

In at the Deep-End (thoughts on mapping multiple studies into CDISC for an ISS/ISE)

Andrew Noller, Quintiles, Bracknell, England

ABSTRACT

I have been working within a team mapping data from multiple studies for an ISS/ISE. These studies have consisted of different designs (e.g. parallel 'vs' cross-over), different domains (e.g. diary data, labs, ecgs) and different data structures.

In this paper we will share my experiences of interpreting the CDISC guidelines and mapping data into SDTM format, focusing in particular on some initial teething problems, coding and formatting issues and deriving study design datasets. Highlighting some of the added complexities involved with mapping multiple studies at the same time and our strategy in dealing with this.

INTRODUCTION

If you have ever had to map legacy data for a single study into SDTM then you will know some of the hurdles that have to be overcome in order to get your data into the correct CDISC format. Now consider that you have many studies that you wish to combine for an ISS and/or ISE and conversion to SDTM first would allow you to create an integrated SDTM database platform to work from. In this paper I aim to show how we approached this and how we overcame the many challenges.

The scope of the ISS/ISE was based on 37 studies being mapped for 20 SDTM domains not including supplemental qualifiers. For this paper it helps if you have had a little experience of the CDISC SDTM Implementation Guide as this will be referred to throughout.

This package of studies would be used as a basis for including pooled analysis of safety and efficacy data that is put together to submit to the regulatory authorities such as the FDA and EMEA. When we started working on the studies not all of the databases had locked (a large handful of the studies were ongoing) and with the approx 25 studies we had already been given, we soon discovered that we had at least 6 different database structures. We were to program derived datasets from scratch for these studies as we did not have programs available from the legacy studies.

I also had the personal challenge of having to very quickly get up to speed with the CDISC Implementation Guide (we were using version 3.1.1. at the start) and I was fortunate enough to have some internal formalised training from one of the CDISC SDS team.

DATA INTEGRATION STRATEGY

Before any work was started on the mapping and integration of any of the studies a Data Integration Plan was written. The purpose of this document was to provide background information on the data that was to be combined into two integrated analysis databases that would be used to summarize safety and efficacy for the investigational drug. The integration and manipulation of these data, as well as the algorithms for the created variables would be documented in detail in the Data Definition Tables (DDTs) and corresponding derived data set specifications. All statistical methodology, as well as details on the statistical summaries to be created from the integrated databases would be documented in the Integrated Summary of Safety (ISS) and Integrated Summary of Efficacy (ISE) Statistical Analysis Plans.

The Data Integration Plan document was intended to provide (1) an understanding of the types of safety and efficacy data that would be integrated; (2) the methodologies used for integration; and (3) details of any overall manipulation and transformation of data for integration. In addition, details on the data transfer, receipt, storage, and retrieval process were provided in the integration plan.

PhUSE 2008

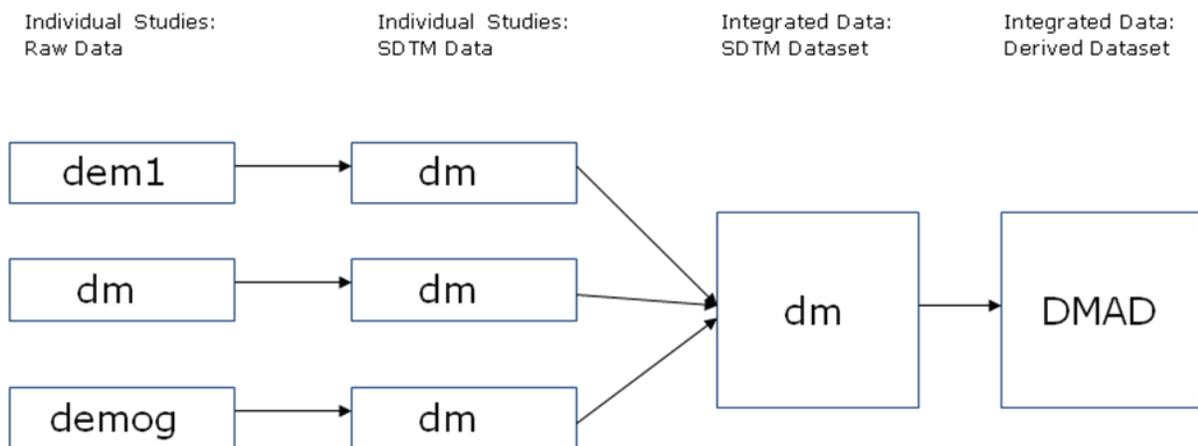
We created Study Data Tabulation Model (SDTM) data sets (for the selected domains required for the ISE and ISS) for each of the individual studies from their original raw datasets. These individual study SDTM data sets would then be combined into one integrated SDTM data set per domain. One Data Definition Table (DDT) would be produced with each integrated SDTM dataset but it had been considered to produce one DDT for each study domain. A DDT is a CDISC specification that details what variables are in the raw CDISC dataset for that domain (and associated derivations). Within each DDT, details of how variables are mapped into a CDISC format would be captured on a study specific level if needed. Example DDT variable definition below:

| Variable Name | Variable Label | Type | Controlled Terms or Format | Origin | Role | Sponsor Comments |
|---------------|----------------|------|----------------------------|--------|------------------|---|
| SEX | Sex | Char | **M, F, U | CRF | Result Qualifier | Studies a,b,c,d,e,f: Convert SEX1C using SEX1V format and select first letter Studies g,h,i,j,k: Convert SEX using SEX format and select the first letter Study p: Convert SEX using SEX1v format and select the first letter Study r : First letter of variable GENDERX Studies s,t,v : Convert SEX using SEXF format and select first letter |

Practicality was an issue when thinking about the creation of the DDT's as when submitting to the FDA you need to have one DDT per study but for ease of programming we decided to have 1 large DDT rather than 37 and at the end this could be split into 37 documents.

The integrated SDTM data sets would be used to create derived data sets for efficacy and safety. Figure 1 illustrates the stages of the process for the demography domain.

Figure 1: Four Stages of Data Integration



Below is a table of the domains that were required for the summary's of safety and efficacy:

Table 1: SDTM Domains Required

| Domain | Description | Comment | ISS | ISE |
|--------|------------------|---|-----|-----|
| TA | Trial Arms | | Yes | Yes |
| TE | Trial Elements | | Yes | Yes |
| SE | Subject Elements | Key data set. This defines what study element (i.e. | Yes | Yes |

PhUSE 2008

| | | | | |
|----|-------------------------|--|-----|-----|
| | | treatment) a subject was in and when. | | |
| TV | Trial Visits | Required if visit windows to be used | Yes | Yes |
| SV | Subject Visits | Required if visit windows to be used | Yes | Yes |
| DM | Demographics | | Yes | Yes |
| SC | Subject Characteristics | Baseline subject characteristics | Yes | Yes |
| EX | Exposure | | Yes | Yes |
| DA | Drug Accountability | | Yes | Yes |
| DS | Disposition | By trial epoch. Screening, Titration, etc. | Yes | Yes |
| CM | Concomitant Medications | | Yes | Yes |
| AE | Adverse Events | | Yes | Yes |
| MH | Medical History | | Yes | |
| EG | ECG | | Yes | |
| LB | Laboratory Tests | | Yes | |
| VS | Vital Signs | | Yes | |
| YD | Diary Data | | | Yes |
| QS | Questionnaires | | | Yes |

One consideration was would all possible raw data be mapped to SDTM or only that which was required for the integrated databases? Also, data from ongoing long-term studies (e.g. open label extension studies) would be included based upon a predetermined data cut-off date for each study. Efficacy data for the ongoing blinded studies would not be included until study completion.

As an example of the process for one of these domains consider the Demographics (DM) integrated dataset. Individual studies would have their original raw demographic data sets integrated into one final combined derived data set (by following the four stages shown in figure 1). This final integrated derived data set then contains data from all of the original studies and can be used to provide summaries for each individual study group and for any combination of studies/groups.

The studies being integrated were based on the same study compound for a single customer. However, as the studies were collected over a number of years and a handful of indications (for the same compound) there were changes in CRF templates, data collection standards and in data management from different companies. This highlights one of the main reasons why it was useful to map the data into SDTM format before attempting to integrate the datasets to conduct a pooled analysis.

Due to the nature of integrating databases there could be differences between the analysis results produced from data in the integrated databases and the individual study results. Such discrepancies may be due to any one of the following:

- Data issues are handled using a uniform approach across all studies rather than in any study-specific approach
- Derivations of parameters in the ISE and ISS differ from those in the original studies
- Populations defined in the ISE and ISS differ from those defined in the original studies
- Data is recoded using a standard terminology/code (e.g. AEs, race, country, medications etc.)

With tight timelines on a critical path we identified those domains that we needed for all the pooled analysis so that we could return to the unmapped variables at a later date that could be in the form of a different domain (CO for comments) or a supplemental qualifier (SUPPEG for extra info on ECG's). We were going to have a separate team who would ensure all domains on the CRF were mapped.

The unique subject ID (unique across all studies in the integrated database) would be derived as a concatenated Study ID, Site ID, and Subject ID variable. Open label subjects will have the same USUBJID across both protocol numbers (i.e. subjects from study A which continue into study B will have the same USUBJID but different STUDYID) would be derived as a concatenated Study ID, Site ID, and Subject ID variable.

Open label subjects would have the same USUBJID across both protocol numbers as this is the same subject within the submission

PhUSE 2008

| STUDYID | SITEID | SUBJID | USUBJID |
|---------|--------|--------|---------------|
| ABC987 | 0001 | 0001 | 987-0001-0001 |
| ABC986 | 0001 | 0001 | 987-0001-0001 |

MAPPING AND FORMATTING ISSUES

One of the trickier domains was LB due to differences over the course of all the studies in lab test name as well as the original unit. Thus at the SDTM stage in order to have consistency for when we create one overall LB dataset we required a unique Lab test name (LBTEST) as well as a unique standardised unit (LBSTRESU). In order to process this information we put unique combinations of the raw test names and units into a spreadsheet, which could be read into the SAS program. It should be understood that in some cases the raw datasets had different parameter names for the lab test and unit in some of these studies so these had to be mapped to unique raw terms before the spreadsheet could be used. The spreadsheet also contained a conversion factor from the raw data to the SI units that would be chosen for the ISS. Below is an example of three parameters in this document:

| RAW TEST NAME | RAW UNIT | LBTEST | LBSTRESU | LBTESTCD | CONVFACT |
|-----------------|----------|-------------|----------|----------|----------|
| GLU | MMOL/L | GLUCOSE | mmol/L | GLU | 1 |
| GLUCOSE | MG/DL | GLUCOSE | mmol/L | GLU | 0.05551 |
| GLUCOSE (URINE) | MG/DL | GLUCOSE | mmol/L | GLU | 0.05551 |
| GLUCOSE | MMOL/L | GLUCOSE | mmol/L | GLU | 1 |
| HEMATOCRIT | L/L | HAEMATOCRIT | % | HCT | 100 |
| HAEMATOCRIT | FRACTION | HAEMATOCRIT | % | HCT | 100 |
| HAEMATOCRIT | % | HAEMATOCRIT | % | HCT | 1 |
| HAEMOGLOBIN | G/DL | HAEMOGLOBIN | g/L | HBN | 10 |
| HAEMOGLOBIN | G/L | HAEMOGLOBIN | g/L | HBN | 1 |
| HAEMOGLOBIN | GM/DL | HAEMOGLOBIN | g/L | HBN | 10 |
| HAEMOGLOBIN | MMOL/L | HAEMOGLOBIN | g/L | HBN | 16.5 |
| HAEMOGLOBIN | G/DL | HAEMOGLOBIN | g/L | HBN | 10 |

Another programming area for concern was the formats received on the raw data for certain variables. There are no formats in the SDTM conventions therefore to map raw data we need to map formats into suitable categories. This is easy when you have one study but things can get complicated when you have multiple studies.

Examples are as follows:

AE – Related (AEREL)

Action taken (AEACN)

Outcome (AEOUT)

DS – Patient withdrawal categories (DSDECOD)

DM – Gender (SEX)

race (RACE)

To illustrate, in demography (DM) the variable RACE was described by five different formats across all the studies that would then require mapping into one CDISC option at the SDTM stage so that the studies could then be integrated as one. The table below was used at the end of our DM specification:

| CDISC variable name | Original raw data CRF Value [format value] | CDISC mapped value |
|---------------------|--|--------------------|
| RACE | 2 [BLACK] | BLACK |
| | 3 [ASIAN] | ASIAN |
| | 6 [WHITE] | WHITE |
| | 88 [OTHER] | OTHER |
| | | |

PhUSE 2008

| CDISC variable name | Original raw data CRF Value [format value] | CDISC mapped value |
|---------------------|--|----------------------------------|
| RACE | 1 [CAUCASIAN] | WHITE |
| | 2 [BLACK] | BLACK |
| | 3 [ASIAN] | ASIAN |
| | 88 [OTHER] | OTHER |
| | | |
| RACE | 1 [Caucasian] | WHITE |
| | 2[Afro-Caribbean] | BLACK |
| | 3[Oriental] | ASIAN |
| | 4[Asian] | ASIAN |
| | 5[Other] | OTHER |
| | | |
| RACE | 1[Europid] | WHITE |
| | 2[Oriental] | ASIAN |
| | 3[Asian/Indian] | ASIAN |
| | 4[Afro-Caribbean] | BLACK |
| | 5[Other (specify)] | OTHER |
| | | |
| RACE | 1 [Black] | BLACK |
| | 2 [White] | WHITE |
| | 3 [Hispanic] | WHITE |
| | 5 [Native American] | AMERICAN INDIAN OR ALASKA NATIVE |
| | 6 [Asian/Pacific] | ASIAN |
| | 8 [Other] | OTHER |

We created tables like this at the bottom of our DDT's to assist programming but not sure if we would have been able to submit the DDT's to the FDA like this.

Another aspect of mapping to SDTM that occurred was the creation of supplemental qualifiers in order to capture non-standard variables and their association to the parent domain. This could be for extra coding variables found in adverse events (AE) for example). As we were planning to bring in one study at a time to the SDTM integrated database as each study locked, it meant we could not pre-empt exactly what supplemental datasets were required in planning as the last study locking added to the database could have non-standard variables that were particular to that study.

STUDY DESIGNS

The Integrated summaries of safety and efficacy were to be based upon a total of 37 studies. These were not all studies of similar type either; a combination of phases 1, 2 and 3; parallel, cross-over and dose-escalation. Add to these a great variety of planned treatment groups and studies with single or multiple dosing regimens.

All of this information plus other minor considerations meant that the integration of these would be difficult and an effective way of comparing these was in production of the CDISC trial design datasets for each study. These include Trial Arms (TA), Trial Elements (TE), Trial Visits (TV), Subject Elements (SE) and Subject Visits (SV) as the main trial design datasets used to describe each study.

From the CDISC SDTM Implementation guide the following are defined. A Trial Arm (TA) dataset contains one record for each occurrence of an Element in each Arm of the trial, a Trial Element (TE) dataset contains one record for each type of element in TA and a Trial Visit (TV) dataset describes the planned visits in a trial. Subject Elements (SE) and Subject Visits (SV) datasets take the structures of TE and TV respectively and then include actual subject data to the theoretical trial design.

The studies we were working with had never been mapped into SDTM before so we started by creating an individual TA, TE, TV, SE and SV for each study by the study identifier. Once these were produced for all the studies they could only be combined by appending one upon another. We also had just one specification available for each of these domains.

PhUSE 2008

CODING CONVENTIONS

In combining legacy and current studies it is clear that different versions of coding dictionaries will have been used for adverse events and medications. Thus, a consistent version must be chosen before the majority of the studies are re-coded. In addition, in focusing on the AE and CM domains for this I will highlight any other difficulties we had in the creation of these including supplemental qualifiers for these.

ADVERSE EVENTS

All AEs and SAEs would be coded using the MedDRA version 10.1 dictionary. The integrated database would contain at least the following levels of coding: Lower Level Term (LLT), Preferred Term (PT), System Organ Class (SOC), Higher Level Term (HLT), and Higher Level Group Term (HLGT).

Studies where AE and SAE source files do not include MedDRA version 10.1 coding would be re-coded prior to inclusion to ensure consistency of coding across the integrated database. The individual study domain files would retain the study-specific coding version used, while the integrated SDTM database would reflect the single version coding of events, while retaining the original coding values in order that comparisons with the individual study report could be made.

The SUPPAE dataset contained both original MedDRA coding and the updated MedDRA coding for the integrated database so that original coding values could be checked against individual study reports.

PRIOR AND CONCOMITANT MEDICATIONS (INCLUDING PD MEDICATIONS)

Prior and concomitant medications would be assigned to a generic drug name using the World Health Organization Drug (WHODRUG) dictionary. Subsequently, the drug names would be matched to the respective Anatomical-Therapeutic-Chemical (ATC) classification system.

The same version of WHODRUG would be used throughout. The medications would be re-coded in a similar manner to the adverse events with the original coded value retained in the integrated SDTM database alongside the latest version allowing comparison with the individual study CSR.

CONCLUSION

In working on this sizeable project it allowed those working continuously on it to become more familiar with the intricacies of the SDTM implementation guide than they might have done just mapping a single study. Indeed, such was the enormity of some of the domains that different team members became expert in certain domains that would come in handy for future CDISC involvement.

This project also helped to highlight some of the shortcomings of the implementation guide in helping to define an ISS/ISE for certain domains. However, the guide and the whole concept of SDTM were extremely useful in enabling a standard to conform to. Doing the integration without using CDISC would have been a more challenging prospect.

In terms of the data integration strategy we wanted a consistency of analysis so that we could ensure that data manipulation had been conducted the same way in each study (e.g. that baseline flags had been derived in exactly the same way). However, would need to go back and check that we get the same results for the original study analysis conducted on the CDISC SDTM data as the original raw data. By mapping into SDTM first we were then going to re-derive all the analysis.

Timelines were a consideration with our approach as we would have to wait until we'd mapped at least some of the data into CDISC before we could start work on the derived datasets. With tight timelines, this was risky.

An alternative approach could have been to utilise derived datasets that already existed for the studies and if they were consistent then reconciliation could have been more direct while we mapped to SDTM in parallel to this, but we had 40 studies that did vary. Another approach we did consider was to completely independently map the raw data into SDTM in parallel to the derived datasets (which would be pooled from the original derived datasets), however, we felt that the advantages out-weighed the disadvantages

Overall I would say the biggest issues that we came across specific to the SDTM mapping were the inconsistent data structures and the creation of the diary domain (expand on this). I feel that in a project of this size that there was a great deal of necessary planning of the programming approach and studying of the implementation guide to the extent that presented parts of it to each other to get a better understanding. The learning curve for those that had never seen these standards was steep but well worth the effort considering the number of studies that would use SDTM data in future.

PhUSE 2008

REFERENCES

CDISC SDTM Implementation Guide (Version 3.1.1)

ACKNOWLEDGMENTS

I would like to thank Stephanie Noller for all her assistance and review with this paper

RECOMMENDED READING

CDISC SDTM Implementation Guide (Version 3.1.1)

CONTACT INFORMATION

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

Andrew Noller
Quintiles Ltd
Station House
Market Street
Bracknell
Berkshire / RG12 1HX
Work Phone: +44 (0)1344 708646
Fax: Not available at this time
Email: Andrew.noller@quintiles.com
Web: www.quintiles.com