CDISC Implementation on a Rheumatoid Arthritis Project Partnership

Patricia L. Gerend, Olivier Leconte, Chris Price, Michelle Zhang

ABSTRACT:

In late 2007 Genentech of South San Francisco US and Roche of Welwyn UK decided to take a risk and begin a new collaborative rheumatoid arthritis (RA) project using Clinical Data Interchange Standards Consortium (CDISC) structures. With little experience among us at the start, the effort has been both fun and challenging. This presentation will address our intelligence gathering approaches, documentation standards, SDTM modeling conventions and conversion processes, ADaM specifications and structures, plans for electronic submission to the US Food and Drug Administration (FDA), and thoughts on efficiencies gained.

INTRODUCTION:

As is customary for many pharmaceutical and biotechnology companies, Roche and Genentech sometimes collaborate on clinical projects, including RA. Since one of these RA projects was just getting started in 2007, we decided to begin its data analysis in CDISC for several reasons:

- CDISC would possibly be required by the FDA when the project was ready for submission
- Data sharing between the two companies would be easier if we used a common structure
- Neither company had any submission CDISC experience and this is rapidly becoming a needed area of expertise

The decision to perform data analysis and submit to the FDA in CDISC was made after case report forms (CRFs) and operational databases were already designed, so our programming team needed to convert existing operational data to SDTM rather than having this provided to us.

We also needed to identify ways of obtaining input on best SDTM modeling practices, ADaM approaches, and preparation of electronic submission materials. The only experience either company had before this was that Genentech had performed two non-production pilot CDISC conversion projects, one with vendor Meta-Xceed in 2006 using SDSv3.1 and another with vendor PharmaStat in 2008 using SDTM-IGv3.1.1. Although these projects in no way made us experts, they did allow us to recognize that although CDISC does have pre-specified structures, there are still many ways of modeling the same data while conforming to SDTM and ADaM. Documenting our metadata could also be done in a variety of ways. We would need information on which potential approaches were best.

Once we had an idea where we were going, we would need to set up documentation standards, model SDTM, convert the operational data, design and program ADaM, and package up our information for submission to the FDA.

INTELLIGENCE GATHERING:

Our first step in readying ourselves for this new CDISC project was to get some training. The Roche contingent attended SDTM and ADaM trainings in Europe, while the Genentech staff attended similar trainings in the US, both in person and on-line. Information on such training can be found on the CDISC web site as well as from independent vendors.

At Genentech, we were also fortunate to be a part of the Bay Area CDISC Implementation Forum, which was founded by John Brega of PharmaStat. Attending their meetings and hearing how other companies were addressing various issues was very informative, and both John Brega and Jane Diefenbach of PharmaStat were extremely helpful in answering our questions and in supplying ideas for approaches.

In addition, some staff from both Roche and Genentech who were not assigned to the RA project had CDISC exposure at prior companies or had developed CDISC expertise by participating on formal CDISC teams. Their contributions to our plans were also helpful.

SDTM MODELING:

The first question to address when performing SDTM modeling is which version of the Implementation Guide to use. When this project began its modeling, v3.1.2 was under review and nearing finalization, so we chose that. We are aware that the FDA has not yet migrated from v3.1.1 to v3.1.2, so we will identify a date by which, if the FDA has not yet converted to v3.1.2, we will retrofit our data to v3.1.1.

Another important question is how to use CDISC Controlled Terminology (CT). This document changes more than any other CDISC deliverable as it is continually being updated with new values. Since SDTM versions are not tied to any particular CT version, there is a lot of leeway on CT implementation. Our decision was to implement the CT version that was active just prior to the first database lock on our project so that this first study incorporates the most current CT and that subsequent studies will be consistent with it.

SDTM involves numerous pre-defined data domains, but data will not always fit into them smoothly. Therefore, it is allowable to create user-defined data domains. These must follow the basic structures of the 3 general observation classes: interventions, events, and findings. All SDTM domain dataset names contain only 2 characters. SDTM recommendations are to start names of user-defined domains with X, Y, and Z. We took this a step further and used the names X_ for interventions, Y_ for events, and Z_ for findings to facilitate recognition of the type of data in the domain. Some examples are XP for previous procedures, YI for previous immunizations, and ZJ for tender and swollen joint counts (part of the composite primary efficacy endpoint for RA).

One seemingly innocuous concept on many CRFs that can cause trouble in modeling SDTM is "Other, specify" and similar questions. When there is only one answer in the "specify" field, the solution is fairly easy: put the information into SUPPQUAL. However, in some cases, there is more than one related question and response. An example is "Was the infusion completed without interruption? If no, specify the following: how was the infusion changed, what was the start time of the change, what was the stop time of the change, and what was the reason for the change?" Luckily, the final version of the SDTMv3.1.2 IG offers the new domain FA (Findings About) which can be used with the RELREC domain for such cases. Without this, you are left to embed sequence numbers into text strings to try to hold together the various responses so they can be linked back to the parent record in the standard domain. This is ugly database design at best, and problematic to use at worst. In addition to using the FA domain, the new variable xxPRESP (pre-specified) can sometimes help to identify records that came from "Other, specify" questions.

In any project, there will also be other questions that can map to various SDTM fields as well as various ways of using ID and category variables. Having open discussions about the different possibilities, with both internal staff and external advisors, served as a good model for making mapping decisions.

Given that this project consists of multiple similar studies, we realized that we needed to document the modeling conventions we were using in order to model efficiently and consistently across studies and to avoid losing time re-inventing the wheel. We developed a document called "SDTM Modeling Information" containing the following sections when we began modeling our first study:

- Conventions for SDTM Modeling
- CRF -> SDTM Domain Map
- SDTM Domain -> CRF Map
- Changes to Annotations Since First Draft

Please see excerpts below for samples of the documentation.

Conventions for SDTM Modeling: Sample

"A Controlled Terminology spreadsheet contains the domain and variable names for any items that have a known, limited number of potential values. In most cases, there is a test name and a test code. In these cases, generally the test name is on the CRF and the test code is developed from that information. There are sometimes other variables with controlled terminology, such as terms, term decodes, units, supplemental qualifier information, and other values. For 1:1 value pairs, such as TEST and TESTCD, TERM and DECOD, QLABEL and QNAM, a sequence number is provided to link the pairs."

CRF -> SDTM Domain Map: Sample

CRF #	CRF Name	Domain
26	Physical Exam	PE
27	Physical Manifestations of RA	ZA
28	Rheumatoid Nodules	ZA
29	Physical Exam	PE

Domain -> CRF Map: Sample

Domain	CRF Name	CRF #
PE	Physical Exam	26
PE	Physical Exam	29
ZA	Physical Manifestations of RA	27
ZA	Rheumatoid Nodules	28

Changes to Annotations since First Draft: Sample

Date	CRF #	Change Description
27 June 2008	12	Added CM.CMPRESP as Y or null
	32	Changed ZX.ZXSTDTC to ZX.ZXDTC
	88	Added DS.DSENDTC, which will have the same value as DS.DSSTDTC.

Controlled terminology is something that, if not documented well, can quickly get out of hand. We used the following spreadsheet to identify which terms were used for specific questions. This allows identification of all terms being used for any particular variable across CRFs. Both values that had official controlled terms and values that conformed to proprietary company standards were included in the spreadsheet. See below for some sample records:

Controlled Terminology: Sample

Domain	Variable	Seq	Label	Original Value	CDISC Std Value
AE	AECAT		Category for Adverse Event	INFUSION RELATED REACTION SYMPTOM	
AE	AEOUT		Outcome of Adverse Event	UNRESOLVED	NOT RECOVERED/NOT RESOLVED
LB	LBTEST	1	Lab Test or Examination Name	HEMOGLOBIN	HEMOGLOBIN
LB	LBTESTCD	1	Lab Test or Examination Short Name	HGB	HGB

An issues log proved to be quite helpful. As someone was going about their business writing SDTM conversion programs or ADaM specs, they would occasionally stumble upon an SDTM implementation that caused problems or that seemed inappropriate to them. They put the issue into a master log, and it was addressed by the team. Any decision made was applied to all relevant studies.

Issues Log: Sample

Issue Detail	CRF Page/ Domain	Raised by	Date Raised	Actioned by	Date Actioned	Status	Resolution Comments
For QS pages, change values of QSEVLINT to ISO8601 format.	Multi	Chris Price	7/15/2008	Patty Gerend	7/24/2008	Resolved	
Ensure that all values of xxPRESP are either Y or null.	69, 72, 76	Chris Price	7/15/2008	Patty Gerend	7/24/2008	Resolved	

SDTM CONVERSION:

Initially our team had planned to use a commercial GUI tool to convert the operational data extract into SDTM. However, the tools we had available were found to be insufficient, so in the end, we settled upon developing and using SAS® template programs. Given that we are all SAS programmers, this choice made it easier for us to understand whether or not our SDTM maps were implemented correctly and also provided us with more control over the conversion process and timings.

Specification spreadsheets were used to bridge the SDTM-annotated CRFs and the operational database extracts. These specifications itemized exactly how to set up the SAS code to implement the SDTM maps. Each variable for a standard domain is assigned a value, an algorithm, or identified as “not mapped”. Please see below for an abbreviated sample specification:

SDTM Conversion Specs: Sample

Domain PE	
STUDYID	PEPE.STUDY
DOMAIN	“PE”
USUBJID	Concatenate PEPE.STUDY, PEPE.CRTN, and PEPE.PT separated by dashes
PESEQ	Unique sequence number of PE observation per subject
PEGRPID	*Not mapped*
PESPID	PEPE.DOCNUM
PETESTCD	“PE”

Domain PE	
PETEST	"PHYSICAL EXAMINATION"
…	
VISITNUM	PEPE.VISIT
VISIT	PEPE.CPEVENT
VISITDY	*Not mapped*
PEDTC	PEPE.DCMDATE formatted as a CDISC ISO 8601 date
PEDY	▪ If PEDTC is on or after DM.RFSTDTC, then PEDY is date part of PEDTC – date part of DM.RFSTDTC + 1. ▪ If PEDTC precedes DM.RFSTDTC, then PEDY is date part of PEDTC – date part of DM.RFSTDTC.

Based on the conversion specifications, SAS code using base SAS tools such as the data step and basic procedures was written and generalized as much as possible for use across all studies in the project.

The SAS conversion process yielded standard SDTM, including the ISO8601 dates and SUPPQUAL datasets. Since ISO8601 dates are not easily usable in analysis, and since sometimes important variables end up in SUPPQUAL, the standard SDTM was converted to more analysis-friendly datasets by converting the ISO8601 dates to SAS dates and by adding the SUPPQUAL records to the appropriate observations in the parent domain before beginning ADaM programming. These interim datasets will not be submitted to the FDA.

ADAM SPECIFICATIONS AND STRUCTURES:

Having made the decision to use CDISC, we were later faced with another decision: should we use the ADaMv2.1 and Implementation Guide v1.0 vertical structure with its parameters and flags? Theoretically this is not necessary since the ADaM model is comprehensive enough to not require this. We could have simply conformed to ADaM naming conventions, created ADSL, and stuck a few ADSL variables onto datasets structured in whatever way we wished. We decided to go with the vertical parameter and flag model for efficacy datasets since, after all, why just take the plunge part-way? We are aware that ADaMv2.1 and the Implementation Guide v1.0 are still fairly new and are currently only in draft format. Also, they are not yet used much at the FDA or even within sponsor companies. Our hope is that the FDA reviewers will be familiar with this model, but if not, we plan to provide ample documentation and training for them. ADaM safety datasets will be a similar structure to SDTM, with data from multiple domains combined into a single dataset and with further derived variables to ensure that they are analysis ready. Additionally, ADSL variables will be added on, such as age, sex, race, and treatment.

The biggest challenge we faced in designing our ADaM datasets was the use of analysis flags. Deciding which were needed was fodder for some very interesting discussions. One hot topic was how robust to make the flags. Should we just create flags to produce the outputs we know will be required for our report, or should we create additional flags in case the FDA wants to perform different analyses? We came down on the side of robust analysis flags since the FDA is just as much a customer of the database as we are.

We also faced other challenges such as how to use the PARAM and DTYPE variables correctly, when to add new parameters (rows), and when to add new variables (columns). The answers are not always obvious, so we spent a lot of time discussing this to make sure our approaches would be compliant with the ADaM-IG and robust enough to meet all needs.

In the pre-CDISC world, FDA expectations of metadata documentation were the provision of a dataset list and a variable list. When using vertically-designed datasets, however, describing derivations for parameters is challenging since each parameter value may have a different derivation and it is awkward to cram so many different derivations into one table cell. Our options for dealing with this included continuing to use 2 metadata tables (datasets and variables) and expanding the documentation to include a third table for value-level derivations. For this project, we settled on using a 2-table approach for datasets and variables. In part, this decision was based on the existing software we have that produces define.pdf. It is possible that for future projects we will include a third table for the value-level metadata. Please see tables below for a sample of the 2-table ADaM specifications where the first is a partial list of datasets and the second is a partial list of joint count variables.

Data List: Sample

Order	Dataset	Description	Structure	Purpose	Key	Location
	ADSL	Subject level analysis dataset	Analysis – one record per subject	Analysis	USUBJID	ADSL.xpt
1	ADJCT	Tender and swollen joint counts analysis dataset	Analysis – one record per joint count per visit/date per subject	Analysis	USUBJID, PARAMCD, AVISITN, ADT	ADJCT.xpt

Variable List for ADJC (Joint Count): Sample

Variable Name	Variable Label	Type	Code	Origin	Derivation Rules
PARAM	Parameter Description	Char	$40	Derived	Descriptions of 66/68 individual joint counts and their total joint counts. The values of PARAM are based on ZJTEST, ZJCAT, and ZJLOC. **Individual joint parameter examples** Right Tender Shoulder Right Swollen Shoulder Left Tender Shoulder Left Swollen Shoulder **Total joint counts parameters** Total 66 Swollen Joints Total 68 Tender Joints Total 28 Swollen Joints Total 28 Tender Joints See the PARAMCD section for all the values of PARAM and its one-to-one mapping with PARAMCD values.
AVAL	Analysis Value	Num		Derived	**Individual joint parameters** Set AVAL values according to AVALC values respectively. AVALC: ‘Y’, ‘N’, ‘ND’, ‘NE’. AVAL: 1, 0, missing, missing.

Variable Name	Variable Label	Type	Code	Origin	Derivation Rules
					For total joints parameters Tender total joint counts are calculated in the same way as the swollen joint counts. **At post-baseline** When PARAM='Total 66 Swollen Joints', two records are generated for this parameter: observed total and LOCF total. Totals calculated here are at all visits and for all patients. (etc.)
TOTTYPE	Total Type	Char	$15	Derived	**Set for total parameters only** For observed totals, set TOTTYPE='Observed' (See AVAL observed total section). For LOCF totals, set TOTTYPE='LOCF' (See AVAL LOCF total section). For Baseline records (where ABLFL is 'Y') set TOTTYPE='Baseline'. For observed totals occurring between screening and study day 1 which are not the baseline result, set TOTYPE='Baseline period'.
ANL7FL	Ana Flg 7 (LOCF,Incl Rescue & Wdrawl)	Char	$1	Derived	**Total joint counts parameters:** For non-missing LOCF total records at post-baseline visits (TOTTYPE='LOCF' and AVAL is non-missing), set ANL7FL='Y' for the visit nearest to the target study day within an AVISIT window. (If one visit record is within an AVISIT window, set ANL7FL='Y' for this record. If there is more than one visit within an AVISIT window, set ANL7FL='Y' for the record with the date closer to the planned visit day. If there is a tie, take the latest record.) If TOTTYPE='Baseline' then set ANL7FL='Y'.

As with SDTM, we created an ADaM Modeling Conventions document to help ensure consistency across studies. Also as with SDTM, we created an issues log to store information on modeling issues/problems as well as the corresponding solutions. These were similar in format to those shown previously for SDTM.

ELECTRONIC SUBMISSION PLANS:

Given that the SDTM model is quite specific, we plan to use WebSDM™ from PhaseForward, which we have licensed, to check the structures of the SDTM data that we created for these studies. If the data does not load into SDTM, the FDA may reject it, which is of course not our desired result. The new version of WebSDM will be able to load SDTMv3.1.2 data and will also

be able to automatically generate associated define.xml documents. It will be important to have an xml document corresponding to our SDTM data for it to load into the FDA's Janus data warehouse where FDA staff can mine it for safety trends across compounds and companies. However, since we are aware that define.xml style sheets sometimes have problems and we suspect FDA reviewers will be more comfortable with define.pdf documents, we will generate those as well.

For ADaM data, however, generation of the define.xml is not automatic and would require considerable resources. Since this is not as important as generating define.xml documents for SDTM, we will only generate define.pdf for ADaM on this project.

EFFICIENCIES GAINED:

The first study that we converted to SDTM took quite a bit of elapsed time (not necessarily full time): approximately 8 months. The second study, however, which was based on modeling, specifications, and conversion programs from the first study, was converted in about 3 months of elapsed time. The third study, which is in progress, is proceeding even more quickly.

Developing ADaM datasets for the first study took about 4 months, including design, specification creation, programming, and QC. These datasets for the second study, although not yet complete, appear to be consuming about half that time.

As with most endeavors, we are seeing establishment of existing software and processes yield efficiency pay-offs. In addition to this, we will have established some data sharing standards across the two companies, which promise additional efficiencies in the future by alleviating the 6 months it generally takes to convert proprietary data structures from one company to those of the other.

CONCLUSION:

The somewhat risky decision of having SAS programmers new to CDISC perform their work in this new structure is succeeding. Teams from both companies are learning a lot about the new structure and are documenting their decisions and rationales. This is helping their respective companies make long-term comprehensive plans around issues such as SDTM and ADaM modeling and electronic submission formats. The support from internal and external experts has made this a doable exercise, and we look forward to a successful filing with the FDA.

CONTACT INFORMATION:

Patricia L. Gerend
Senior Manager, Statistical Programming & Analysis
Genentech, Inc.
South San Francisco, California, USA
gerend@gene.com
650-225-6005

Olivier Leconte
Programming Team Leader
Roche Products Limited
Welwyn Garden City, UK
olivier.leconte@roche.com
+44 (0) 1707 36 5710

Chris Price
Senior Programmer
Roche Products Limited
Welwyn Garden City, UK
chris.price.cp1@roche.com
+ 44 (0)1707 36 5801

Michelle Zhang
Senior Statistical Programmer Analyst
Genentech, Inc.
South San Francisco, CA, USA
zhang@gene.com
650-225-7414