PhUSE US Connect 2018

Paper SI15

The First and Last Words of ADaM

Susan J. Kenny, Maximum Likelihood, Inc., Durham, NC, USA

ABSTRACT

Over the years, the use of the ADaM standard has expanded beyond the original focus of supporting statistical review. Sponsors now use ADaM for nearly all types of analysis summaries. Regardless of this expanded use, the original principles of ADaM still apply yet it appears that adherence to these principles has gotten lost amidst the focus on using standard data structures. This paper provides a walk thru the history of ADaM and how the standard and implementation strategies have grown over the years. We will begin and end with a discussion of what is truly important for ADaM implementation and how we may be better able to achieve the fundamental principles in the future.

INTRODUCTION

It has long been recognized that creation of analysis data sets to support a clinical submission is a process that is tailored to the specific protocol and requires the creation of derived variables and records. Because the creation of these protocol specific analysis data sets involves many decision rules and derivation logic, the fundamental principles of ADaM highlight the importance of clear and unambiguous metadata. This paper will discuss the view that the primary focus of ADaM implementation should be the creation of informative metadata and that adherence to standard structures and compliance checks, while important, are secondary.

THE FIRST WORD

When CDISC was first established in the late 1990's, the development of standards for the collected clinical data known as the Study Data Tabulation Model (SDTM) was the first focus. During the discussions regarding SDTM content, it was well established that derived data, other than simple operationally derived content such as age or study day, would not be included in SDTM. This decision was based on a number of reasons, predominately that the often complex derivations required for analyses could not easily be represented in normalized structure of SDTM. Additionally, and perhaps more importantly, it was recognized that statistical programming and analyses were not performed by statisticians and statistical programmers, not data management, and the creation of the analysis data sets may be conducted long after the clinical data base was designed.

The expectation at the time was that with this standardized clinical data, a data warehouse would be designed and regulatory agencies such as the Federal Food and Drug Administration (FDA) would load submitted data into the warehouse and software tools would allow the clinical safety reviewer to perform their regulatory review. Now that there were data standards to support the clinical reviewer, the CDISC Analysis Data Model (ADaM) team was formed in order to develop standards to support the statistical reviewer. Due to the complex nature of statistical analysis and review, the initial focus of the CDISC ADaM team was to develop standards for the documentation pertaining to the content and the use of the submitted analysis data sets. The goal of developing standards for documentation was to provide clear and unambiguous communication regarding the analysis decisions made by a sponsor in the creation of the analysis and results specified by the statistical analysis plan. With clear and standardized documentation, the statistical reviewer should be able to understand the analyses performed and efficiently perform validation and exploratory analyses.

In 2006, the CDISC ADaM team formalized these nascent standards in a document entitled 'General Considerations v1.0'. This document outlined best practices with respect to creation of analysis data sets and defined types of metadata that should accompany all analysis data sets. The types of metadata described by this document are still supported today by the ADaM model and these include data set level metadata, variable level metadata, and analysis results metadata. However, there was another recommended standard document with was entitled 'Analysis Dataset Creation Documentation'. The purpose of this documentation was to clearly communicate key issues in analysis data set creation, such as description of derived variables, reasons for added observations or omitted observations, imputed data, use of visit windowing, etc. It is interesting to note that when this General Considerations document was updated to become the Analysis Data Model V2.0, all mention of this data set creation documentation was removed.

The focus of the ADaM standard during these first years was to develop standards for documentation and this emphasis is encapsulated in the ADaM fundamental principles. Therefore, the first word of ADaM was METADATA.

CDSIC / FDA PILOT PROJECT

During the time frame of the development of the early ADaM documentation standards, CDISC and the FDA embarked on a collaborative pilot project to test how well the submission of CDISC compliant data sets and metadata met the needs of the both medical and statistical reviewers. This pilot project used de-identified data which the project team converted to SDTM and ADaM. A small statistical analysis plan was created and set of specified analyses were created. All metadata was represented in the new Define V1.0 model.

This pilot project provided useful feedback which impacted the evolving ADaM standards. Primarily, the project underscored the importance of traceability for analysis data sets. To support traceability, the ADaM standard defined the variables DTYPE and ANLzzFL and these variables were the first instance of data point traceability. The ADaM team then updated the first underlying principle of ADaM by including additional text regarding the importance and the requirement for metadata traceability. It is important to remember that the ADaM model requires metadata traceability, with data point traceability being useful when feasible.

ADAM STANDARD STRUCTURES

During the time of the CDISC FDA pilot project, there was a push within CDISC for ADaM to be more definitive with respect to definable data standards. This is reasonable since to be part of a family of end to end data standards, ADaM would need to constitute a standard that is predictable and enforceable. Attempting to create standards for documentation was not viewed as a high priority compared to creation of standards that defined data set structure. It was during this time that the ADaM standards for the Subject Level Analysis Data set (ADSL), the Basic Data Structure (BDS), and the standard for adverse events analysis data set (ADAE) (which is now termed the occurrence data structure (OCCDS) were finalized.

This emphasis on standard structures had a large and lasting impact on the ADaM standards. With the concurrent development of compliance checks, which focused on adherence to the rules of the standard structures, there was a shift away from consideration regarding the standardization of documentation and metadata. There was no longer mention of the importance of data set creation documentation. The prevailing thinking was that the use of standard structures and variables that are defined in ADSL, BDS and OCCDS along with the metadata represented in define.xml would address all requirements for informative and clear communication. Although the ADaM model requires metadata traceability and not data point traceability, many implementers appear to focus on data point traceability as a primary way to achieve communication, regardless of the complexity. The definition of a good and compliant ADaM submission became synonymous with the use of BDS, OCCDS, define.xml, and passing of the compliance checks.

It is important to note that although the fundamental principle of ADaM regarding clear and unambiguous communication still exists, no effort was made within the ADaM team to develop standards or documented best practices with respect to how to represent metadata within define.xml, especially for complex variables. Although the advent of parameter value level metadata was a helpful addition to Define V2.0, the basic difficulty of representing ADaM metadata in define.xml is that the primarily content is variable level metadata yet many ADaM variables are derived from other variables and/or record selection from the source data may differ depending on the variable. With this focus on individual variables, it is nearly impossible to have the define.xml contain metadata in a manner that would allow a reviewer to understand the derivation logic for a data set as a whole or any important cross-variable or cross-record dependencies that may exist. It is unfortunate that the original concept of Analysis Data Set Creation documentation was abandoned because it could address some of these gaps within the current representation of ADaM metadata within define.xml.

At present, many people think of the ADaM standard in terms of standard structures with the use of define.xml as a way to address the fundamental principle of ADaM. Thus simply put, ADaM has become synonymous with structure rather than clear communication.

CURRENT AND FUTURE STATE

While it is important to set goals and look forward, it was Kierkegaard who said ‘Life must be understood backwards, but it must be lived forwards’. Much can be learned about the present and future by reviewing the past. In the case of ADaM, it is important that we revisit the initial focus on metadata and communication so that we align or realign the standard development to fulfill the needs of the primary consumer, that of a reviewer. The following are issues that could be considered in an effort to refocus ADaM standards on fulfilling the fundamental principle of clear communication and traceability.

ADAM STRUCTURES

There are many aspects of the BDS and OCCDS structures that work very well. These structures work especially well when the analysis data set is derived from a single SDTM domain, such as AE to ADAE or LB to ADLB. As a generality, in these use cases, the analysis data set extends the SDTM domain with the addition of analysis ready derived variables that are applicable to the SDTM topic variable. In many instances, there is an easy to trace association between xxTEST and PARAM. When xxSEQ is provided in the analysis data set, it is easy for users of the analysis data set to clearly see how the ADaM data set compares with the source SDTM data set.

However, it is unreasonable to think that the current ADaM standard is without gaps or implementation challenges. One area where ADaM structures do not work well is when there are analysis dependencies between records. Not only is there no mechanism to identify records that are dependent or related to each other but there is no way to express this metadata within define.xml. In some cases, these related records are not analyzed themselves but are present in the data set to provide traceability for how values on another record are derived. Many implementers create these records because of the constraints of the BDS rules yet in practice, having these variables be present on the record where they are used for the derivation of other value(s) that are analyzed would provide easier to follow traceability and easier to document metadata. To address these pain points, it would be helpful if CDISC ADaM developed a mechanism to identify records that are related to each other as well as to consider clarifying that BDS Rule 1 pertains only to variables that are analyzed.

As exemplified in conference presentations or questions on LinkedIn ADaM groups, there appears to be a heightened focus on using only the defined ADaM structures, predominately BDS and OCCDS, for all analysis data sets. Using ADaM OTHER, a sponsor defined structure, is considered by some to be suboptimal. Implementers often 'jump through hoops' to force the derived data to adhere to BDS or OCCDS and seldom give a thought to resulting convoluted metadata. The proposal to allow ADaM OTHER only to be used if it is itself created from a BDS or OCCDS structure is a particularly troubling since it focuses all of the attention on adhering to a defined structure with no focus on clarity or metadata. Standard structures, through their predictability, do support a level of communication but the use of standard structures can never be the sole mechanism for clear communication. Throughout ADaM development, the 80/20 rule has always been used to help guide decisions. The use of ADaM OTHER falls into the 20 (or even 10) percent portion and this sponsor defined structure may be the best for these less frequent situations. To perpetuate the view that it is a substandard way to address a particular analysis need is not only misguided but it has the potential to damage ADaM reputation by forcing the use of a structure that only serves to complicate communication and metadata.

IMPLEMENTATION VARIABILITY

It has always been recognized that ADaM implementation needs to be flexible since the supported analyses are protocol dependent. When accompanied by informative, easy to find metadata, variations in ADaM implementation should not cause too much difficulty. However, there are some simple changes that could be made to the ADaM standard which would foster harmonization across sponsors.

The first of these would be to conditionally require (as opposed to permit) the inclusion of xxSEQ in all ADaM data sets. Having the xxSEQ variable provides an easy way to compare, on a record basis, the content of ADaM versus SDTM. In many ways, using xxSEQ is preferable over the use of the SRCDOM and SRCSEQ variables since it is easier to implement, self-documenting, and easier to use for traceability programming. Additionally, it could be beneficial for ADaM to review the use of SRCDOM and SRCSEQ variables when there are multiple SDTM sources. Retaining the individual xxSEQ variables is self-documenting when there are multiple SDTM domains that are used to source one or more records in ADaM as opposed to the creation of the SRC* variables. This is especially true if variables on the record are sourced from more than one SDTM data set. In this case, multiple pairs of SRCDOM, SRCSEQ variables would be necessary while the simple provision of each individual xxSEQ variable would be sufficient and informative. Though these changes are relatively small in scope, having uniformity across sponsors with respect to the primary variable used for traceability would support metadata clarity and tool development.

Various presentations by regulatory reviewers have indicated that clinical reviewers use both SDTM and ADaM, which is not surprising since ADaM contains all of the derived data that is helpful for review and the promise of tool development based on SDTM has proved to be difficult to achieve. With this in mind, consideration could be givent to industry standards or at the very least strong conventions that would apply to the common safety analysis data sets and to the associated metadata. This would include defining one way to identify the maximum / minimum value on treatment, which in the current ADaM model allows for two approaches. In addition, it would be helpful to have harmonization in whether DTYPE is used for the imputation of values that are above or below a quantifiable limit. In this situation, a record in SDTM exists with xxSTRESC containing a string such as '<3' while xxSTRESN would be blank. In ADaM, AVAL is populated using a pre-defined imputation logic to create a numeric value from the '<3' value. Some implementers use DTYPE to identify this imputation while others do not. Other standard convention could be relatively simple, such as retention of certain SDTM variables, consistent use of analysis variables for the common paired concepts with SDTM, such as ASEV and AESEV, etc. Developing standards for metadata associated with some of the important concepts, such as treatment emergent derivation, comparison of EPOCH and APERIOD, identification of records used for analysis, identification of records that are duplicated in ADaM and the reason for the duplication, etc would be very helpful. One difficulty with imposing these conventions is that there may need to be two versions of these analysis data sets; one to support safety review and another to support the creation of the entire set of tables included in a submission. Assuming that the full analysis data set is an extension of the targeted safety review data set, this should not create an undue burden for sponsors.

WHERES THE METADATA?

As stated above, define.xml is a useful model yet there are challenges not only in the development of the content of define.xml but also in providing a cross-data set, cross-record understanding of the content of the data set. For

SDTM, these challenges are minimal but for ADaM, they are substantial. Given the total lack of standards regarding metadata content and format, it is no surprise that regulatory presentations have indicated that the define.xml often does not provide an adequate level of communication.

The Analysis Data Reviewers Guide (ADRG) was created in part to help document issues that are hard to express within the define.xml. When reviewing the completion guidelines published with the ADRG package, one cannot help but recognize that many of the issues that are recommended to be explained are those that were suggested in the long forgotten Analysis Dataset Creation documentation as defined in the initial ADaM document. For new versions of the ADRG, it is recommended that the template include required questions that must be answered about analysis data sets, predominately questions pertaining to how the ADaM data set differs from the source SDTM data set(s). Despite the recommendations in the ADRG completion guidelines, the content of the some ADRG's is devoid of important information, such as the presence of adverse event records being replicated in ADAE or reasons why a large number of records from SDTM are not included in ADaM, or having one sentence under the header of an analysis data set that reads 'This is the analysis data set used to analyze <<topic>>'. These obvious gaps in ADRG's suggest that the ADRG is being treated as a check-box exercise. This is a troubling trend and must be curtailed before the reviewer guides lose their value.

One challenge with metadata is that there are now so many places where a concept can be documented. For example, the definition of treatment emergent should be in the statistical analysis plan but it should also appear in the define.xml for the required ADaM variable TRTEMFL. It may also be present in the ADRG. If SUPPAE contains AETRTEM, then one would expect the define.xml for SDTM to provide derivation logic. Because there are no standards related to metadata, there is no consistency or predictability for where to find the metadata related to a given concept. And if metadata is provided in the define.xml, there is no consistency in how this metadata is formatted. It may be English readable statement, complex programming code, or some hybrid. Some sponsors will use the 2-level name for each variable (e.g. 'ADSL.TRTSDT') while others do not. Even simple concepts such as the use of the define elements of 'Assigned', 'Derived', and 'Predecessor' are poorly understood, not always utilized, and inconsistent across sponsors for the same variable. If there is any promise that SHARE will support metadata needs of ADaM, we must have definable standards that help align how sponsors express their metadata in define.xml.

It is recommended that CDISC ADaM reduce the focus on developing new data standards or new variables and focus attention on developing detailed guidance on how to create meaningful metadata, in both content and format. Consider the case of the newly proposed ADaM variable PARQUAL. This variable signifies a departure from the current BDS structure by allowing another variable to qualify one or more parameters. The record selection and/or derivation logic may differ for parameter variables when different values of PARQUAL are used. Will this metadata be represented in define.xml as another layer in the parameter value level metadata, at the variable level, the data set level? The ADaM team has spent countless hours discussing the use of this variable but never once discussed best practices for how to represent the metadata for this complex topic. The initial scope of this proposed metadata guidance does not have to be extensive since there are many basic concepts that need to be harmonized. Crowd sourcing and developing some meaningful metadata examples from real analysis data sets would be very helpful.

THE LAST WORD

CDISC ADaM was founded with the primary principle of communication. Whereas data standards are important, they can only go so far in communicating important analysis concepts. In all respects, CDISC ADaM has moved away from developing standards for metadata and this has the potential to significantly reduce the value of the ADaM standard. The suggestions provided in this paper are modest and certainly not an exhaustive list of items that CDISC ADaM could focus on to enhance the standard so that it better supports clear communication. Since analysis will always be variable and complex, we would do well to build an appreciation that the last word of ADaM must be the same as the first – METADATA!

CONTACT INFORMATION

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

Susan J. Kenny
Maximum Likelihood, Inc.
Email: susankenny@maxlikelihood.com