PhUSE 2016

Paper CD12

Pattern based Metadata Repository: toward high quality data standards

Alan Cantrell, PAREXEL®, Sheffield, United Kingdom
Julius Kusserow, PAREXEL®, Berlin, Germany
Julie James PAREXEL® Informatics, Nottingham, United Kingdom
Deb Copeland, PAREXEL®, Durham, North Carolina, USA
Natraj Patro, PAREXEL®, Billerica, Massachusetts, USA
Isabelle de Zegher, PAREXEL® Informatics, Wavre, Belgium

ABSTRACT

In most organizations, data standards are maintained in silos: data collection standards, SDTM and ADaM are managed by different groups. Mapping between these standards remains a challenge across disjointed governance processes.

Within PAREXEL, we followed a different approach based on three principles

- End-to-End approach, from protocol to submission
- CDASH, SDTM and ADaM are different views of the same object
- Mapping obey standards patterns, related to data types (ISO21090)

We implemented these principles in PAREXEL's Clinical MDR

- Domains are defined through hub & spoke; hubs contain generic variables while each spoke contain standard specific variables
- Mapping through the hub & spoke is defined by re-usable patterns

This approach brings several benefits

- Centralized management of standards
- Automated and consistent mapping with data lineage
- Automatic generation of SDTM and ADaM for each study, following specification of data collection

INTRODUCTION

Many companies have deployed – or are in the process of deploying – a metadata repository (MDR) to manage their data standards, including data collection (CDASH), data submission (SDTM) and derived data (ADaM) standards.

In most organisations, data standards are maintained in silos: data collection /CDASH standards are maintained within Data Management, SDTM is governed by the clinical programmers and ADaM is managed by the statistical programmers. These different groups collaborate to maintain proper mapping between the different standards, but this remains a challenging process across separate groups with disjointed governance processes. In addition, mapping between the standards remains an “art”, based on manual interpretation and experience from the programmers. In this paper, we explain how we implemented a different approach, based on end-to-end data standards, in order to increase efficiency and provide a complete data lineage as required by FDA.

While the MDR brings value in increasing quality and consistency of data standards, it is truly effective when data standards are used in the context of a study. And this is the second challenge faced by organizations deploying an MDR: there is indeed limited (or no) integration with the downstream functions & systems that need to consume data standards for a specific study. In a separate paper, presented at the PhUSE 2016 conference, we focus on this second challenge: i.e. how to use PAREXEL's Clinical MDR to support the transformation of (part of) the protocol into a machine readable format that can be utilized by all the downstream systems used in trial execution.

WALKING THE LINE

So often we work in a linear way, waiting for the process before us to be completed before we can move forward – e.g. we need to build a database before we can enter data; that data needs to be entered before we can produce tabulation data (SDTM); and the tabulation data needs to be available before the analysis datasets are created.

What could happen if we didn't need to wait? How can we change our process to avoid waiting?

We need an end-to-end (E2E) approach to standardize the information, so that when a data collection form is designed we can know how it will impact the tabulation and analysis data. If we can know the structure of the data collection forms, and how that is connected to the structure of tabulation data, and consequently the structure of analysis datasets, we can reduce the waiting time.

INTRODUCING PAREXEL CONCEPTS

To explore this approach at PAREXEL, we looked at a specific module of data: adverse events (AE). We considered that a module is composed of 3 levels:

1. Data collection (from for EDC and other data collection instruments)
2. Tabulation data (some from data collection, some derived, some assigned)
3. Analysis data

We noted during this review process that certain aspects overlapped: AESEV (CDASH); AESEV (SDTM) and AESEV, AESEVN, ASEV and ASEVN (ADaM)

So if an element exists in 2 more places, it is (or should be) the same "thing". After identifying the different unique "things" present in the AE module, we created a grouping of them in a "concept".

At PAREXEL, we view a concept as the group of all the "things" or unique variables identified across the three standards, defined in a standards agnostic way within a Domain Definition, and linked with the standards-specific variables in different Spokes. We explain throughout this paper how we reached this definition and how we are using it in data operations.

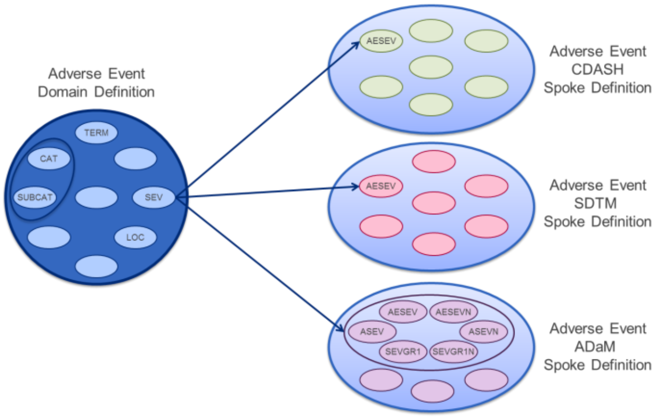


Figure 1. A concept includes the Domain definition and different spokes

RELATIONSHIP PROBLEMS

Once we had a concept and a list of variables contained within it, we then checked how these variables should be used. Should they occur in SDTM, in ADaM, or in data collection uniquely or all together? If they appear in data collection, are they mandatory, e.g. do they have to always be present in the definition of a data collection form?

We also noted that we still had variables in the concept that were related to each other – for example: the AESEVN variable in ADaM requires the AESEV variable within the tabulation data. So in a situation where the data manager decides to not collect severity in the CRF, we would not want the analysis severity variables appearing in the ADaM dataset. Without a concept approach, we would not be able to identify this before the ADaM programmer realizes that he does not have the needed data to compute AESEVN. However, something was needed to counsel the variables and render explicit these implicit relationships.

EXPLORING THE RELATIONSHIPS WITHIN A CONCEPT

As we explained in the section above, some variables are grouped together and should be managed as a whole. Instead of having all these variables in the concept, we could have a single variable which represents each grouping. We reduced the list of variables to these single elements, called “Domain Definition variables”.

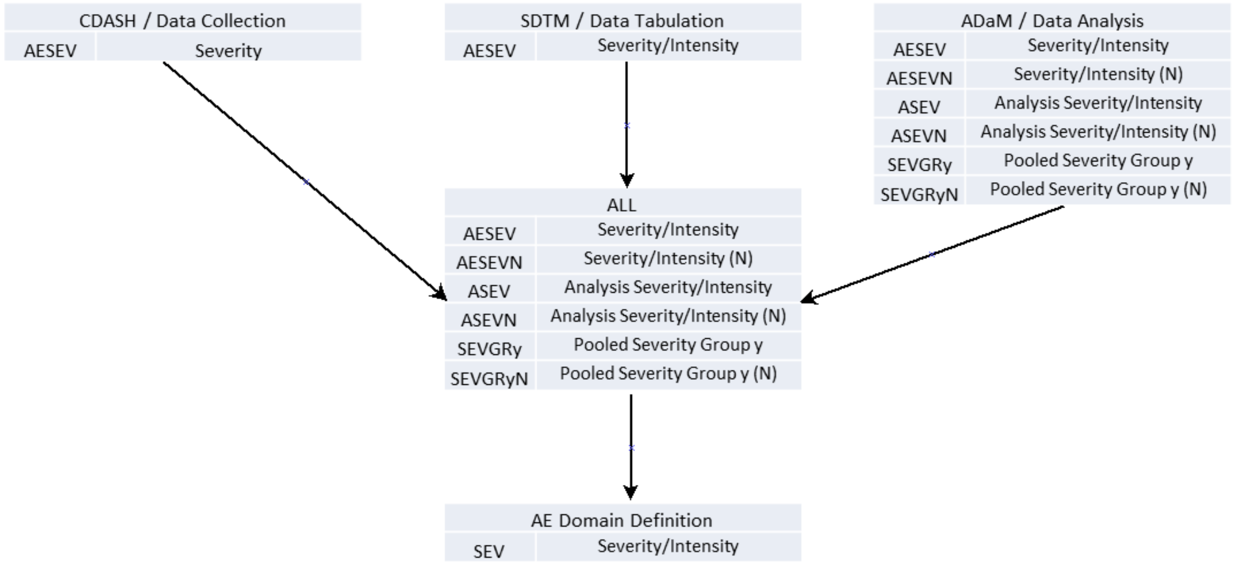


Figure 2. A Domain Definition variable represents the meaning of variables across the different CDISC standards

This greatly reduced the number of variables required to represent the collection, tabulation and analysis. But how do we populate ASEVN in AE ADaM and AESEV in the AE SDTM domain from this?

We need a way of moving back and forth between these elements. Currently, most mapping is done in a linear fashion (red arrows) between Data Collection to Data Tabulation, and then between Data Tabulation to Data Analysis. This method requires the prior steps in the process to be completed before the programmer can be moving to the next.

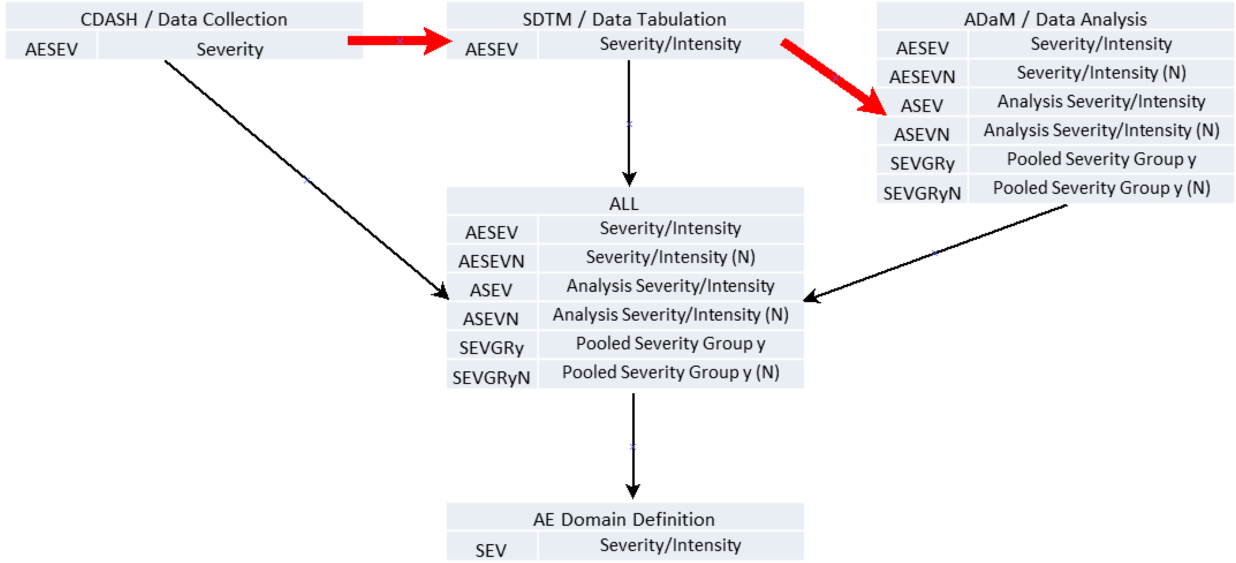
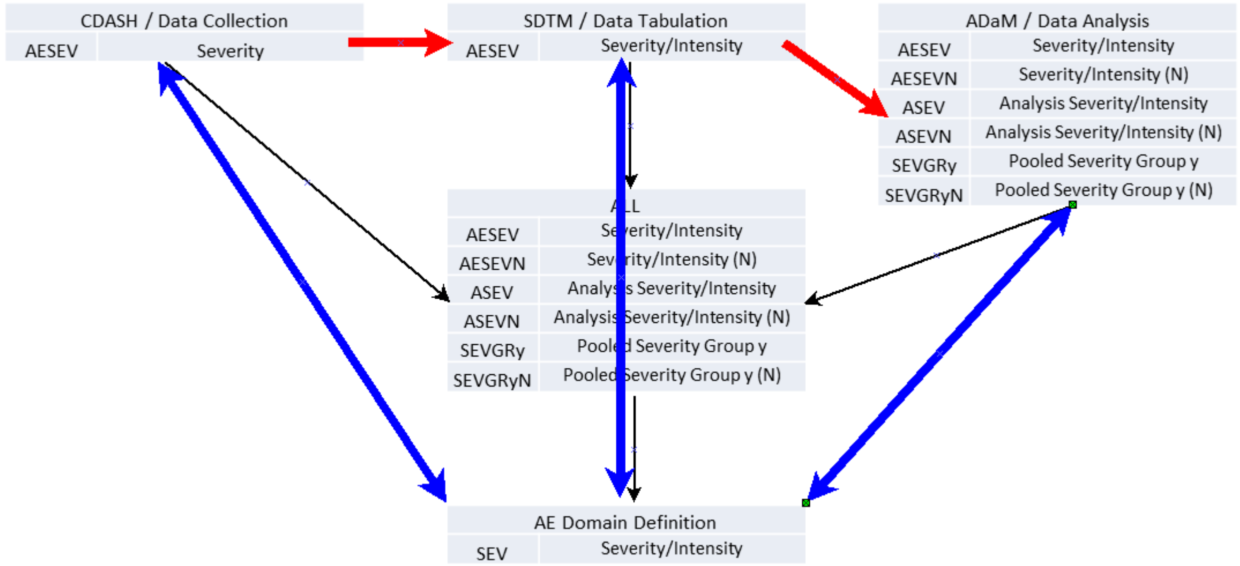


Figure 3. Linear mapping approach (red arrow)

If we move to an approach where the data collection, data tabulation and data analysis variables are all connected to a unique domain definition variable, this method does not require prior steps in the process to be completed before moving to the next. Once we have defined these domain definition variables, we are able to understand the information in the context of data collection, tabulation or analysis.



CDASH / Data Collection	
AESEV	Severity

SDTM / Data Tabulation	
AESEV	Severity/Intensity

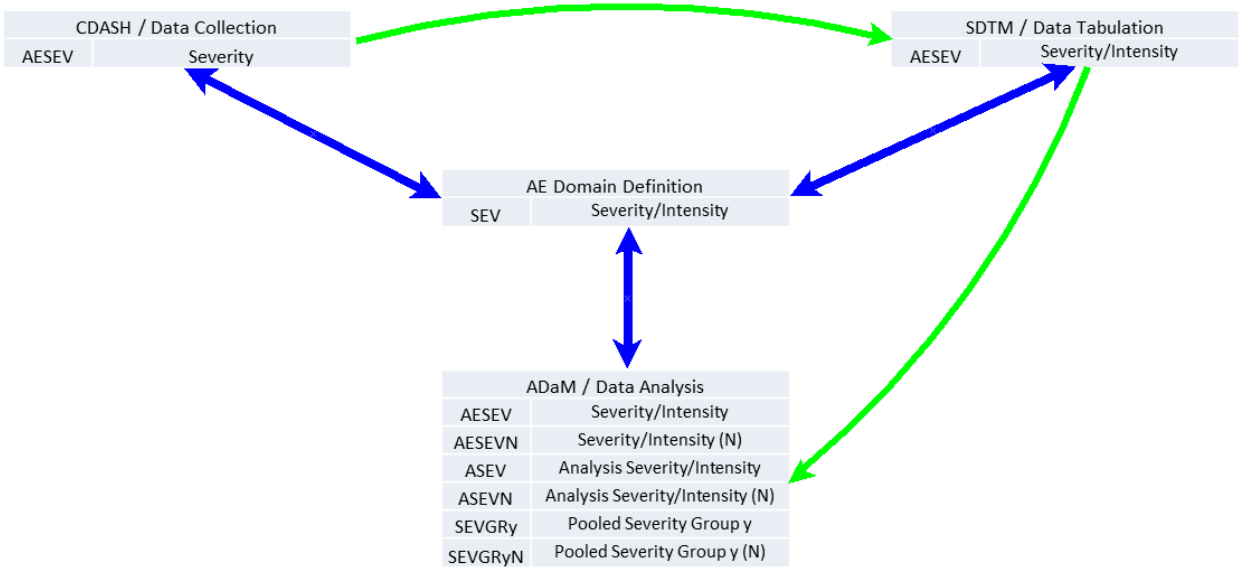
ADaM / Data Analysis	
AESEV	Severity/Intensity
AESEVN	Severity/Intensity (N)
ASEV	Analysis Severity/Intensity
ASEVN	Analysis Severity/Intensity (N)
SEVGRy	Pooled Severity Group y
SEVGRyN	Pooled Severity Group y (N)

ALL	
AESEV	Severity/Intensity
AESEVN	Severity/Intensity (N)
ASEV	Analysis Severity/Intensity
ASEVN	Analysis Severity/Intensity (N)
SEVGRy	Pooled Severity Group y
SEVGRyN	Pooled Severity Group y (N)

AE Domain Definition	
SEV	Severity/Intensity

Figure 4. Connecting the Domain Definition variable with the different standards

By using the connections between the domain definition variable and the data collection, tabulation and analysis variables, we can recreate the current linear approach to mapping.



CDASH / Data Collection	
AESEV	Severity

SDTM / Data Tabulation	
AESEV	Severity/Intensity

AE Domain Definition	
SEV	Severity/Intensity

ADaM / Data Analysis	
AESEV	Severity/Intensity
AESEVN	Severity/Intensity (N)
ASEV	Analysis Severity/Intensity
ASEVN	Analysis Severity/Intensity (N)
SEVGRy	Pooled Severity Group y
SEVGRyN	Pooled Severity Group y (N)

Figure 5. The Domain Definition variable at the center, with the different standards – considered as spoke

So, how do we expand upon this? The above example is just one part of adverse events – we need to collect the remaining elements for adverse events, as well as all the other elements for demography, disposition, medical history, exposure, etc...

INTRODUCING PATTERNS

We have seen above the creation of a domain definition variable called SEV. The variable is not just a character string datatype or numeric datatype; it has to have the capability of representing multiple datatypes related to the different variables linked to this domain definition variable.
For instance:

- The data collection and data tabulation element AESEV is governed by controlled terminology
- The data analysis element AESEVN is a numeric representation of the AESEV variable
- In addition, ASEV and ASEVN is an imputed (e.g. in case of missing severity values in the data collection) variable pair representing AESEV
- SEVGRy and SEVGRyN can represent a pooled grouping of AE severity in the analysis based on ASEV values

Based on this, it is clear that specifying a CD datatype (Coded Descriptor in ISO20190) to the SEV domain definition variable does not allow full definition of the multiple datatypes needed. We require more details on the controlled terminology for the variable.

- The controlled terminology for AESEV is MILD, MODERATE and SEVERE from the code list C66769 (Severity/Intensity Scale for Adverse Events)
- The data collection and tabulation components may take value from this terminology
- The numeric equivalent value would use values of 1, 2 and 3 to represent these text values
- The analysis values need populating also, e.g. the values of proper cased and if missing set to "Severe"
- We also need to define the groupings for the analysis, e.g. values of mild / moderate are in group 1 and values of severe are in group 2

Here is an example of these data components:

Collection / Tabulation	Analysis	Ordinal	Analysis Group	Codelist	Codelist Code	NCI Preferred Term
MILD	Mild	1	Mild/Moderate - 1	C66769	C41338	Mild Adverse Event
MODERATE	Moderate	2	Mild/Moderate - 1	C66769	C41339	Moderate Adverse Event
SEVERE	Severe	3	Severe - 2	C66769	C41340	Severe Adverse Event
<null>	Severe	3	Severe - 2			

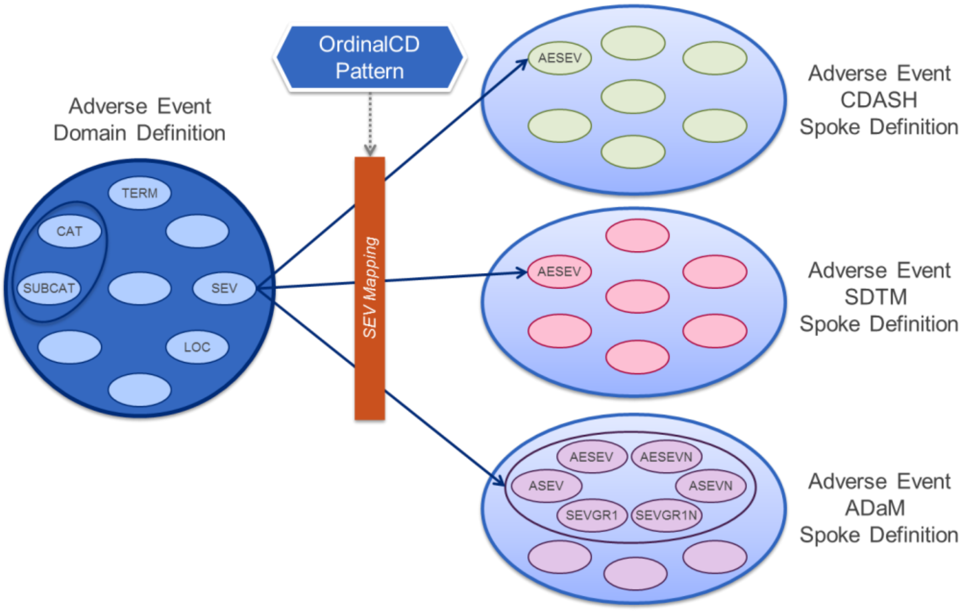
These coded values are linked together in a logical way. This logic can be defined building on the code list definition and some SAS logic for imputing the analysis value and defining the grouping. An example of a value "MODERATE" selected in the collection data could then produce the following:

.codelist().displayName()	MODERATE	AESEV
.codelist().value()	2	AESEVN
.codelist().imputed().displayName()	Moderate	ASEV
.codelist().imputed().value()	2	ASEVN
.codelist().group().displayName()	Mild/Moderate	SEVGR1
.codelist().group().value()	1	SEVGR1N

We can identify all of the values for collection, tabulation and analysis through the logic defined in the table above that we call "complex datatype" as they enable representing multiple data types from the different values. When this complex datatype model is connected to the domain definition variable 'SEV' above we obtain a single variable with a complex datatype to represent the collection, tabulation and analysis for the data.

This complex datatype can also be reused – it is not unique to "Severity" but it can be used in the aspects of collection, tabulation and analysis for "Causality" and "Toxicity Grade" in AE and in many other variables across domains. So we call this complex datatype a pattern and we provide a pattern for each domain definition variable. Each pattern has a name – in the example mentioned above we call it OrdinalCD.

In this instance, we can now construct a *domain definition* called **Adverse Event** with a *domain definition variable* **SEV** which has a connected *pattern* **OrdinalCD**. Likewise, we can also add domain definition variables of **REL** and **TOXGR** to this domain definition with the same pattern.


OrdinalCD Pattern
Adverse Event Domain Definition
TERM
CAT
SUBCAT
SEV
LOC
SEV Mapping
AESEV
Adverse Event CDASH Spoke Definition
AESEV
Adverse Event SDTM Spoke Definition
AESEV
AESEVN
ASEV
ASEVN
SEVGR1
SEVGR1N
Adverse Event ADaM Spoke Definition

USING THE MODEL

We have seen above an example of a pattern **OrdinalCD** that can be applied to a domain definition variable. We have identified so far 20 patterns that cover the majority of CDISC standards. Now consider a domain definition which contains multiple domain definition variables – each of these variables is connected to a pattern which allows the definition of complex datatype elements linked to data collection, tabulation and analysis elements. Developing a collection of domain definitions related to the CDISC domains enables a standard model which is central to the collection, tabulation and analysis models. We call this the “Hub and Spoke” model.

We are in the process of implementing all CDISC domains within PAREXEL's Clinical MDR following the pattern approach. At the time of writing this paper, we have defined 10+ domains with this approach.

The main lessons learned from the work performed to date have been:

- The ability to have a flexible adaptive approach to defining the patterns. What a pattern is named is not important – the structure is.
- To produce visual examples of the patterns when discussing with the team – the theory of the pattern approach can be difficult to follow without these.
- To test the patterns – when a pattern is defined, does it work? What challenges are faced with the pattern defined?
- To be agile in development – what was originally a linear mapping approach has developed into a combined list of variables and finally into domain definitions with patterns. There is always room for improvement in standards, and testing new ideas promotes this.

As described in the table below, there are benefits to this approach versus the classical linear approach when managing data standards. In addition, we believe that the re-usable patterns can add tremendous value in the generation of high quality Therapeutic Areas standards: the C-Maps being generated by the scientific experts can be used as the basis of the domain definition and the patterns can be used to generate the corresponding CDASH, SDTM and ADaM variables.

	Linear approach	“Hub & spoke” approach with patterns
Description	• Data Standards (CDASH, SDTM, ADaM) maintained independently • Mapping between data standards is managed manually (one by one)	• Data standards are maintained through concepts, composed by a domain definition with the domain variables (hub), linked with the representation of the variables in the different CDISC standards (spoke) • Linkage between hub and spoke is done through re-usable patterns
Benefits	• Flexibility, supporting independence across different data standards groups within an organization	• Consistency and quality in maintaining data standards – approved CDISC ones and emerging TA standards • Efficiency in managing mapping across different standards and different versions of standards • Specification of SDTM and ADaM are generated

	Linear approach	"Hub & spoke" approach with patterns
		automatically when specifying data collection forms, ensuring that everything collected is used and no missing data to compute SDTM or ADaM variables • Efficiency in generating data lineage
Concerns	• Stepped approach in specifications and in execution • Risk of inconsistency in mapping for same type of variables across domains • Risk of inconsistency in mapping across studies • Workload to generate high quality mapping for studies • Workload to generate E2E data lineage	• Requires definition of the different patterns • Requires definition of the "Hub" concept and mapping patterns • Process change within the organization • Requires MDR tool that support management of patterns and "hub & spoke" approach

CONCLUSION

The PAREXEL data standards are being developed with a wide team of experts from data management, clinical programming, SDTM programming and ADaM programming groups. We started with the specification and the development of the Clinical MDR, in collaboration with Sycamore Informatics, to facilitate efficient development of data collection, tabulation and analysis elements. Patterns were developed to facilitate management of standards while maintaining traceability between collection and analysis, and ensure high quality of data standards and therefore underlying data.

We are in the process of deploying PAREXEL's Clinical MDR, and will begin use in sponsor studies by end of 2016.

REFERENCES

[1]. Effective use of a Metadata Repository across data operations: the need for a machine readable form of (part of) the protocol. PhUSE 2016.
Isabelle de Zegher, Michele Gray, Mark Sullivan, Michael Goedde.

[2]. E2E data standards, the need for a new generation of metadata repositories. PhUSE 2015.
Isabelle de Zegher, Alan Cantrell, Julie James.
www.phusewiki.org/docs/Conference 2015 DH Papers/DH04.pdf

ACKNOWLEDGMENTS

The work presented in this paper is part of a wider initiative within PAREXEL around E2E Data Standards management, including the PAREXEL's Clinical MDR, a Statistical Computing Environment (SCE) supporting generation of SDTM, ADaM and TLFs, and a Data Consistency Checker ensuring that the data delivered from data collection instruments are delivered in the format defined in the MDR and expected by the SCE.

CONTACT INFORMATION

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

Alan Cantrell
PAREXEL International
1 South Quay Drive
Sheffield / S11 9JN
Work Phone: +44 114 225 1351
Email: alan.cantrell@parexel.com
Web: www.PAREXEL.com