PhUSE EU Connect 2018

Paper DS01

A SDTM Legacy Data Conversion

Markus Stoll, German CDISC UN Lead Member, Muehltal, Germany
Laura Phelan, Cytel Inc., Paris, France
Angelo Tinazzi, Cytel Inc., Geneva, Switzerland

ABSTRACT

We would like to summarize a SDTM/ADaM legacy data conversion project which was performed in 2 tranches of each 5 trials from mid-2016 until begin 2018. The conversion was mapped and programmed by an external data service provider, based on the sponsors SDTM/ADaM standard specifications and guidelines by following the sponsor data governance/maintenance approach. The presentation will show the planned approach, the obstacles and hurdles faced in the project lifetime and the resulting preventing actions. A special focus will be on:

- Project Management and Conversion Approach (planned step-by-step approach, delivered documents, timelines versus the real-world experience)
- Quality Control (Standard specifications/documents provided, QC steps and tools used)
- Review Approach (initial and changed review approach based on timelines, resources and QC review experience).

TABLE OF CONTENTS

INTRODUCTION

Ever since the FDA announced that by Dec 2014 it will become a requirement to use the CDISC standards for the submission of study data and published its binding guidance documents, the conversion of legacy study data into the CDISC submission standards has become a major task in the preparation of submissions to the FDA´s CDER or CBER division[1]. Depending on whether the study start date is before or after a certain date specified in the Study Data Standards Catalogue[1], the CDISC standards are either the required or the preferred structure for the submission. Even though the challenges for different data conversion projects are diverse and primarily related to the source data structure, the respective documentation and some lessons learned might be of common interest.

The intention of this paper is the identification of generally valid factors by summarizing the challenges we faced during this conversion project. That´s why we report mainly high-level issues avoiding details on study level.

The main focus will be on project management and timelines, the conversion approach, differences in expectations and the interpretation of guidelines. Lessons Learned will be summarized providing some recommendations and some support for future conversion projects.

CONVERSION APPROACH

A leading global science and technology company, located close to Frankfurt in Germany and further referenced in this document as the 'sponsor', was the pharma company responsible for a planned FDA CDER submission in 2018. To facilitate the regulatory authority´s review process it was planned to convert 9 legacy Ph-II and Ph-III studies and 1 ongoing observational study into the CDISC data submission standard. As the available timeslot for the conversion project was limited, an experienced Data Service Provider Cytel Inc., further referenced in this document as 'DSP', was contracted in early 2016. The DSP was responsible to set up a project plan and to create the SDTM and corresponding ADaM data packages for 5 selected studies and SDTM data packages only for the 5 other studies. In addition, the SDTM tabulation data was pooled at the ADaM datasets level to be used as input for the ISS/ISE, giving a total of 10 SDTM packages and 6 ADaM packages with their corresponding tables, listings and figures (including the ISS) to be delivered by the DSP.

Programming for the SDTM and ADaM data was based on the following sources:

- sponsor SDTM CRF annotation Standards
- sponsor SDTM Standard Guidelines as extension to the SDTM IG v3.1.3/ SDTM Model v1.3
- sponsor SDTM Model including sponsor custom domain definitions, derivation rules and algorithms etc.
- sponsor defined extended Controlled Terminology and VLM definitions
- sponsor defined ADaM Guideline and Standard Template

Validation of the SDTM datasets – which were programmed according to the mapping specifications based on the CRF annotations – was performed with the Pinnacle21 validator in order to ensure compliance to SDTMIG v3.1.3. The resulting issues were addressed by respective updates of the SDTM datasets. In a QC step SDTM SAS datasets were double-programmed and any issues between production and QC programming were resolved, -in cooperation with the Sponsor's team if applicable. After creation of the define.xml and define.pdf a final run of the Pinnacle21 was performed.

Validation of the ADaM datasets followed an almost identical process to the SDTMs – programming was performed according to mapping specifications, double-programming for QC, and P21 checks were run on the define alone, and again with the ADaM XPTs to ensure compliance with ADaM v1.0.

All the original study report outputs were re-generated for full traceability from the ADaMs, and reconciled with the original CSR outputs. Any discrepancies found were discussed with the lead statistician and Sponsor's team as needed, and if not resolved, were described in the ADRG accordingly.

[1] See, References' on page 12

DATA DELIVERY PACKAGES

The data delivery packages were directly uploaded by the DSP into the sponsor's data warehouse. The content of a data package comprised the following components:

SDTM Delivery Package

- SDTM XPT files (Submission ready domain datasets) according to the sponsors SDTM Implementation Guidelines based on CDISC's SDTM Implementation Guide v3.1.3. This includes both subject domain datasets & Trial Design datasets
- define.xml (v1.0)[1] including value level metadata definitions etc. and a corresponding printable define.pdf
- SDTM annotated CRF, bookmarked and hyperlinked chronologically per visit and alphabetically per domain
- P21 community (v2.2.0)[1] validation reports on the XPT files (SDTM 3.1.3) and define file
- Sponsor defined validation checklist
- Legacy data mapping specifications:
  - Legacy source data to SDTM domains
  - Mapping specification of Legacy Terms to NCI SDTM Controlled Terminology
- Listing of all required deviations from the sponsor's standard SDTM model and CT specifications
- Clinical Study Data Reviewers Guide (cSDRG) draft in MS Word and pdf format based on the PhUSE SDRG Template v1.2[1]. The draft cSDRG included already the study specifics, comments on the P21 findings in the Issues Summary section, as well as the Legacy Data Conversion Plan and Report including a description of resolved and outstanding issues. Attached as an appendix was the CT mapping.
- SAS Programs used to create the SDTM datasets
- File Notes to describe inconsistencies in the legacy data if applicable

ADaM

- ADaM XPT files (Submission ready domain datasets) according to the sponsors ADaM Implementation Guidelines based on CDISC's Analysis Dataset Model ADaM v2.1 and the ADaM Implementation Guide v1.0
- TLFs generated from the ADaM data
- define.xml (v1.0) including value level metadata plus stylesheets etc. and a corresponding printable define.pdf
- Commented P21 community (v2.2.0)[1] validation report on the ADaM and define.xml
- Analysis Study Data Reviewers Guide draft in MS Word and pdf format based on the PhUSE ADRG template v1.1 with in addition an appendix describing any eventual discrepancies compared to the original study reports
- SAS Programs used to create the ADaM datasets and TLFs

The legacy data mapping to the SDTM domains and variables was initiated by annotating the blank version of the original eCRF following the SDTMIG v3.1.3 rules and including multiple versions of an CRF module/page to reflect changes due to Protocol Amendments. Based on this aCRF, and external data transfer specifications, the mapping specifications were written to define the content of the SDTM datasets as basis for the programming process. The Trial Design domains were setup based on the available information in the final version of the study protocol and data management related documentation. In the project plan two sponsor review cycles were planned for the aCRF as well as for the complete SDTM or ADaM delivery packages.

Non-resolvable issues identified in the source data or issues and special cases in the mapping process were to be documented and explained in the respective cSDRGs.

Based on the stable SDTM domains, ADaM/TLF datasets were programmed and the TLF datasets were compared electronically with the original clinical study report TLF datasets for the relevant studies.

[1] See, References' on page 12

The 10 studies to be converted were grouped into 2 buckets of 5 studies each. Bucket-II was to be started once the Bucket-I studies were considered almost final.

BUCKET-I AND II APPROACH

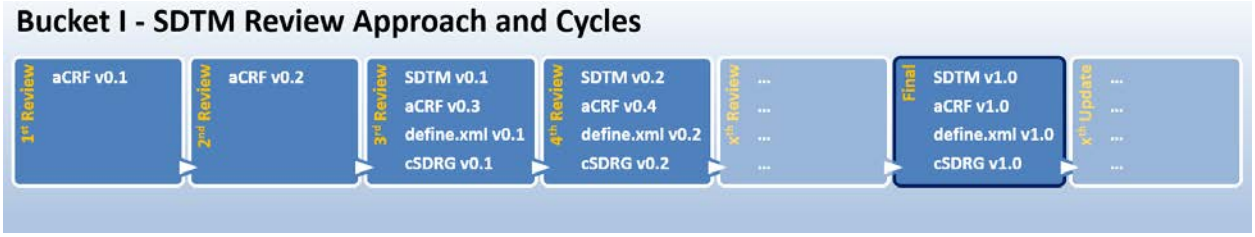
For Bucket-I and II a sequential delivery and review of the study specific deliveries was agreed. Review steps as shown (see Table 1 and Figure 1) were planned for each study. For the 5 Bucket-II studies no ADaM deliveries were planned so none of the ADaM and TLF processing steps were required.

Table 1 – Prospected Data Conversion Process Steps for Bucket-I

Step	Start on Project Day	Duration [days]	DSP	Sponsor
1	0	7	draft aCRF v0.1	
2	7	7		1st Review aCRF v0.1
3	14	7	create aCRF v0.2	
4	21	7		2nd Review aCRF v0.2
5	28	7	draft SDTM v0.1 + aCRF v1.0	
6	35	7		1st Review SDTM v0.1 + aCRF v1.0
7	42	7	create SDTM v0.2 (+ aCRF)	
8	49	7		2nd Review SDTM v0.2 (+ aCRF)
9	56	14	create SDTM v1.0	
10	70	7	draft ADaM v0.1	
11	77	7		1st Review ADaM v0.1
12	84	7	create ADaM v0.2	
13	91	7		2nd Review ADaM v0.2
14	98*	28-56*	create ADaM v1.0 and TLF Reprogramming	
15	126** (70)***	14 14		Review ADaM v1.0 + TLF outputs Review SDTM v1.0 + aCRF

* TLF Reprogramming efforts can vary, these were the estimated figures and latest prospected start date
** Based on actual efforts required for TLF Reprogramming the start date for this task may vary.
*** For Bucket-II Studies no ADaM steps would apply and the SDTM v1.0 review start project day would be day 70.

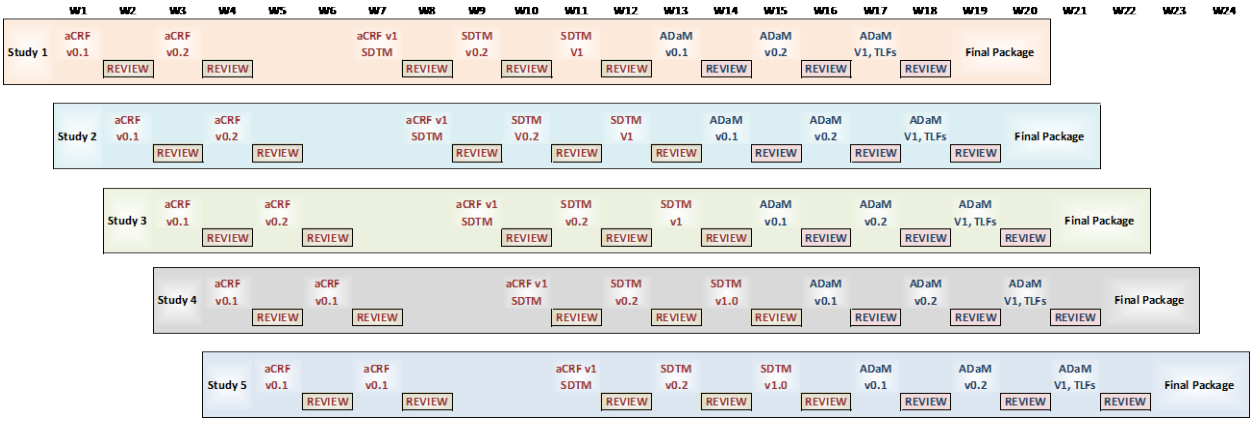
Figure 1 Review Approach and Cycles for Bucket I



For the Bucket-I studies, work on the individual packages was started with a week offset. According to the project plan (see **Figure 2**), this would require a parallel review of multiple tasks. The same approach was originally planned for the processing of the Bucket-II studies.

Based on the experience and challenges in the course of the conversion for Bucket-I this approach was adapted to ensure increased consistency between the individual studies and to reduce the number of review cycles in the later course of the conversion process.

Figure 2 - Staggered Approach for Bucket-I



SPONSOR CHECKS OF DELIVERABLES

The following quality controls were carried out on the deliverables by the sponsor. For SDTM deliverables two sub-teams, each consisting of an SDTM expert and a Data Manager reviewed the deliverables from multiple studies in parallel whereas one of the experts was responsible to consolidate all findings across studies. For ADaM: 2-3 statistical programmers reviewed the ADaM deliverables and each of them was responsible for a certain study.

SDTM-aCRF only

- Review of Table of Content in aCRF versus TOC in Subject PDF
- Visual review of annotations versus sponsor SDTM annotation guidelines and SDTM IG
- Spot checks of annotations in CRF against data in XPT files
- Various consistency checks across studies

SDTM package

- Transfer of XPT datasets was validated by comparison of MD5 hash sums, which were to be created for each file in the DSP´s programming environment and in the sponsor´s data warehouse.
- Consistency check of results in delivered P21 report against results P21 report created by Sponsor
- Check of findings in P21 report against commented issues in cSDRG
- Review of trial design domains and check against cSDRG and Protocol/Amendments
- Manual review (spot checks only) of XPT dataset content versus source datasets
- Visual cross check of SDTM mapping by comparison of SDTM datasets against selected subject PDF´s.
- Spot checks on XML code (Metadata Header) in define.xml
- Manual check on define metadata by visual review
  - computational algorithms
  - VLM definitions
  - dataset definitions
  - hyperlinking in browser view of the define.xml
- Manual consistency checks of define.xml VLM against listed QNAM´s in cSDRG
- Automated consistency checks: define.xml VLM versus annotations in aCRF
  - All metadata in define with origin = aCRF
  - Annotated metadata in aCRF not specified in define file
- Automated checks of define and XPT metadata against CDISC SDTM metadata definitions
  - Controlled Terminology
  - Structural metadata: SDTM Model 1.3 + IG 3.1.3
- Various consistency checks of metadata and mapping across studies

ADaM package

- Review of the ADaM specs for consistency with the SAP
- Review of the ADaM datasets for consistency with the specs
- Review of the ADaM metadata for consistency with Merck ADaM standard
- Internal tools were used to check ADaM compliance in addition to P21
- Review of the analysis results for consistency with the original study reports - manual review of selected TLFs for traceability, as well as programmatic comparison of tables

ISSUE TRACKING

To efficiently collect and track the review results, an issue log was created in the form of an Excel table. For simplicity, one central issue log was planned at the beginning where findings from all studies would be collected if/when the same issue applied to several studies, the affected studies would be flagged in this central log to reduce double commenting. However, this approach was quickly abandoned due to parallel sponsor review and DSP comment implementation in multiple studies, and it was replaced by an issue log per study instead. Ownership of a study issue log was then limited to

- the sponsor upon receipt of a deliverable from the DSP for commenting during their review
- and then to the DSP once each deliverable's set of comments were sent out by the sponsor

Thus, only one "live" issue log existed per study/per SDTM/ADaM package and updates on "outdated" issue logs were avoided.

CHALLENGES AND ISSUES

In this chapter we focus on general challenges or issues either on sponsor and/or DSP side detected in the conversion project. Issues might have affected both the sponsor and the DSP or might have been specific to one party or the other (Details on any specific issues detected in the data processing are out of scope for this paper). Despite the great impact of the one or the other issue and the time pressure, the team atmosphere was always highly professional and friendly. The sponsor and DSP team quickly grew together which was a main contributing factor to the fact that the teams could always stay focused on the common target and issues were quickly resolved.

ISSUES FOR DSP AND SPONSOR

Clinical Study Report Discrepancy Handling

The comparison of the original study report outputs with the CDISC converted output was fortunately automated due having TLF datasets existing for both. However, detecting the actual cause in case of discrepancies proved more time consuming than originally planned, because if differences were found, the source or cause of discrepancy had to be detected and could have one of many:

- a SDTM programming mistake/misinterpretation of the legacy data?
- or an ADaM programming mistake/misinterpretation based on the SDTM data?
- a programming error in the original analysis?
- a misinterpretation of the SAP in the original analysis vs during conversion?
- or simply unclean/strange data which needed special handling?

Issue Log Handling

It revealed to be difficult to follow up on some issues if updates/changes tracked in the Issue Logs were not maintained properly. For example, in some cases no feedback comments were provided for issues reported or with missing details, or the update comments were very high level and not meaningful enough. This ended up on time

consuming checks of documents or data by the reviewer to understand what changes applied – for instance it would have saved time to answer an issue with “SV domain updated for VISIT” vs “Implemented”. If not resolvable, it ended up in additional queries back to the Issue Log provider. Such challenges were directly addressed in the weekly alignment calls among the practitioners.

Standardization and QC Tools

Some of the review tasks could not be automated and revealed to be more time consuming than expected or simply were not considered for manual review in the original time planning (e.g. manual review for mapping of legacy terms and collected free text on CRF to NCI CDISC Controlled Terminology). In addition, the legacy studies were setup and finished before any CDASH compliant CRF standards were in place. In consequence none of the existing standard compliance checking tools was optimized for this project and only a few general compliance checks against CDISC metadata could be automated. The manual efforts required for the compliance review was therefore much higher than originally expected.

Data Quality

Another factor might have been remaining data issues in the data. The time needed to generate the SV and SE domains, and notably EPOCH, was underestimated by the DSP due to for example non-chronological visits. Almost none of the colleagues involved in the conduct of the study were still available, so time consuming explorations, followed by discussions among all experts on the project where required to resolve the findings. The “age“ of the studies, some >15 years old, also added difficulty in the interpretation of information and data structure, notably for the Bucket-II studies.

Review Approach vs Timelines

These challenges had a massive impact on the originally planned timelines and led to delays. Timelines for this project were very tight and as the conversion for the first 5 Bucket-I studies was done more or less sequential, any issues detected in the later phase which impacted all studies (e.g. issues detected during ADaM programming) had an additional impact on the timelines and efforts. Already ‘finalized’ packages had to be updated again which, in consequence, led to unplanned additional reviews.

Depending on the complexity of the changes and the impact on the deliverables, this meant a significantly higher number of reviews had to be done in parallel. Since the expert resources available for reviews were limited more than 2 parallel reviews were not possible on sponsor side. The additional review tasks could only be performed one after another and led to an accumulation of the tasks to be processed. This problem could only be successfully solved through the mid-term allocation of additional resources and a change in conversion strategy.

All in all, however, this led to a significant delay in the completion of the Bucket-I studies. Instead of the originally planned 5-6 months per Bucket it took already 10-11 months just for the completion of the Bucket-I studies. As a consequence, the review approach for Bucket-II was discussed and changed in addition to an additional expert resource which had already been assigned to the sponsor team. It was agreed that for the remaining 5 studies all aCRF´s should be reviewed in parallel to facilitate the consistency across the studies and to reduce the number of late change requests.

The same should apply to the SDTM data review, whereas the SDTM data package delivery was split into 2 review cycles, 1st cycle comprising 2 studies and the 2nd review cycle with 3 studies 3 weeks later. Although the 5 studies were still divided into two parts, the consistency across all studies was significantly higher than in the Bucket-I Studies (Bucket-I > 10 data package deliveries, Bucket-II only 5-6). By adapting the Bucket II approach and the additional resource, the review timelines for all Bucket II studies were met without any further delay.

QC Expectations

During the Bucket-I conversion the need and the effort required to keep the consistency across studies for QNAM´s, Categories or Subcategories or Controlled Terminology definitions was underestimated and ended up in additional updates required late in the conversion process. The main reason for these inconsistencies was the different expectation of the sponsor and the DSP about Quality Checks to be performed before data package delivery. Once this was discovered the team immediately discussed and aligned on the QC expectations for the data package delivery. This important change had also a positive impact on the review cycles required for the Bucket-II conversion.

Review of External Data Mapping

The review of the proper mapping of external data not collected on the CRF was not properly calculated in the review timelines. The consistent mapping and assignment of proper Test Codes of the external data was much more complex than expected as a lot of Biomarker data was collected. One reason was that the naming of Biosimilar Tests in the Data Transfer specifications was quite different and it required an expert review to crosscheck and identify similar tests across the different specifications.

CONCLUSION, RECOMMENDATIONS

TECHNICAL MEETINGS

A fundamental aspect for the successful completion of this project was the good cooperation and collaboration within the project team across companies. A key factor for this good collaboration was the respectful communication among the team members and, if required, the flexible set up of ad-hoc telephone conferences (TCs) for problem-oriented discussions in addition to a regular planned slot. We recommend scheduling a regular technical meeting with the experts involved and to separate this slot from any project planning or status recap meeting.

From experience these meetings are of different nature and should be planned accordingly with the different focus groups. In our project we found having this technical meeting on a weekly basis, with impromptu ad-hoc meetings, and having the project status meeting monthly, to be perfect. At the beginning both the SDTM and ADaM experts attended the technical meeting, which proved to be of benefit as SDTM issues with potential impact on ADaMs could be discussed and immediate decisions taken.

Later in the project, notably when the focus was on the outputs, define and reviewers guides rather than the data conversion, the technical meeting was split for the SDTM and ADaM members.

ISSUE HANDLING

For a successful issue handling, it is highly recommended to use a supporting tool. This tool should enable the team to ensure the proper tracking and processing of all open issues and it should allow recording of the issue processing over time as well as the information about the affected components and the people involved. In this project a simple excel table (see **Figure 3** and **Figure 4**) provided by the sponsor was used, which was also due to the short time frame available. At the end it turned out to be a fortunate coincidence as it easily enabled us to keep the agreed Trial Design draft as excel tables, and the external data mapping table and other information in one document.

Figure 3 - Issue Log Example Part 1

	B	C	E	F	G	H	I	J	K	L	M	N	O	P
1	Deliverable		Bucket I							Question/Finding/Review Feedback				
2					Study 1	Study 2	Study 3	Study 4	Study 5					
3	Deliverable	Date	Num	Status						Document Reference	Page Reference	Feedback	Initial	Date
4	aCRF	01. Jan 16	1	New						domain or annotated CRF etc.	annotated CRF page	If "Finding/Review Feedback" is set to "New" the status cell is highlighted in green and bold text		01. Jan 16
5	aCRF	01. Jan 16	2	Open						domain or annotated CRF etc.	annotated CRF page	If "Finding/Review Feedback" is set to "Open" no cell formatting applies		01. Jan 16
6	aCRF	01. Jan 16	3	Open						domain or annotated CRF etc.	annotated CRF page	Independent of a conditional format applied to cells with the status set in column F "Finding/Review Feedback"... If Status under "Merck Serono Evaluation" is set to "To be Confirmed" the columns H and L are highlighted via condditional format in orange cell shading.		01. Jan 16
7	aCRF	01. Jan 16	3	Closed						domain or annotated CRF etc.	annotated CRF page	If "Finding/Review Feedback" is set to "Closed" cells in column E to T are formatted by grey shading and fading grey text to indicate this task is done		01. Jan 16

Figure 4 - Issue Log Example Part 2

R	S	T	U	V	W	X	Y	Z
Issue Evaluation					Resolution / Follow-up			
Status	Initial	Date	Comments		Status	Initials	Feedback	Date
		01. Jan 16	Example Line - Please delete content					01. Jan 16
		01. Jan 16	Example Line - Please delete content					01. Jan 16
To be Confirmed		01. Jan 16	Example Line - Please delete content					01. Jan 16
		01. Jan 16	Example Line - Please delete content					01. Jan 16

DEFINE AND AGREE UPON QC EXPECTATIONS

To avoid gaps in the QC checking process on the data it is recommended to agree on the QC checks which are expected to be performed on the data packages before delivery. To assure that all expected checks are done a QC checklist can be used which should be completed by the DSP and delivered with the checked data to the Sponsor.

SOURCE DATA MAPPING CHECKS

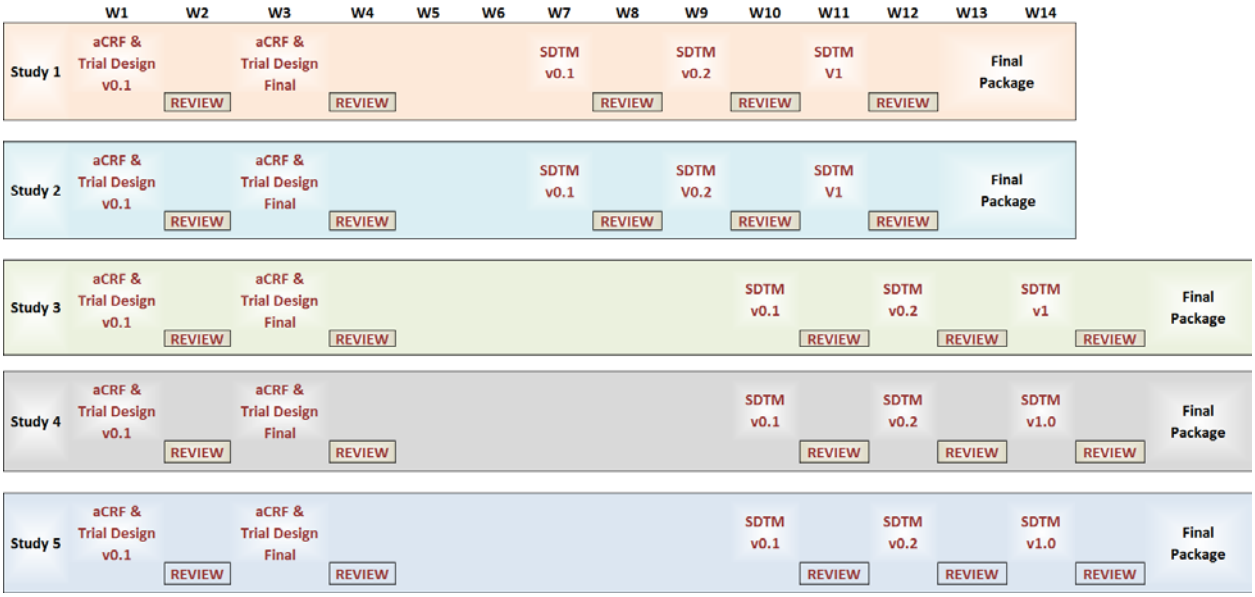
To assure a proper mapping of all data from the source database spot checks were done by comparing selected subject pdfs against the available information in the SDTM database. This turned out to be a very important QC step since issues like incomplete or incorrect mappings could be identified by these manual checks. For example, it was detected that some external data files were missed in the legacy database provided to the DSP. It may also make more sense (time-consuming and economic), for projects to develop more thorough/automated cross checks between the source data and resulting SDTMs rather than double program the entire SDTM set, selecting perhaps a subset of key SDTM domains for double programming instead.

PARALLEL REVIEW BASED ON DATA TYPE VERSUS SEQUENTIAL

Based on the complexity of the staggered review approach faced in the Bucket-I studies in the first part of the conversion project we adjusted to review specifications or deliverables grouped by type across studies rather than sequentially by study. The parallel review of different type of specification documents for multiple studies by multiple team members in a short manner of time is a quite complex task and requires close collaboration as well as high communication efforts and alignment activities. Especially if no automated tools are available to support the consistency checks of the mapping across studies. The review of the annotated CRFs is an important milestone and should be done for all studies in parallel or at last shortly after another to reduce the risk of inconsistencies in the annotations and consequently the mapping to SDTM. In this process it can also helpful to split up the aCRF, group the content by type (e.g. demo, disposition and safety for all studies).to build multiple small review packages which can be reviewed in parallel if resources are available.

Additionally, priority may be considered, if agreed with the ADaM team, to focus first on domains used in the ADaMs/TLFs so they are stable sooner and allow analysis programming to begin while remaining less analysis relevant SDTM domains are finalized.

Figure 5 - Sequential Review Approach for Bucket-II studies

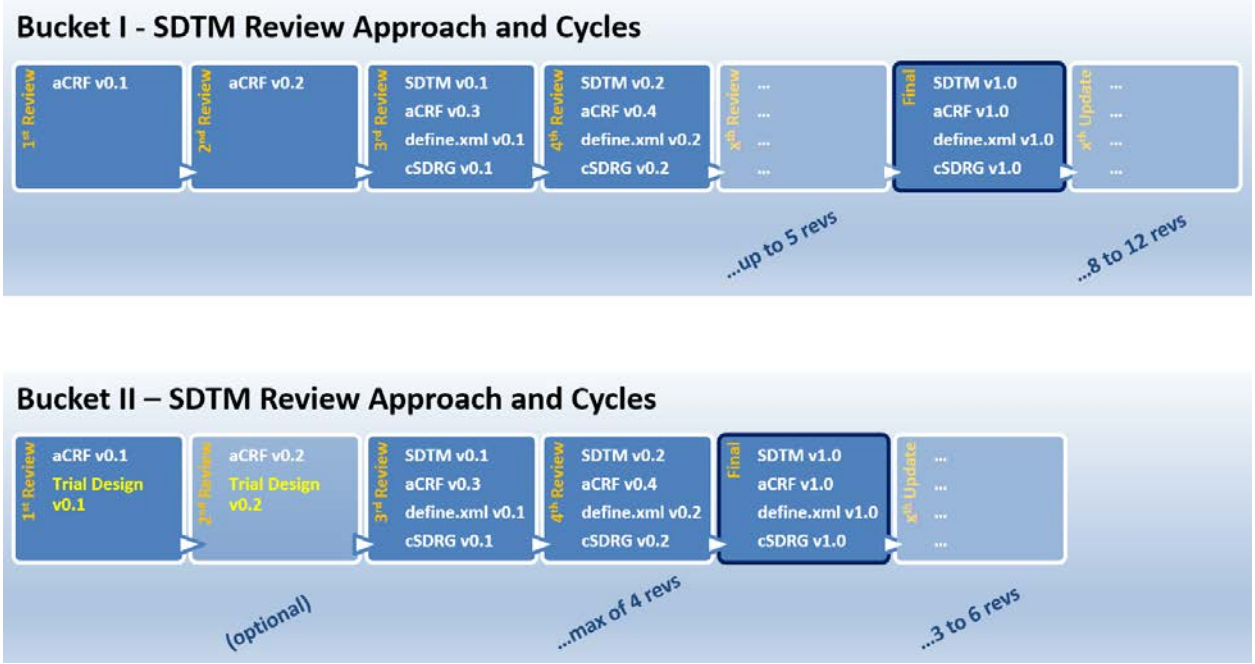


TRIAL DESIGN PRE-REVIEW

To avoid effort intensive re-programming on the SDTM data it can be very beneficial to do a pre-review of the trial design in form of drafted trial design specifications in an excel table before programming starts. Changes in the trial design can have a big impact on all domains and should be avoided wherever possible. Setting up a trial design draft does not require any upfront data mapping, but mainly a protocol review and the knowledge of some other key data and could be done as one of the first steps. Ideally the trial design draft is delivered with the first annotated CRF draft - if not before - to allow time for review/follow-up of outstanding questions while CRF annotation progresses.

Paying attention to trial design at the start also helps considerably programmers in understanding the whys which may only arise once SDTM programming begins, such as: why several versions of specific CRF pages were needed and reason behind the specific changes, how having several protocol amendments which may explain better the different study periods attended across subjects in the same trial.

Figure 6 - Compare of Bucket I versus Bucket II review



CSDRG AND ADRG

To facilitate the Reviewers Guide reviews the order of the findings in the P21 report should be consistent with the order the issues are presented and commented in the respective Reviewers Guide's (SDTM or ADaM). This is a great benefit not only for the reviewer on sponsor side but also for the reviewer at the regulatory authority

LEGACY CT MAPPING

Due to the labor intense mapping and review of legacy terms to CDISC CT, treating the 10 studies raw databases as a whole, i.e. using one central CT mapping excel file, would have proven time-saving. This would also have ensured consistency across the 10 studies.

MD5

We would recommend using the MD5 hashtag (see Figure 7) which was generated by the DSP on the SAS datasets before conversion to XPT outside of the sponsor's data warehouse, and again from within the warehouse on the SAS datasets re-generated from XPTs and a comparison performed. This allowed the team to detect for example, a last minute change in 1 SDTM which was not transferred with the rest of the package which otherwise could have slipped though the net.

Figure 7 – MD5 hashtag

DATASET	UNIQID
AE	ad72578c8a916dd1aad8caaa3421b20c
CE	200c7f311314b763b33416c8c1d1b943

AUTOMATED CROSS CHECKS OF QNAMS/VLM

Was implemented later in the project life-cycle, if had to be redone would be initiated from the start to avoid inconsistencies in QNAM´s, Categories or Subcategories.

EXTERNAL DATA TRANSFER SPECIFICATIONS

In the preparation of a legacy data conversion it is fundamental to get familiar with the respective protocol and the case report form of the study. This allows an experienced expert to get a quite good overview already, but there are still some unknown variables. One such example is the external data which might be available in different formats in the database or in separate locations but is not covered in the aCRF. Thus should also be considered in the mapping. The protocol might not always allow a concrete conclusion on the actually available external data collected especially if multiple external data sources are available. A detailed review of the available external Data Transfer Specifications (DTS) early in the process and annotating the contents can close this gap. The resulting mapping documentation is also of great support for the SDTM database programmer.

FDA INTERACTION

A mock submission towards the end of the project lifecycle was submitted which allowed us to address any findings before the actual submission, findings which may or may not otherwise have led to a refuse to file. For example:

- Sponsor standard to use define xml v1.0 was normally not accepted as support for this version ended by 18th March 2018 – but we were able to challenge this with the FDA team and receive a waiver to continue to use this version with some “tweaking” to explicitly describe VLM
- Initially the FDA were reluctant to accept our use of AP (Associated Person) domains for the Observational study – where a lot of data was captured about the subjects families – but we were able to argue the pros and cons of this approach vs an alternative custom-domain model, and we finally received approval for the AP approach

We encourage early and continual contact is made with the FDA to address issues which are perhaps not typical but which have pragmatic technical solutions may bring compromise.

TEAM SKILLS AND MANAGEMENT

As mentioned, both, the sponsor and the DSP had a positive interaction which saw them through the many challenges. It was also recognized that using several members of the same programming team in both SDTM and ADaM mapping is quite beneficial as they were aware of any SDTM updates and could interject with alternative solutions or approaches which will meet the SDTM needs and in consequence can also reduce possible ADaM/TLF reprogramming efforts. By this approach the team also knows faster the impact of a SDTM change in the ADaMs, though not all companies have people familiar enough with both standards to be able to work this way.

Ideally, as goes for any project, in as far as possible, using the same programming team who performed the original CSR analyses could prove very helpful to understanding the entire data flow circle from legacy->original analysis vs legacy->SDTM->ADaM->TLFs comparison.

FINAL NOTE

In summary, crucial to the success of any legacy conversion includes planning the activities including a proper gap analysis, e.g. ensuring all material is available, protocol vs raw and external data provided, if sponsor standards exist, the DSP needs to assess these carefully so expectations can be clearly defined up-front, agreement on the full nature

of the sponsor surveillance of the DSP and establishing a good collaboration between the two parties.

REFERENCES

Clinical Data Interchange Standards Consortium:
https://www.cdisc.org/

cSDRG Template in PhUSE Wiki:
https://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide

define.xml ODM specifications:
https://www.cdisc.org/standards/data-exchange/odm

FDA Study Data Standard Resources:
https://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm

FDA Data Standards Catalogue:
https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM340684.xlsx

About Message-Digest Algorithm 5 (MD5) in Wikipedia:
https://en.wikipedia.org/wiki/MD5

Pinnacle 21 Community Homepage:
https://www.pinnacle21.com/downloads

Md5
https://en.wikipedia.org/wiki/MD5

LIST OF ABBREVIATIONS

aCRF = annotated Case Report Form
ADaM = Analysis Data Model
ADRG = Analysis Data Reviewers Guide
CBER = Center for Biologics Evaluation and Research
CDER = Center for Drug Evaluation and Research
CDISC = Clinical Data Interchange Standards Consortium
cSDRG = Clinical Study Reviewers Guide
CT = Controlled Terminology
DSP = Data Service Provider
DTS = Data Transfer Specification
ISS/ISE = Integrated Safety Summary and Integrated Summary of Efficacy
MD5 = Message-Digest Algorithm 5
P21 = Pinnacle 21, Small Media Enterprise and creator of P21 Community Validator
QC = Quality Checks
SDTM = Study Data Tabulation Model
TLF = Table Listing Figures
TOC = Table of Content
VLM = Value Level Metadata
XML = Extended Markup Language

CONTACT INFORMATION

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

Markus Stoll
Data Standards Expert
German CDISC User Network Lead Member
64367 Muehltal, Germany
Email: m@rkus-stoll.de

Laura Phelan
Cytel Inc.
63, Avenue des Champs Elysées
75008 Paris, France
Email: Laura.Phelan@cytel.com
Web: www.Cytel.com

Angelo Tinazzi
Cytel Inc.
ICC, Building H, 2nd floor
Route de Pré-Bois
20 C.P. 1839
1215 Geneva 15, Switzerland
Email: Angelo.Tinazzi@cytel.com
Web: www.Cytel.com