Case Study: Analysis and Metrics of End-to-End Legacy Data Conversions into SDTM, ADaM, and Define.xml

Robert T. Stemplinger, ICON Clinical Research, Redwood City, CA

ABSTRACT

Implementation of the CDISC standards for legacy data conversion presents many difficulties and challenges, and is a time and resource intensive process. Legacy data conversions offer a myriad of issues; incomplete or inconsistent data across studies, varying data structures, inconsistent application of coding dictionaries and/or terminology, all exacerbated by closed sites that offer no opportunity for querying for clarification. This paper will present, analyze, and discuss the implementation process for three projects consisting of 19 studies, reporting metrics from each project and trends that developed as learnings from one project were carried to the next.

INTRODUCTION

Prior to discussing and analyzing the results for each of the three projects, it is necessary to first define scope and provide a brief overview of the implementation process that is followed at ICON Clinical Research.

SCOPE

Of the models that comprise the CDISC standard, the most relevant for legacy data conversion are the Study Data Tabulation Model (SDTM), the Analysis Dataset Model (ADaM), and the Case Report Tabulation Data Definition Specification (CRT-DDS). Included with the SDTM domains would be the necessary trial design data sets, supplemental qualifier data sets, and the related records data set (if dictated via natural relationships in the data). Any full implementation would require the inclusion of at least these elements, though it is still possible for an abbreviated or study specific implementation employing only one or two of them. For the purposes of this paper, a full implementation is assumed.

PROCESS OVERVIEW

The process followed for implementing the CDISC standards is presented below.

CRF ANNOTATION

Any legacy data conversion must begin with the mapping of the legacy data to the SDTM domains. The most efficient use of time and resources dictates that this is begun by annotating the CRFs with the appropriate SDTM domain and variable names. This is often the most difficult and tedious, albeit most important, step in the conversion process.

SDTM/ADaM SPECIFICATION

Once the CRFs have been annotated with the appropriate SDTM domain and variable names, a specification is created that dictates the contents on the SDTM data sets. Included in the specification are all the SDTM domains required to fully model the legacy study data, including standard and user defined domains, and the variables and variable attributes for each domain. Additionally, all controlled terminology is applied in the specification. In the absence of CDISC published terminology, the variables values are standardized across all studies in the project with close consultation with the sponsor.

Once the SDTM specification is sufficiently robust, and an associated Statistical Analysis Plan (SAP) is available, the ADaM specification is created. Much like the SDTM specification, the ADaM specification details what data sets, variables and associated attributes, terminology, and formulas and/or derivations are necessary to create the ADaM data sets.

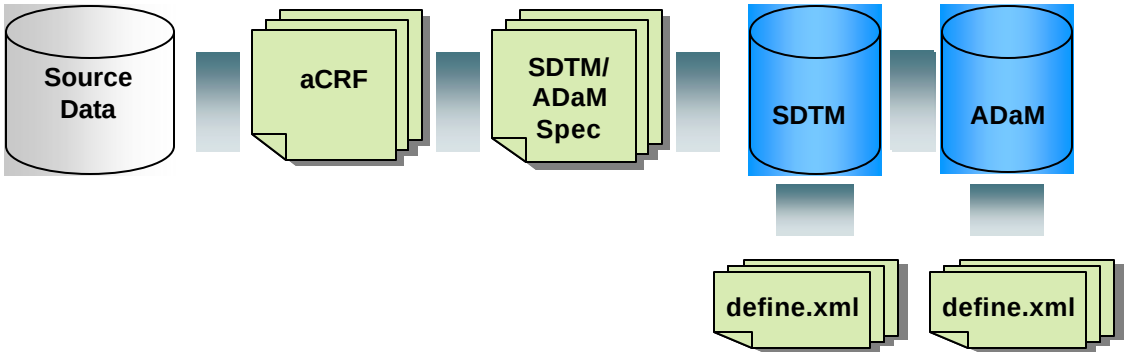
PROGRAMMING IMPLEMENTATION/VALIDATION

Programming of the individual data structures is undertaken using the annotated CRFs, the SDTM and ADaM specifications, and the SAP. Controlled terminology is applied post creation of the SDTM data sets. Validation consists of verification of the structure and content of each data set. Structurally, the data sets are validated against the published SDTM domains. Integrity and conformance checks are automated. Content is validated via review of individual variable values against controlled terminology (either published by CDISC or provided by sponsor), and a visual inspection against a percentage of the source data.

CASE REPORT TABULATION DATA DEFINITION SPECIFICATION (CRT-DDS)

The CRT-DDS, or define.xml, for the SDTM and ADaM data sets are created once all data sets are final. Metadata for define.xml is collected directly from the final SDTM/ADaM data sets. ICON has developed an automated SAS based tool for its creation, augmenting the contents of the document with the necessary comment, formula, and/or derivation information.

Figure 1. Implementation Process



PROJECT OVERVIEW

The projects for each of the legacy data conversions are described below.

PROJECT SCOPE

The first project is comprised of seven oncology trials, with starting dates in 1997 and completion dates in 2007. The study designs are a mix of two double-blind, randomized comparator trials with five open label trials. In total, there are 473 unique CRFS and 372 raw data sets.

The second project consisted of eight sleep disorder trials, all but one a randomized, double-blind design, with start dates as early as 2002 and completion dates in 2008. There are 343 unique CRFs and 205 raw data sets.

The final project consists of four studies of randomized, double-blind design for the treatment of post operative pain, with start dates spanning from 1999 and completion dates in 2008. There are 119 unique CRFs and 164 raw data sets.

Table 1: Project Scope

Project	Studies	Unique CRFs	Source Data Sets
A	7	473	372
B	8	343	205
C	4	119	164

For all three projects, the intention was to apply the linear approach to create the SDTM, ADaM, and Tables, Listings, and Graphs (TLG) deliverables. SDTM data sets were to be created for each study, but ADaM and TLGs were to be created within the confines of an Integrated Safety Submission (ISS). This differs slightly from a strict study by study implementation where each of the deliverables is created for each study, but the lessons learned and the metrics reported are still valuable.

PROJECT STAFFING

Staff assignments consisted of members of database programming (DB) and statistical programming (SP). Experience levels across each group ranged from junior, entry level staff, who experienced their first exposure to SDTM, to more senior level staff with prior CDISC experience. The senior staff had additional responsibilities for reviewing and harmonizing the SDTM structures. The table below details the staff assignments.

Table 2: Project Staff

Project	DP Programmers: Junior	SP Programmers: Junior	SP Programmers: Senior	Project Lead	Total
A	2	1	3	1	7
B	2	2	3	1	5
C	2	0	2	1	3

For SDTM, staff responsibilities included annotation of the CRF, creation of the SDTM specification, programming and validation of the domains, and generation of the define.xml document. For ADaM, activities included creation of the ADaM specification, programming and validation of the data sets, and generation of the define.xml document.

PROJECT DELIVERABLES

Project deliverables were numerous and are shown below.

Table 3a: Project Deliverables - SDTM

Project	Annotated CRFs	Domain Data Sets	Supplemental Qualifier Data Sets	Trial Design Data Sets	Controlled Terminology	Define.xml
A	473	97	53	49	815	7
B	343	143	51	56	1200	8
C	119	56	26	28	612	4

Table 3b: Project Deliverables - ADaM

Project	Data Sets	Define.xml
A	-	-
B	9	1
C	2	1

As each of the projects was ongoing, decisions were made that affected the direction of the project, and caused deviations from the original plans. In Project A, less than positive news was received from FDA at about the time the SDTM programming was being completed, and the project was put on hold and eventually cancelled. Thus, no define.xml for SDTM or any ADaM deliverables were produced. For Project B, it was decided that an abbreviated set of ADaM data sets would suffice as input to the ISS analysis. Consequently, the metrics for ADaM implementation are not as robust as had been hoped for.

ANALYSIS OF METRICS

The metrics for each of the legacy data conversions are shown below.

Table 4: Metrics

Project	Staff	Task	SDTM (hrs) (Avg)	Define.xml (hrs) (Avg)	ADaM (hrs) (Avg)	Define.xml (hrs) (Avg)
A	7	CRF Annotation	273 (39.0)	-	-	-
		SDTM/ADaM Specification	31	-	-	-
		Domain Creation	1162 (166.0)	-	-	-
		Harmonization	27	-	-	-
		Total	1493 (213.3)	-	-	-
B	5	CRF Annotation	224 (28.0)	-	-	-
		SDTM/ADaM Specification	22		30	
		Domain Creation	1224 (153.0)	53 (6.7)	160	36
		Harmonization	36	4	16	-
		Total	1506 (188.3)	57 (7.1)	206	36
C	3	CRF Annotation	72 (18)	-	-	-
		SDTM/ADaM Specification	12		9	
		Domain Creation	308 (77)	18 (4.5)	50	14
		Harmonization	15	2	-	-
		Total	407 (101.8)	20 (5.0)	59	14

SDTM METRICS

With respect to CRF Annotation, an expected difference is noted between the projects. Project A, a collection of very dissimilar studies in oncology, presented many challenges in mapping to the standard. This is reflected by an average time of 39 hours to complete the annotation. A drop of nearly 10 hours is noted for Project B, where four of the seven studies were very similar and only the initial three were disparate. Project C consisted of 4 very similar studies, and the 18 hours for CRF completion very nearly met the ICON metric for task completion. The data then would seem to indicate that study design complexity, and differences across individual studies within a project, have a very real effect on the time needed for annotating

the CRF. Since the experience level of the staff who performed the annotations was roughly equal across projects, this is not an unexpected finding.

Similar findings are noted for creation of the SDTM/ADaM specifications. Since a single specification applied to all studies within a project, there are no per study averages. But it is noted that the complexity of study design and variation among studies had an effect on the time needed to complete the task.

Domain creation consists of SDTM/ADaM data set programming and validation. For SDTM, this is inclusive of the trial design data sets and any necessary data sets for relating records (SUPPQUAL and RELREC). There is less of a distinction here than for the CRF annotation and specification tasks between Project A and B, with a real drop noted on Project C.

Harmonization of the SDTM data sets for the individual studies within each project consisted of alignment of data set structures as well as their content. Again, integrity checks were automated. As expected, the time taken to complete this task seems to be dependent on the number of studies in the project, with design complexity and disparate CRFs having no relevance.

For Projects A and B, the time taken to complete the deliverables for fully compliant SDTM structures exceeded ICON internal targets, with both projects hovering around 5 weeks for completion. Project C did meet that expectation with 2.5 weeks needed for completion. From the data, then, we can gather that legacy data conversions for projects consisting of very dissimilar designs take a good deal longer to complete. The closer the study designs, the more quickly that conversion can proceed. This is not surprising.

ADaM METRICS

With the differences in approaches to ADaM implementation between the two projects that had deliverables created, it is difficult to draw comparisons or conclusions from the metrics collected. For Project B, a full set of nine ADaM data sets was created for the ISS analysis. For Project C, only two ADaM data sets were created. Average time for completion per data set is roughly 17 and 30 hours, respectively. This is a somewhat significant difference. It would appear then, that ADaM creation is not dependent upon study design, but rather some other factor. It would be interesting to consider complexity of analyses to determine its affect, but that is beyond the scope of this paper.

CONCLUSION

LESSONS LEARNED

Perhaps the most significant lesson learned from the undertaking of each of the individual legacy data conversion project was an appreciation for the amount of time required to successfully, and completely, implement the CDISC standards. From annotating the CRF and creating the specifications, to implementing the actual programming, sufficient time is needed to create and review each deliverable – and of equal importance, provide mentoring and training to all staff. With projects consisting of many studies, additional time is needed to ensure consistency across all studies, and subsequent harmonization in the event of differences.

Study design and complexity are two major factors that determine the amount of time needed for completion of tasks. For projects consisting of studies all of varying complexities and all with greatly differing CRF designs, there is virtually no opportunity to expedite the annotation, specification, or programming of SDTM structures from one study to the next. The impact is mitigated for the ADaM structures, which are reliant upon the SDTM structures and seemingly affected more so by the complexity of analyses. This finding is supported by the differences noted between Projects A and B, and Project C, where although study designs varied, CRF designs did not, and much of the annotation and code was portable across studies within the project. Interestingly, Project B was a sort of hybrid of Project A and C, in terms of study design complexity and variation of CRFs, and this is reflected by the metrics as well.

One possible finding that was difficult to tease out of the data, and offered only anecdotal evidence at best, was the effect of programmer experience. It would seem logical that more experienced programmers would generally complete tasks in shorter periods of time, but that was not always the case, especially for the CRF annotation and specification creation tasks. It is reassuring to note, however, that with more exposure and experience built upon a strong foundation of training in the standards, staff quickly increases the speed at which they can implement the steps necessary for a successful legacy data conversion to SDTM, ADaM, and define.xml.

CONTACT INFORMATION

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

Robert T. Stemplinger
ICON Clinical Research
555 Twin Dolphin Drive, Suite 400
Redwood City, CA 94065
Phone: (650) 620-2165
Fax: (650) 591-0611
Email: stemplingerr@iconus.com