 to update type	Company 2
ADCM	ASTDT	x	x	End Date/Time of Medication	Analysis Start Date	x	Num	Num		Cond	Cond		yymmdd10.		x	Company 1 to update label, Format used is the same by manual review	Company 1
ADCM	ASTDY	x	x	Analysis Start Relative Day	Analysis Start Relative Day		Num	Num		Cond	Perm	x				Company 1 to update core	Company 1
ADCM	ASTTMF	x	x	Analysis Start Time Imputation Flag	Analysis Start Time Imputation Flag		Char	Char		Cond	Cond		TIMEFL	TIMEFL			
ADCM	CMSEQ	x	x	Sequence Number	Sequence Number		Num	Num		Req	Req						
ADCM	TRTP	x	x	Planned Treatment	Planned Treatment		Char	Char		Req	Cond	x				Company 1 to update core	Company 1
ADCM	USUBJID	x	x	Unique Subject Identifier	Unique Subject Identifier		Char	Char		Req	Req						
ADCM	ACAT2		x	Analysis Category 2			Char			Perm						Variables are functionally equivalent, just in different format	
ADCM	ACAT3		x	Analysis Category 3			Char			Perm						Variables are functionally equivalent, just in different format	
ADCM	AENTM		x	Analysis End Time			Num			Cond			time5.			Sponsor specific variable. No action needed	
ADCM	TRTA		x	Actual Treatment			Char			Cond						Company 2 to evalute to add it	Company 2
ADCM	ATCTM1		x	Highest drug classification level			Char			Cond						Sponsor specific variable. No action needed	
ADCM	ATCTM2		x	Second highest drug classification level			Char			Cond						Sponsor specific variable. No action needed	
ADCM	CMPT		x	Dictionary-Derived Term			Char			Cond						Sponsor specific variable. No action needed	
ADXX	ACATY	x			Analysis Category y			Char			Perm					Variables are functionally equivalent, just in different format	
ADCM	ARANEDY	x			Analysis End Day Rel. Randomization			Num			Perm					Sponsor specific variable. No action needed	
ADCM	CMDISRFL	x			Disallowed Medication Record Flag			Char			Perm					Discuss to harmonize custom flag conventions in ADCM.	Both
ADCM	XXSEQ	x			Sequence Number			Num			Req					Discuss to harmonize the use of XXSEQ or SRCSEQ	Both

Figure 2: Programmatic comparison report for documenting findings and decisions

EXAMPLES OF DIFFERENCES

Below are examples of differences identified during the harmonization process.

Units of baseline measurements in ADSL: Certain measurements, such as height, weight, BMI collected at baseline are stored in ADSL. One standard has separate variables in ADSL for the value and the unit (for example, BMIBL for BMI at baseline and BMIU for the unit). The other standard has only one variable including the unit in the label (for example, BMIBL with label Baseline Body Mass Index (kg/m**2).

Character variables with numeric versions: One standard has all ADaM IG specified numeric versions of character variables, whereas the other standard only includes them when needed for sorting.

Generic variables and descriptions: One standard has generic variable and a description variable to give the context for the generic variable. For example, ADSL has the generic grouping variables GROUPxx and GROUPxxD, with GROUPxx to group subjects and GROUPxxD to describe what the group is. The other standard described the group in the variable name and label, similar to the ADaM variables AGEGR1 and RACEGR1 instead of creating another description variable.

Population and non-population flags: Both standards have sponsor-defined population and non-population flags in ADSL. There is a need for a harmonized naming convention for handling custom population flags and non-population flags in ADSL. Example, all population flags have “POP” in the variable name.

SEQ variables: One standard has the SDTM xxSEQ variable when only one SDTM domain is an input to an ADaM dataset, and has the variable SRCSEQ if multiple SDTM domains are an input to an ADaM dataset. Other standard has a simplified approach to use only SRCSEQ to capture SDTM sequence values regardless of the number of input datasets.

ADJUDICATION OF DIFFERENCES

The output from the comparison of metadata gives a clear understand of the similarities and differences between both company standards. Review teams specializing in a specific standard can hold meetings to facilitate decision-making and document the next steps. In most cases, the case report forms will guide resolutions for differences in the SDTM model. In the ADaM model, the table shells will help determine how the conflicting standards are resolved. Companies should try to harmonize the standards wherever possible. However, there will be instances where differences cannot be resolved. For example, findings arising from difference in study design, data collection or analysis methodology cannot be harmonized. Differences in data collection may lead to differences in controlled terminology. Some companies may choose not to harmonize QNAM and QLABEL values to have traceability to naming used in data collection. ADaM standards usually have more than one correct way of implementation and that can lead to more differences as against SDTM. Company specific derived variables in ADaM datasets can also lead to differences.

Each company publishes their standard to their study teams after all outstanding differences have been resolved/documented by the adjudication committee. A joint harmonization governance team with representation from both companies facilitates the adoption of harmonized standards by ongoing and upcoming studies.

ADOPTION OF HARMONIZED STANDARDS

Project teams can refer to the published implementation guides while working on their studies. Their familiarity with harmonized standards lead to easy adoption especially as studies change hands from one company to another. This, in effect, results in big time savings as well as increased accuracy when creating submission packages and data supporting integrated summaries. Project teams can document additional findings during implementation pertaining to both data structure and controlled terminology.

LESSONS LEARNED

The standards team should organize lessons learned sessions with individual study teams at study closeout to capture the findings from implementation. This can be used to understand differences arising at the execution stage which may not be apparent during standards development. Every subsequent iteration of standards development can incorporate this information to achieve a higher degree of harmonization within study level data.

CONCLUSION

Companies should use existing standards amongst them as a starting point in developing new standards. Since each company has already gone through the process of implementing CDISC standards in a way that fits their needs, these existing standards are an excellent starting point in developing a new, harmonized standard. That will increase efficiency in development and subsequent harmonization. Companies can implement the same programmatic approach to re-harmonize standards that change with time and evaluate any new differences and ensure that previous differences were resolved. A joint governance process can also be defined to reduce ongoing harmonization activities. Harmonized standards can be stored in a central metadata repository accessible to both companies. Adoption of harmonized standards will largely eliminate variation in study data and enable efficient pooling and faster reporting of clinical trial data to regulatory agencies.

ACKNOWLEDGMENTS

We would like to thank Steve Kirby for his review and feedback.

CONTACT INFORMATION

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

Keith Shusterman
Chiltern
Keith.Shusterman@chiltern.com

Prathima Surabhi
AstraZeneca
Prathima.Surahbi@astrazeneca.com

Binoy Varghese
MedImmune
vargheseb@medimmune.com