

Using CDISC ODM to Migrate Data

By Alan Yeomans

Abstract

The migration of data from a legacy system to a new EDC system poses both technical and regulatory challenges. System architectures differ widely, database structures are not compatible and it is rare that a simple copy-paste type of solution can be applied. This paper describes the choice of CDISC ODM as a mechanism to migrate data from a legacy system that did not support CDISC standards, and the work entailed.

Conversion, cleaning, archiving, migration, export to a legacy data analysis system and validation of the data were all aspects that were included in the migration project. The migration took approximately 6 months to prepare, and 3 days to perform.

Setting the Scene

The sponsor wanted to update a proprietary database solution containing post marketing information comprising over 33 million data points collected over 15 years to a new EDC solution. The data points were derived from 1.1 million forms corresponding to 116,000 visits for 15,000 patients. The study covers 31 countries around the world. The existing study consisted of paper forms (filled out manually onsite) being returned to the sponsor for entry into the database. Laboratory samples with centralised analysis had also been performed. Reports and analyses of the data were then produced and sent back to the sites.

The new solution specified that all sites would use a web-based interface to directly enter patient information. Reports and analyses would then be automatically generated online for users to download directly.

As the original process was largely paper-based, the sponsor decided on a 30 day stop for data entry into the system by investigators. During the first two weeks of that stop data cleaning would be performed on the existing database solution, and then that database would be frozen/locked and the migration to the new EDC solution would commence. All data collected by investigators during this 30 day stop would then be entered by the investigators themselves retroactively once the new system was taken into operation.

During the 15 years that the study had been in operation a number of protocol amendments had been performed. This resulted in data that was relevant but had only been collected in later years, historic data that was no longer collected but was still of use for reference, and historic data that was no longer collected and was not of interest. It was decided that as well as migrating the data, the CRF would be updated (i.e. a new protocol amendment performed) and data that was no longer of interest would be sorted out of the data set and not migrated at all. The sponsor had an advanced collection of SAS scripts for analysis of data in their analysis database, and the decision was made to continue using the analysis database even after the migration to perform advanced analyses.

Data Cleaning

There were a number of data processing steps required, the first of which was data cleaning. This involved completion of data and answering of open queries, basically bringing the existing database fully up to date before starting the migration. From the point of view of the migration process there were only two requirements resulting from data cleaning:

1. The data cleaning must be completed before the rest of the migration process can commence (which was a sponsor activity and responsibility).
2. It must be possible to clean data in the new EDC system even after migration – we do not live in a perfect world and we must make allowances accordingly.

Define Data to be Migrated

Most of the data was to be migrated, but a small portion was judged to be not of interest and was not migrated. This resulted in the following categories of data to be taken into account during the migration:

- Historic data that was not migrated
- Historic data that was migrated, but only for reference purposes – it should not be possible to modify in the new CRF
- Historic data that was migrated, was not included in the new CRF, but which should be editable (allowing investigators to update or correct old data)
- Data that was part of the new CRF (irrespective of whether it was migrated from the old system or directly entered by the investigator into the new system) and which was to be handled accordingly. The majority of the data was in this category.

Archiving of Historic Data

Archiving became a little simpler when it was decided not to migrate all data to the new EDC system. Using the new EDC system to archive historic data was no longer an option. In the end it was decided that the entire existing proprietary database and associated tools and macros would be mothballed and stored by the sponsor in such a fashion that it can be extracted from storage and re-installed if necessary.

Data Mapping

Data is not just data - there were a number of special cases discovered when the detailed mapping of variables was started.

- The first special case that arose was metadata. This included information about clinics as well as detailed information about the data entered (audit trail information and more). As the old solution was basically a paper process there were no electronic signatures or equivalents to migrate.
- Data formats had sometimes been changed in the past, and some were being changed as a result of the protocol amendment being performed when the new EDC system was implemented. Conversion rules for each format change were defined and implemented.
- In some cases independent data fields in the historic data were being combined into one common data field in the new EDC system, and in some cases these did not even have the same format initially. One example was a single numerical field in the new system, which had two numerical fields and one text field from the old system all being mapped into it. This required analysis and conversion of the text field.
- The opposite was also being done – two or more new data fields were being implemented in the new protocol for data that had historically not been differentiated. This required careful mapping so that every individual instance in the old system was migrated into the correct field in the new system.
- The new EDC system allowed specification of allowed values and range checks, something that was used extensively in the new protocol. Many old data values did not comply with these allowed values or were out of range and rules for handling these values were drawn up. In some cases these values were allowed only for historic data, in other cases they were marked as invalid and the investigator was prompted to correct these values (by automatically generated queries) when the new system was taken online.

Data Export from the Old Solution

As the original data was stored in a proprietary database (without support for export in CDISC ODM format) a special application was written to extract all data. The extracted data was then converted to CDISC ODM format and written into files containing a maximum of 500 patients each. This resulted in 31 files to be imported into the new EDC system.

This was the only new programming performed especially for the data migration, and as such Computer System Validation (CSV) of the program must be considered. As this was a once-off, throw-away program (it will never be used again) the question arises about whether or not it is sufficient to validate the data (can the program be considered to be part of the data migration procedure) or is CSV of the export program a requirement? As the data being migrated was from a post marketing study and not a clinical study we decided against CSV this time, but it does raise an interesting question for the future.

Data Import to the New EDC System

The new EDC system contains CDISC ODM import and export as standard functionality, which eased the rest of the process. CDISC ODM import was a little quicker than the export program, so the export program was started first and then after a few hours the import was started. As soon as the export

program produced a new file (of the total of 31) it was transferred to the import program. The entire export/import run took approximately 36 hours.

The new protocol specified a number of calculated values to be produced based on the data entered by the investigator. These values did not exist in the old system, so the calculated values had to be created by the new system for migrated data after it had been imported. Functions and scripts were triggered after all data had been imported and these values were produced.

As the sponsor will continue to use their SAS-based analysis database (which was even used for the old system) a SAS export set is produced by the new EDC system for export to the analysis database.

Validation of Data

The sponsor's decision to use the same analysis database with the new system as for the old system eased the way for a comprehensive and detailed data validation procedure that did not cost much in terms of extra effort. The decision was made to compare the SAS-export from the old database to the SAS-export from the new database and use that to verify that the data had been migrated correctly.

The goal was to be able to perform a file compare of the output from both systems, but in order to do that SAS scripts had to be written to take into account the data mappings and conversions described above. From a validation point of view this resulted in an independent (in terms of programming) set of scripts to validate the data migrated.

The result of the SAS file comparison still produced a number of deviations (that could not be programmed into SAS). Each of these deviations then had to be checked and accounted for in the final validation report.

Finally end-to-end data validation still had to be performed to confirm that the original source records had been correctly migrated into the new EDC system, and for this validation even the data cleaning had to be taken into account.

Why CDISC ODM?

The choice to use CDISC ODM format for the data migration was simple when we considered the following advantages:

- Reuse of import functionality. The export function was a once-off, throw-away product and appending a CDISC ODM format conversion to that product was additional work. But by using CDISC ODM for the import format we anticipate that we will be able to reuse the import functionality many times in the future, and thus save design effort in the long term
- The use of the CDISC ODM format resulted in a preliminary data correctness check before import – the data had to fit into the CDISC ODM format or there was something wrong with the initial data
- Separation of export and import eases debugging and validation. The fact that human readable CDISC ODM files were produced as the interface between these processes made everyone's life easier
- Computer System Validation of the new EDC system covers the import functionality, and therefore we only need to perform CSV of our product once for many data migrations

Good Practice

The migration taught us a lot about how to migrate data and (surprisingly enough) not too much about how not to migrate – the choices we made for this data migration were confirmed by the success of the operation. The first users started using the system the same day it went online and 6 new patients were entered into the database during the first two weeks of use.

Other major factors (besides the use of CDISC ODM) that contributed to the success of the data migration were:

- The production of a detailed mapping of all variables including format conversions and data field combinations and splits
- The production of a data decision document in advance of the migration, deciding how to handle different non-compliances before they arose
- Multiple test runs of the data migration (starting at one patient, one clinic, one country, etc.) to identify special data cases that need to be handled and other problems well ahead of time

The next steps

Where do we go from here? This time there were a number of areas that we could ignore as we were migrating data from a paper based process even though the data came from a database.

We have already commenced our next data migration project and it is a step up in complexity – the migration of data from a proprietary web-based eCRF system with clinics, users, audit trail, queries, comments and electronic signatures. This system does not support CDISC ODM so we will use a similar methodology to that described above, however there is a lot more regulatory information (21 CFR Part 11 compliant information) that must be included and handled correctly this time.

After that the next project will involve migrating data from another commercially available EDC system, one that does support CDISC ODM. We suspect that this may not be as simple as it sounds – CDISC ODM does allow users a certain degree of freedom in specifying customised fields, and the sponsor may decide to perform a protocol update at the same time. Therefore we suspect we may still need to perform a certain degree of “post-processing” on the CDISC ODM files exported from that system before they are ready for import into our system. Or maybe CDISC ODM really will solve all our problems...