PhUSE 2010

Paper CD06

Traceability between SDTM and ADaM converted analysis datasets

Florence Somers, Business & Decision Life Sciences, Brussels, Belgium
Michael Knoessl, Boehringer Ingelheim, Ingelheim, Germany

ABSTRACT

The FDA CDER is requesting sponsors to submit collection tabulation data in the CDISC SDTM format. Often, the clinical data was not collected in the SDTM format and study data are retrospectively remapped. What happens with analysis datasets which are not ADaM compliant?

Business & Decision Life Sciences and Boehringer Ingelheim are sharing their experience on converting analysis datasets for completed studies into ADaM analysis datasets.

Traceability between SDTM and ADaM converted analysis datasets is a major concern for the FDA and the sponsors. Therefore, a detailed mapping sheet which describes all conversion specifications from the analysis datasets to the ADaM analysis datasets needs to be created. The programming of the ADaM conversion is based on this mapping sheet; focusing on the following items:

- The SDTM variables kept in the ADaM analysis datasets cannot be changed. This means same name, same type, same label and same content.
- The documentation of the conversion of derived analysis dataset variable(s) into the respective ADaM variable(s).
- The original computational algorithm to derive the analysis datasets variable(s) is updated: SDTM dataset(s) and variable(s) have to be used instead of the clinical database dataset(s) and variable(s). The ADaM define.xml has to contain this updated algorithm.
- The creation of new variables: add the unique subject ID variable from the SDTM database, remove sponsor defined formats, create flags for data point traceability, etc.

The conversion of analysis datasets into ADaM analysis datasets is not a simple task. The different steps to take have to be described in a detailed action plan. The ultimate check after converting analysis datasets into ADaM analysis datasets is to (partially) re-perform the statistical analysis. Using the original analysis datasets or the ADaM analysis datasets must give the same results on the statistical outputs.

INTRODUCTION

It is expected that SDTM will be “required for FDA submission” within the next two years. The FDA CDER department is already requesting sponsors to submit clinical raw data in the SDTM format. The FDA CBER department is accepting SDTM submissions since May 2010.

Sponsors convert their original clinical databases into SDTM in preparation for a submission but, what happens with the analysis datasets? The CDISC Analysis Data Model (ADaM) together with its Implementation Guide (ADaM-IG) became effective as of 17 December 2009. Rather recent, therefore, sponsors are currently in a transition period to adapt to ADaM. Thus, for analysis data, it needs to be negotiated and decided between the FDA and the sponsor, whether analysis data will be submitted in the ADaM structure or in the sponsor specific analysis structure for each particular submission. Submission of analysis data in a sponsor specific analysis data structure does not necessarily provide the ADaM required link and traceability between raw data as collected and submitted in SDTM structure to the analysis data as submitted in a non-ADaM structure.

For a recent submission, we created the clinical raw data in SDTM and the analysis data in ADaM structure and, also provided the data point and metadata traceability between SDTM and ADaM datasets as required by the ADaM-IG.

To achieve this, two implementation approaches were investigated:
The first approach (Figure 1) would consist of generating ADaM analysis datasets directly from the SDTM database as created by conversion from the Boehringer Ingelheim's raw database structure for collected raw data. This approach would involve a complete re-doing of the whole statistical analysis required to produce the analysis data, but it would guarantee a high level of traceability from ADaM back to SDTM.

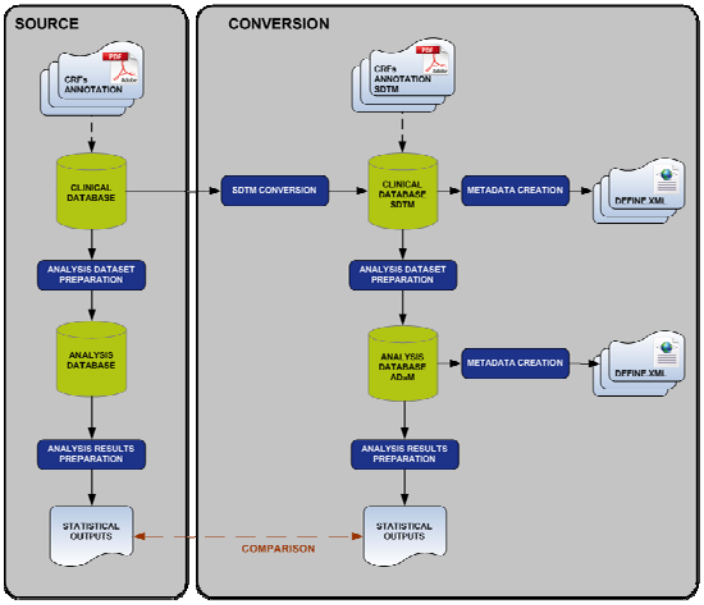

Figure 1: Creation of ADaM datasets from SDTM

The second implementation approach (Figure 2) would consist of generating ADaM analysis datasets from the existing, original analysis datasets similarly to the conversion of the raw data from the clinical database to the SDTM. In addition to the transformation of the analysis data, the traceability between the newly created ADaM datasets and the converted SDTM datasets would have to be manually investigated and established.

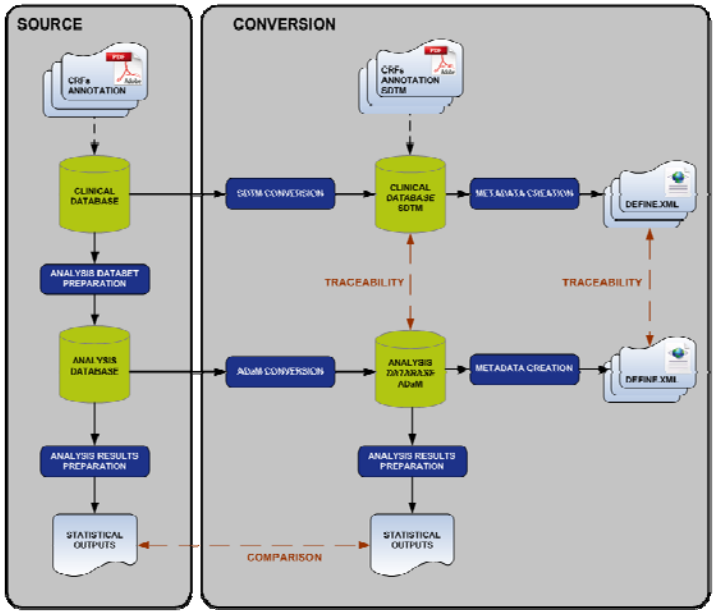

Figure 2: Conversion of original analysis datasets into ADaM datasets

Traceability between SDTM and ADaM analysis datasets is both a CDISC requirement and a major concern for the FDA. Traceability means understanding the relationship between the analysis results, the analysis data (ADaM) and the collected data (SDTM). Traceability is establishing the path between an element and its immediate predecessor. There are two levels of traceability: metadata traceability and data point traceability.

The metadata traceability provides an explanation of how any analysis data information has been created from the raw data as its predecessors by describing the algorithms and derivations used as metadata (e.g. in the define.xml file). The data point traceability, on the other hand, is related to the predecessor record(s): are the records originating from SDTM or are these derived records.

This paper describes the second approach which was chosen for an actual submission by Boehringer Ingelheim and Business & Decision Life Sciences. The process, challenges, solutions and lessons learned from this analysis data conversion are described.

ADAM CONVERSION

Initially, on a trial level, the original analysis datasets were created based on the trial statistical analysis plan from the raw data, both in a company specific data structure, and, in each case with a trial specific flavor. The clinical results were created based on these trial analysis datasets. However, the submission under question encompassed a multitude of individual trials, thus, when preparing this submission, the trial analysis datasets where pooled together into a common structure. This pooling process involved transformation of data values in order to reconcile the trial peculiarities. For example, mapping of trial specific controlled terms to a common, across trial controlled terminology; or mapping of trial specific indicator values contained in the trial analysis datasets that indicated particular derivation routines, or particular uses or purposes of analysis data into a common set of indicator values across all submitted trials. These pooled analysis datasets then served as sole source for the conversion to ADaM datasets. Both pooling and ADaM conversion can be accomplished only with data transformations to some extent. These transformations, however, are attended by the risk of inability in reproducing the trial statistical analysis results as originally created for the clinical trial reports.

The process for converting the pooled analysis datasets into ADaM datasets is represented in Figure 3.

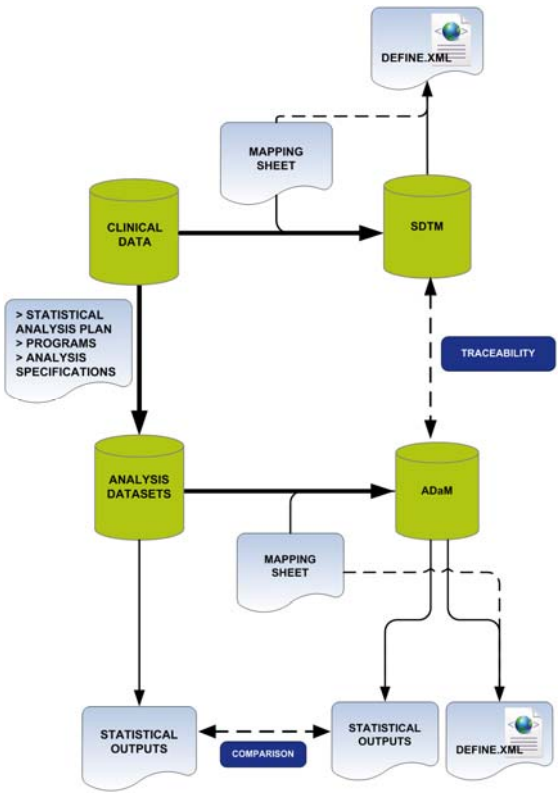

Figure 3: Conversion of pooled analysis datasets into ADaM datasets

The conversion of the analysis data into the ADaM structure started with defining all transformations, and associated rules, required for this conversion, in a separate mapping sheet. This mapping sheet was created by mappers who are ADaM experts. The mappers determined which ADaM datasets and variables were required and determined the specifications on how to convert the pooled original analysis datasets into the ADaM analysis datasets. All of these metadata were then included in the mapping sheet. In addition, the mappers investigated how the derived variables in the original analysis datasets were derived from the original clinical database. Then, for each algorithm a new computational algorithm now based on the SDTM database had to be defined. These new computational algorithms were also included in the mapping sheet and afterwards also included in the ADaM define.xml.

Programmers created the conversion programs based on the specifications from the mapping sheet.

After the conversion, the quality checkers performed the review of the ADaM datasets and the define.xml.

This Boehringer Ingelheim submission consisted of 11 trials for which ADaM datasets were created. Aside from some trial specific datasets, four ADaM datasets were created for each of these trials:

1) ADSL, the subject level analysis dataset,
2) ADaM Basic Data Structure (BDS) for baseline data, containing baseline conditions per subject,
3) ADaM Basic Data Structure (BDS) for treatment data, containing all treatment administration related data,
4) ADaM Basic Data Structure (BDS) for endpoint data, containing the efficacy analyses relevant endpoint data.

As multiple studies were included in this submission; consistency of the ADaM data structure across these trials had to be insured; which was under the responsibility of the data steward.

The last part of the process consisted of comparing the statistical results from the clinical study reports which were based on the original analysis datasets, with the statistical results based on the converted ADaM datasets.

CONVERSION TYPES

The conversions that were needed to transform the pooled analysis datasets into ADaM datasets essentially can be categorized into four different types:

- Typical SDTM variables like the unique subject identifier (USUBJID) had to be created.
  The USUBJID was created by the SDTM conversion team when the original clinical database was converted into SDTM. The ADaM conversion team had to derive this kind of variables similarly.

- Variables for which only a minor conversion was required.
  The content of these variables remained unchanged but the metadata (name of variable, label of variable, etc.) had to be changed in order to be ADaM compliant. For example, the content of the age group variable remained the same but the name and label of the variable had to be changed into an ADaM variable name.

- Variables for which a major conversion was required.
  The content and the metadata (name of variable, label of variable, etc.) had to be changed. For example, the variable containing the gender in the pooled analysis datasets was called SEX and contained 1 or 2. This variable had to be renamed into SEXN and a new SEX variable containing “F” or “M” to reflect the SDTM SEX variable had to be created.

- Datasets that required a transposition, i.e. observations in the source dataset became variables in the ADaM dataset.
  The pooled analysis datasets contained one record per subject per population indicator data value. As the ADaM ADSL dataset may only have one observation per subject, the pooled analysis dataset had to be transposed in order to have the populations as variables instead of observations.

TRACEABILITY

As introduced above, in addition to the conversion specifications in the mapping sheets, the mappers also had to investigate the traceability to SDTM. Two types of variables were identified: variables originating from SDTM and derived variables.

SDTM variables are inherited in ADaM analysis datasets for traceability. When an SDTM variable was inherited in an ADaM dataset, the SDTM variable has to remain unchanged. This means, the variable has to keep the same name, same type, same label and also the same content.

Mappers had to investigate the original computational algorithm for the derived analysis dataset variable(s) which was based on the original clinical database. A new computational algorithm based on the SDTM database was defined, included in the mapping sheet and included in the ADaM define.xml.

QUALITY CONTROL OF THE ADAM CONVERSION

Four levels of QC were defined and performed: compliance of the datasets to the CDISC ADaM model, manual QC of the conversion, QC on consistency and re-production of statistical results from the ADaM datasets.

QC LEVEL 1: CDISC ADAM COMPLIANCE

Business & Decision Life Sciences has developed a tool to check the compliance of the SDTM and ADaM datasets to the CDISC models. This tool contains an expanded and enhanced list of checks which we continue to populate.
A global standard library has been created for the different CDISC models and the SDTM and ADaM datasets are electronically checked against these global standard libraries. The tool generates a report with all non compliance issues found.

QC LEVEL 2: MANUAL QC

After execution of the CDISC compliance checks, a manual QC was performed on a random sample of patient records from the ADaM datasets to check the correctness of the conversion and to check a number of ADaM principles which can not be checked automatically e.g. “one proc away” principle, appropriate redundancy of variables.

QC LEVEL 3: CONSISTENCY BETWEEN TRIALS

A submission often contains multiple studies, and consistency between these studies with regard to data structure and controlled terminology is important. One of the team members was assigned to the Data Steward role for the project in order to ensure this consistency. The data steward runs a report that uses the metadata of all studies included in the project as defined in the mapping sheets. This report shows inconsistencies for the five levels of metadata: dataset metadata, variable metadata, value-level metadata, computational algorithms and code lists.

QC LEVEL 4: RE-PRODUCTION OF STATISTICAL RESULTS

The creation of the ADaM datasets, as explained above, essentially consisted of two data transformation steps:

- firstly, the original analysis datasets of the individual trials were transformed into one set of pooled analysis datasets;
- secondly, the analysis data in the pooled datasets were transformed into ADaM datasets.

These conversion steps, however, are attended by the potential risk of transforming the data such that the statistical results, as originally created from the original trial analysis datasets for the clinical trial reports, may not be reproducible from the ADaM datasets.
Since the same statistical results must be obtained from both the original trial analysis datasets and the ADaM datasets, the final quality control step consisted of checking the quality of the analysis data in the converted ADaM datasets themselves. This was done by reproducing selected statistical results, as reported in the particular original clinical trial reports, now from the ADaM datasets as they were submitted.

The ADaM QC process to check the reproducibility of the statistical results started with selecting more than 200 result tables from the original clinical trial reports of the submitted trials by the project statistician. These result tables encompassed approximately 55 different table types, containing mainly descriptive statistics evaluations but also inferential statistics (e.g. ANCOVA) analysis. These QC relevant tables were extracted from the respective CTRs and compiled in a separate document. In parallel, a separate IT-environment was set up both as a storage place for the SAS® datasets (trial ADS, pooled ADS, ADaM datasets), programs, and ADaM QC related documents and for executing the QC relevant process steps.

A statistical results QC sub team was formed with three statistical programmers. These statistical programmers were not involved in the trial analyses, and thus new and naïve to the project. They had to become familiar with and learn the project and trial details, as well as the ADaM philosophy and details. This learning curve was very time consuming during the entire QC level 4 process. Time to submission deliveries was a critical factor as well. Because of this, the statistical results QC sub team started the QC programming on the pooled analysis datasets rather than on the ADaM datasets, which were created in parallel. Upon availability of the first ADaM datasets, the results QC programming moved towards ADaM as the input source. The results QC itself was performed by ocular comparison of the statistical results obtained from the ADaM datasets with the results from the original clinical trial reports. The progress and status of the ADaM statistical results QC were documented by the statistical results QC sub team in a separate QC plan.

In summary, the ADaM statistical results QC tasks were:

- definition and compilation of the QC relevant trial result tables;
- set-up of a separate QC IT/work environment;

- learning ADaM IG;
- learning the trial and project evaluation;
- QC programming;
- QC itself;
- documentation of the QC process.

COMMUNICATION DURING THE QC PROCESS

The communication between Business & Decision Life Sciences and Boehringer Ingelheim was mainly about the progress of the project and reporting of source data issues, e.g. empty variables that were not converted, the exclusion of screen failures, unclear computational algorithms. Boehringer Ingelheim mainly provided more explanations about the unclear computational algorithms and QC comments.

The information flow between the various sub teams involved in the whole QC process turned out to be of critical importance. A three pillar communication line was established:

1. Weekly meetings with representatives from the three sub teams: statistical and programming project team, ADaM conversion sub team, ADaM statistical results QC sub team;
2. irregular, upon request meeting between the programming project team and the ADaM QC sub team;
3. daily status meetings within the particular QC sub team.

This organization of the communication flow proved efficient as the main types of issues to be resolved were:

1. Questions by the ADaM converters and the QC programmers to the project programmers about particularities about the trial and project evaluations;
2. Questions by the project programmers and QC programmers to the ADaM converters on details about the ADaM datasets.

CONCLUSION: CHALLENGES AND LEARNINGS

Overall, the ADaM creation and QC process depicted above was successful and turned out to be efficient, as all original clinical trial results could be recreated based on the ADaM datasets and the whole QC could be accomplished within the given timeline.

A major success factor was the initial transformation of the individual trial analysis datasets into, though yet non-ADaM, pooled analysis datasets per analysis domain (e.g. one pooled analysis dataset for baseline data, one for efficacy endpoints). This pooled analysis database ensured consistency of the analysis data themselves across the trials. Based on the pooled database, the ADaM conversion and accompanying whole QC process could then, within timeline, efficiently being performed. Moreover, working on a consistent pooled database already granted to a large extent the CDISC and regulatory required consistency within any given submission.

The personnel creating and quality controlling the ADaM datasets was not involved in, thus truly independent from the original analysis of the clinical trials. Therefore, an appreciable amount of time was required to become familiar with the original raw data, the analysis data, the computational algorithms, the medical and statistical rationale of the clinical project and the particular trials. In retrospect, this familiarization was the single biggest challenge the team was faced. Moreover, as ADaM became effective only recently (17 December 2009), learning both the philosophy and the details of the ADaM model and implementation guide was a notable challenge.

Establishing traceability between SDTM and ADaM datasets was of major importance. Facing the tight timelines, the second implementation approach (Figure 2) was chosen: the raw data were converted from the established Boehringer Ingelheim data structure for collected clinical data into SDTM and the analysis data were converted from the established analysis dataset structure into ADaM datasets separately. As a consequence, all original computational algorithms (for deriving and calculating analysis dataset variables based on the original clinical database) had to be investigated and, where necessary, re-defined and adapted to become applicable to data transformation and calculation from the converted SDTM data to the converted ADaM data. These new computational algorithms were included in the ADaM define.xml. Worth mentioning, establishing traceability back to SDTM turned out to be the most difficult part of the whole ADaM conversion process.

With implementing SDTM as the raw data collection format and ADaM as the analysis data format as integral parts of the clinical data process, this additional effort when retrospectively converting clinical data into CDISC structures would become obsolete and a 100% traceability would then be intrinsic to trial and project analysis.

Preparing deliverables for a submission implies working on critical path. This means that a large team needs to be established and has to complete a considerable amount of work within a short time frame. A key success factor was the organized communication path. It allowed all members of the ADaM creation and QC sub teams to stay focused on the topical issues and respond in a timely manner to open issues and questions.

Ultimately, the most important overall outcome was that all original clinical trial results, as selected for the QC, could be reproduced from the newly created ADaM dataset structure, just as from the original Boehringer Ingelheim analysis dataset structure. Statistical programming of the statistical analyses based on ADaM datasets turned out to be straightforward. This ease of use of ADaM was attributed to two major aspects:

- Firstly the normalized organization of the ADaM Basic Data Structure (BDS) was close to the definition of the original trial analysis datasets. Therefore the effort of adapting existing programs to accepting ADaM as input source was moderate.
- Secondly, the redundancy inherent to the "one proc away" philosophy of CDISC ADaM avoided merging of analysis datasets, which we experienced as a clear advantage. Accompanying this aspect of having all analysis data relevant for statistical evaluation stored together in one analysis dataset made it easy to familiarize and access the data and to make use of them for clinical trial and project reporting.

A thorough understanding of the CDISC ADaM model and the ADaM implementation guide was gained by creating and using ADaM datasets. Moreover, being carefully attentive to establishing both metadata traceability and data point traceability between the converted SDTM and ADaM datasets, we better understood about how the regulatory reviewers look and deal with the CDISC submission deliverables.

REFERENCES

CDISC MODELS

- SDTM version 1.2 and SDTM-IG version 3.1.2
- ADaM version 1.0 and ADaM-IG version 2.1
- Web: http://www.cdisc.org/

CONTACT INFORMATION

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

Florence Somers
Business & Decision Life Sciences
Sint-Lambertusstraat 141
1200 Brussels, Belgium
Email: florence.somers@businessdecision.com
Web: www.businessdecision-lifesciences.com

Dr. Michael Knoessl
Boehringer Ingelheim Pharma GmbH & Co. KG
Binger Strasse 173
55216 Ingelheim am Rhein, Germany
Email: michael.knoessl@boehringer-ingelheim.com