PP21: One Approach to a Metadata and Data Standards Repository

Bill Gibson, SAS Institute Inc., Cary, NC, USA

Abstract

While the concept of a place to hold all data about data is sound, what does such a repository look like? In this poster, a list of prioritized requirements necessary to create a metadata and data standards repository for clinical trial data will be presented. This list has been gathered from multiple sponsor and contract research organizations across the industry. Tradeoffs and justifications for this list will be explained. A reference implementation, based on those requirements, will be shown. The poster will conclude by listing lessons learned after creating and using this reference repository.

An Approach

There have been papers presented at industry conferences, including earlier PhUSE Annual Conferences, that explain metadata and the uses of metadata in describing the many parts and processes of a clinical trial. The goals of this paper and poster are to present one approach to defining some of the main categories of clinical trial metadata, managing that metadata across trials, keeping it under change control and versioning, and most importantly connecting it all together.

The reference implementation used SAS® Clinical Data Integration and its metadata repository (MDR). It provided the capabilities to define data structures, terminology dictionaries, business rules defining data standards, management of use and life-cycles for the data standards, change management of the standards, and mapping of data to the standards as well as across standard categories (e.g. collection to tabulation, to analysis).

Define

Anyone using data standards is already using metadata. Metadata defining data standards have various structures and purposes. Some are simple, such as controlled terminology dictionaries from National Cancer Institute (NCI), Clinical Data Interchange Standards Consortium (CDISC), the World Health Organization (WHO) drug dictionary and the Medical Dictionary for Regulatory Activities (MedDRA). Others describe structures that are tabular in nature such as CDISC's Study Data Tabulation Model (SDTM), Standard for Exchange of Nonclinical Data (SEND), and Analysis Data Model (ADaM), or multipurpose standards like CDISC Clinical Data Acquisition Standards Harmonization (CDASH) or other case report form (CRF) representations (tabular, form, annotations). Metadata can define other, more dynamic data like SDM-XML (study definition), CDISC Biomedical Research Integrated Domain Group (BRIDG) unified data model, or medical ontologies describing connections between terms and concepts.

Some of this metadata is pretty well understood and modeled effectively today. CRF metadata is modeled and managed as part of electronic data capture (EDC) systems. SDTM or ADaM can be represented in XML, spreadsheets, and other structures. SDTM and ADaM metadata are also consumed by standard adherence checking programs, while full study definitions and data systems based on BRIDG are less prevalent.

SAS® Clinical Data Integration was employed in this approach to manage the tabular data standards like CDASH, SDTM, SEND and ADaM as well as the controlled terminology associated with these standards. It provided a repository for global standard/cross-study representations, study and therapeutic versions as well as check adherence against an understood set of business rules (SDTM and ADaM checks). A recent representation of CDISC BRIDG was also imported as a set of connected table definitions. The table-based data, including terms, was easily represented. BRIDG relationships, as well as the less regular SDM-XML structures, were more difficult to model using this approach.

Manage

Definition of the metadata is the first step. Managing its use comes next. Managing metadata requires a way to report and understand how the existing metadata structures are being used. Elements needed to keep a well-managed metadata repository are:

- Versions of the metadata are available, in use, in development.
- Control of the metadata lifecycle including deprecation and possibly retirement.
- Adherence of data/representations of the metadata and enforcement of the standards.

Data standards like SDTM, ADaM and CDASH are all intended to be extensible. A metadata repository must be able to delineate different versions of data standards as well as deviations from the standard for individual projects. Lifecycle controls and version management become important the longer the MDR is operational.

The MDR needs to also show how individual projects or studies conform or deviate from the standards – for example, a list of the CDASH forms that are used, modified or added for a specific study; which controlled terminology dictionaries are in use; and whether those are the current/preferred version along with any re-coding logic if needed.

The specific implementation provides lifecycle management for SDTM and ADaM structures through the stages of development, production and retired settings for data standards. It also will report which structures are in use, deviations, and customizations of those structures for each study or project. Controlled terminology dictionaries are grouped into packages and managed as whole groupings, although no specific management of re-coding mappings is provided. Some process and business rules are not software enforced in the sample implementation. For example, the application of rules dictating the selection of metadata that can appropriately be applied to a study or project such as CRF pages, SDTM domains – including customizations, and the elements in the statistical analysis plan. These rules might be based on study-level metadata like therapeutic area, phase, protocol and completion date.

Change

Metadata and data standards will change over time. Ensure that there is a change control mechanism and process in place. Questions to answer are:

- Will there be a committee or a single owner?
- How to make change requests and manage those requests for change, additions and removals?
- How will requests and changes be communicated? Via wiki or intranet?
- How will the business process with rules be defined and enforced?
- Will there be automatic notification of change to interested parties both for review and final versions?
- How often will changes be made? It is recommended by the author that changes be batched and be made four times per year or less.
- How do changes affect existing projects? Are there different rules for prior to and after data lock? Review impact to existing projects.
- When does existing data get moved into updated data structures and re-coded? How often and using what process?

The PhUSE-FDA Working Group 1 produced a Change Control Board charter for study data validation rules. The charter has a good flowchart describing the various parts and processes that would be valid for change management of other standards as well (phusewiki.org/wiki/index.php?title=CCB_Study_Data_Validation_Rules_Charter).

The example implementation does not provide a software enforced change management process, but as part of a full implementation, SAS Clinical Data Integration can significantly improve data and metadata governance.

Connect

The various pieces of metadata can be managed and used as separate entities, but when combined, the full power of these data descriptors can be harnessed. Consider study design elements designating collected data elements and even CRF template versions. A template statistical analysis plan can define not only the tables, figures, listings and graphs but also the ADaM structures needed to drive those reports. Those ADaM structures will likely be derived in a standardized way from tabulation data described in yet more metadata.

Connections occur across metadata/data types within a given project, across projects and across organizations.

Connections within a project or study show relationships between study definition (SDM-XML, protocol), collected data (CDASH), tabulation data (SDTM), analysis data (ADaM), and results (tables, figures, listings, graphs). These can be defined as mappings and transformations for the collected data showing traceability and computational methods.

Connect metadata across studies via similar data standards and employing study-level metadata describing what and how data can be pooled (what studies are alike and in what ways). Use this connective metadata to enable faster meta-analyses. This same pooling metadata can be used to define and map to a pooled data model based on SDTM, BRIDG, or a custom model. An understanding of more than just tables and columns is necessary. Common terminology, test codes, units and values will be important and should be defined as part of the study design and definition metadata.

Once a good pooling model and good definition of metadata is in place, it is possible to connect patient/participant information with the corresponding safety events, status/process information about participants as well as data management, and even the financial data related to the study.

Ultimately, look toward pooling study data with health data via metadata matching or master data management and unstructured text analysis capabilities. With these additional tools and capabilities, dealing with metadata changes will become more automated, and connecting study data to post-marketing safety will become more commonplace. This is an area of interest and research at SAS – watch for these additional capabilities and solutions in the future.

The example implementation made use of connections within a study to map collected data with SDTM, ADaM, and some safety reporting. Through data standardization and monitoring the SDTM adherence check results, an integrated safety summary (ISS) report was also produced with only minor column mapping and simple recoding of terminology. While the toolset had a representation of BRIDG, no mapping into or out of the model was developed for this approach using instead, the SDTM structures for pooling data across studies.

Conclusion

Employing an MDR is now a reality. Approach the implementation with all four capabilities in mind: define, manage, change, and connect. Ensure that data standards are well defined, properly used, complied with, and connected. These capabilities apply to both to metadata used within a study or meta-analysis to ensure metadata and data consistency at the project level and also across these projects to best support pooling of data and metadata for additional analysis and insights.

The example implementation enabled modeling of many of the data standards needed and useful for performing the data collection, management, and analysis of clinical trials. Defining these structures in metadata can be accomplished with many tools. Definition is only a part of the overall MDR concept. It is also necessary to manage how the metadata is used and the ability to control change over time. The example implementation was able to accomplish all three activities. The key, however, is the ability to connect this metadata across a study as well as across studies and meta-analyses. The approach presented connected collected data to SDTM table and analysis ready structures. It also provided connections across studies to pool data for meta-analysis and data warehousing for analytics. As SDM-XML and BRIDG become more mature and common, SAS Clinical Data Integration will support and leverage this, more complex, metadata as well.