# The application of SDTM in a
# disease (oncology)-oriented organization

Angelo Tinazzi, Alessandro Cattaneo, Enrica Paschetto, Sonia Colombini
SENDO-Tech S.r.l., Milan, Italy

## ABSTRACT

Applying CDISC[©] standards represents one of the major current challenge in the pharmaceutical industry community. Among the CDISC standards developed (and "stable") so far, the SDTM is probably the most crucial since it has an immediate impact on how to submit data to the FDA in support of marketing applications (although this is not yet mandatory).

Several approaches for and when implementing SDTM standards have been discussed in many public presentations (e.g. DIA, CDISC, PharmaSUG, PhUSE). After the first set of experiences and following the pilot CDISC/FDA project, we know at least that an organization is free to follow its own approach on how to implement SDTM; we can cite at least three approaches:

- build SDTM entirely in the Clinical Data Management System (pre);
- build SDTM entirely on the 'back-end' (post);
- build SDTM using an hybrid approach (hybrid).

Although the effort put by the CDISC developers in providing detailed guidelines (e.g. SDTM Model Implementation Guide Version 3.1.1) and examples, applying the proposed data-domain is not always an easy task, especially when you have to map data not supported by this standard (e.g. AE for adverse events, CM concomitant medications, VS vital signs, etc.). You may be in the position of solving one of the following issues:

- in which CDISC domain should I save my data?
- should I create a new ad-hoc domain?
- if I have to create a new ad-hoc domain, which 'general class' should I use?
- if I have a variable not mapped to the standard domain I have selected, should I create a new variable or should I put it in the SUPPQUAL dataset?

These are just examples of common issues we have encountered during the implementation of SDTM at SENDO we want to address and discuss in this paper. In particular we will focus the attention on the following aspects, by providing and discussing the solution we found:

- create an ad-hoc domain;
- map variables to SUPPQUAL datasets;
- define relationship.

Furthermore we will make a quick report of problems encountered and the deviations we had to apply during our SDTM implementation.

## INTRODUCTION

CDISC defines the models MOD [1] but it is not always clear how to go about implementing these models in your operational environment, although additional guidelines such as "The CDISC SDTM Implementation Guide" IG [2] add a lot of clarifications and suggestions; in addition, it does not tell you how to get from you "raw" data in your clinical data management system (CDMS) to the SDTM.

The aim of this paper is to report a concrete experience of SDTM application in a disease oriented organization and it assumes the reader has some basic knowledge of CDISC standards for which a lot of material can be found (see references section at the of the paper); therefore the paper will not provide any background information on CDISC/SDTM unless required by the examples we will report.

### SDTM IMPLEMENTATION: POINTS TO CONSIDER

Several approaches emerged so far in some of the company experiences presented [3] [4] [5] [6]:

- build SDTM entirely in the CDMS (<u>pre</u>): although this approach has the advantage to have "raw" data equivalent to your SDTM, and therefore no other data manipulation is required, forcing the CDMS to create SDTM structures may be too cumbersome and may cause problems with the operation of the CDMS which could reduce data quality;
- build SDTM entirely on the back-end (<u>post</u>): with this approach you work around the rigid (sometimes) constraints of the CDMS, by implementing the SDTM in other more flexible tools, such as SAS, where you can transform any proprietary CDMS; however with this approach additional resources are required (statistical programmer) to transform data and validate every single step in order to assure that there are no errors in translating from the "raw" CDMS data to SDTM;
- build SDTM using a hybrid approach (<u>hybrid</u>): with such approach you can anticipate some simple SDTM implementation in the CDMS, such as naming data tables the same as SDTM domains, using SDTM variables in the CDMS, and performing simple derivations (such as age) in the CDMS. Again we can take advantage of the flexibility of tools such as SAS© but again similarly to the "post" approach we will have the data-management team speaking a different "language" (semi-SDTM).

Moreover, if your organization is not new, that means software and internal solutions (e.g. standards, libraries) are already in place, in the choice of what approach to follow you have also to consider the following aspects:

<u>Your current CDMS main characteristics</u>
- metadata libraries support;
- flexibility in clinical database updates;
- flexibility in performing complex derivations;

<u>Your organization</u>
- your data-management team 'language';
- biostat experiences;
- existing data-management and analysis standard tools.

Overall we can say that the decision should be driven by the existing company standards regulating the entire clinical trial process from data collection to data analysis.

## OUR APPROACH TO SDTM IMPLEMENTATION

Due to the existence of a well defined structure for data-management, from standard CRF to standard database tables and screens, the heterogeneity among sponsor approaches and standards and because of the CDMS systems (3) currently in place, we decided to implement SDTM using a hybrid approach (see figure 1). In synthesis these are the steps we have followed.

In the CDMS:
- SDTM tables (standard domains) naming;
- SDTM variables naming, including derived variables derived in SAS;
- simple derivations;

SAS post-processing mapping:
- direct mapping;
- complete variables mapping (rename): for example STDTC→AESTDTC (start date);
- complete derivations;
- combine multiple CDMS variables (e.g. datepart variables);
- combine multiple CDMS datasets:
  - Horizontal vs Vertical when required (e.g. vital signs);
  - match 'temporary' non-standard modules (e.g. chemistry+hematology→LB module);
- standardize format values;
- apply new formats (e.g. ISO date format);
- creation/Update of Trial Design Datasets (see MOD 7.1 [Trial Design Datasets] Description of key aspects of the planned conduct of a clinical trial).

With this approach we are able to create an intermediate data-structure (SENDO-SDTM) for all studies, regardless of the type of CDMS used and sponsor standards; it is only when a sponsor plans to submit data to FDA that we apply complete <u>post-processing</u> in SAS (e.g. removal of non-required variables, uses of variable decode rather than code).

With SENDO-SDTM in place, we can then apply a set of standard tools that can be easily applied to any study, from data-management to analysis. We can for example apply standard data validation reports such as patient profiles or listings.

## TOPICAL CDISC SDTM IMPLEMENTATION TASKS

While in certain circumstances applying CDISC/SDTM standards mainly requires to check whether or not your data structure apply correctly, that means applying the correct module name, variable names and characteristics,

constraints (e.g. required non missing variables), there are other circumstances in which the standard models do not apply, either in the non-presence of a specific variable or the absence of a specific data-model and therefore you are for example required to understand the general classes and how you have to model them to your clinical data.
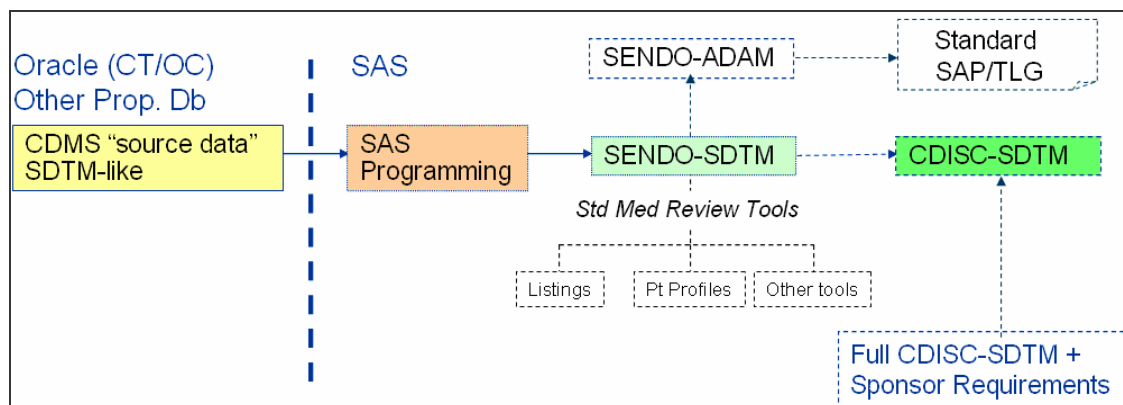


**Figure 1: SDTM Implementation – SENDO Approach**

**CREATING A NEW DOMAIN**
The IG (section 2.6 page 13) tells you how to create a new domain. These are in synthesis the steps you have to follow:
- check that none of the existing published domains will fit the need;
- use the existing models as prototype or choose the general observation class (how to determine where data belong - IG section 8.6 page 129);
- assign a 2–letters domain code: it is suggested but not required by CDISC to call your domain "X_" in order to avoid any overlap with any new models created by the CDISC-SDTM team in the future (check IG section 10.3.1 page 165 for list of "Reserved Domain Codes").

Understand the class your domain belongs to:
- Interventions: Treatments or Procedures intentionally administered either specified by study protocol or other treatments (Investigational Treatments, Concomitant Medications);
- Events: Occurrences or incidents independent of planned study evaluation occurring before or during the trial (Adverse Events, Medical History);
- Findings: Planned evaluations to address specific questions (Physical Evaluation, Laboratory).

For example in our organization we have created new models for mapping prior tumour history such as Prior Systemic Therapies (ST) and Prior Radiotherapies (PR), tumour Response Evaluation (CA), detailed follow-up (FP) and other add-hoc study modules (e.g. lymphoma-non-hodgkin assessment ad-hoc measurement);

Once you have verified that none of the existing modules can be applied and you have chosen the class domain your model belongs to, you start creating your model by adding the variables you need to map as described in figure 2.
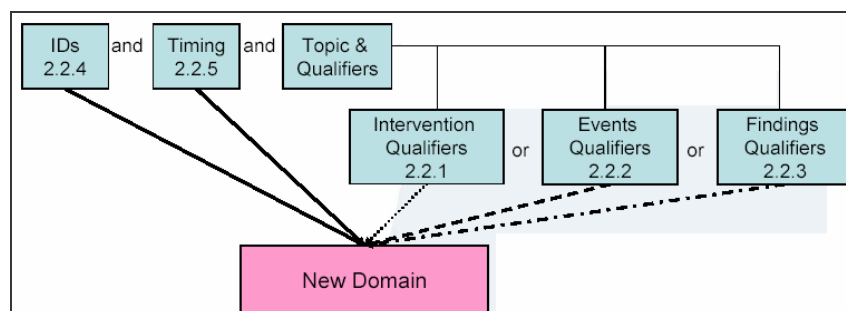


**Figure 2: Schema for adding variable to new domain**

All new domains should contain variables from the Identifier[1] (e.g. subject of the observation) and Timing[2] (e.g. visit

---

Definitions from IG section 2.1 page 9

description, start / end date) variable classes. Furthermore you need to create the Topic and Qualifier[3] variables from the specific general observation class that it was determined to fit, so if you have determined that your new domain belongs to Findings domain, you cannot use Qualifier variables that were defined for the Interventions or Events domain classes.

In creating a new variable, the schema reported in the IG (section 10.4 page 171) can be useful. For example if in your new domain you have to map a variable containing a 'dosage' the suffix (fragment) DOSE or DOS should be used, so that your new variable will be –DOSE, where "--" is replaced by your new domain name.

**AN APPLICATION OF NEW DOMAIN: THE TUMOR ASSESSMENT (CA)**

The aim of the tumour assessment is to evaluate the tumour response to the treatment applied by measuring the tumour lesions at study entry and during the study at regular intervals (usually every two months/visits). Currently, the standards for evaluating tumour response in solid tumours, are those defined with the RECIST criteria [7].

With this criteria, each lesion, as showed in figure 3, is measured (usually a single measure in "mm") using a specific "Method" (e.g. spiral ct-scan) on a specific day (assessment date); the lesion is identified by a sequence nr. (lesion nr.) and a description (and the organ the lesion belongs to). Finally, the lesions are grouped in three main sections: Target Lesions, Non-Target Lesions and New Lesions (those not present at study entry and detected after the enrolment in the study).



**Figure 3: Creation of new domain (tumor assessment)**

After an evaluation of the existing SDTM models we decided none of those were pertinent to our purpose and therefore we have proceeded with the creation of a new model. The Findings class was used and the short code "CA" assigned.

IDs: In addition to the --SEQ variable (CASEQ), we add the additional identifier --SPID[4] (CASPID) in order to store the lesion sequence number so that lesions can be linked across visits.

Timing: The standard timing variables were used (visit identifier plus the assessment date, CADTC).

---

[1] Variables, which identify the study, subject of the observation, the domain, and the sequence number of the record

[2] Variables, which describe the timing of the observation

[3] Topic: variables, which specify the focus of the observation (such as the name of the lab test); Qualifier: variables, which include additional illustrative text, or numeric values that describe the results or additional traits of the observation (such as units or descriptive adjectives)

[4] MOD section 2.2.4: Optional Sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit line identifier on a Concomitant Medications page

Finding Qualifiers: the --CAT variable (CACAT) was used to classify the type of lesion, the --TEST (CATEST) and --SCAT[5] (CASCAT) were used to describe the lesions (description plus organ) and --ORRES (CAORRES) and --ORRESU (CAORRESU) were used to store the results (measure) of each evaluation.

**USE OF SUPPQUAL DATASETS**
According to IG, it is not allowed to add new variables not present in the general-observation-class models (IG section 8.4 page 125). In these circumstances, the Supplemental Qualifiers (SUPPQUAL) special purpose dataset should be used to capture non-standard variables and their association to parent records in domains. In some cases, for example if the SUPPQUAL dataset becomes too large or to capture additional attributions, it is allowed to create a separate supplemental qualifier dataset (SUPP--) for each submitted domain that has supplemental qualifiers.
The EX standard SDTM domain model, a model to record the details of subject's exposure to protocol-specified study treatment (IG section 6.1.2.1 page 47) that belongs to the Interventions Class, has been used to collect details of our study treatments (usually a chemotherapy). Details such as period (scheduled day of treatment) and dosage, together with standard attributes, were easily mapped. However, the variable collecting "Reason for Dose Adjustment" (EXADJ) did not allow to collect multiple / concomitant reasons for multiple / concomitant type of change (see figure 4 where the patient reduced the dosage and delayed the treatment start). We therefore have decided to use a separate domain supplemental qualifier (SUPPEX) collecting each instance of 'treatment modification'.



**Figure 4: Use of a SUPP-- dataset to integrate information in EX module**

The SUPPQUAL dataset is one of the special purpose datasets for representing data relationships (see IG section 8 page 120). With this set of datasets you can represent the relationship among datasets or records (between separate domains or between domains) so that the sponsor can represent data that do not entirely fit within standard domain models[6]. All relationships make use of standard domain identifiers (STUDYID, DOMAIN and USUBJID). The variables IDVAR and IDVARVAL, allow to identify the parent record's. In addition QNAM (the name of the Qualifier Variable), QLABEL (the label for the Qualifier Variable) and QVAL (the actual data value for each record).
In our case the SUPPEX dataset was linked to the EX dataset through the EXSEQ variable. As showed in figure 4, at VISITNUM 2 for USUBJID 01002, there have been two treatment modifications:
- a dose reduction (QNAM=RED, QLABEL=Dose Reduced) due to an AE (QVAL=1);
- a treatment delay (QNAM=DELAY, QLABEL=Treatment Delayed) due to an AE (QVAL=1);

Note that with this structure only circumstances in which there has been a modification, are represented; this means if a for subject it was never required to change the treatment, then no-records will be present in the SUPPQUAL dataset.

**COMMON ISSUES WE ENCOUNTERED IN MAPPING SDTM**
We finally want to make a sort of quick report of the main issues we have encountered in our organization in mapping

---

[5] MOD section 2.2.3: Used to define a further categorization level for a group of related records
[6] An important aspect of the SDTM is that it can be used to re-engineer the entire data-collection process, so sponsor applying CDISC not just in the final submission step has the need to map all collected variables

the CDISC/SDTM structures:

General issues

- CDMS modules (CRF section) that map to one corresponding SDTM domain (e.g. laboratory and normal ranges);
- SDTM variables contain decoded text rather than code (and format/codelist to be applied);
- "other …. Specify" → in SUPPQUAL dataset;
- sometimes it can be considered not analysis-oriented, but this is sometimes due to programmers attitude (e.g. vertical vs horizontal structure);
- it is unclear where endpoints such death and survival should be collected (DS? ad-hoc module? Findings or Events Class?).

Deviations

- sometimes we have created new variables rather than using SUPPQUAL;
- --DY variable = 0 (see IG 4.1.4.4 "There is no study day 0");
- common terminology not fully-integrated;
- use of code rather than decode;
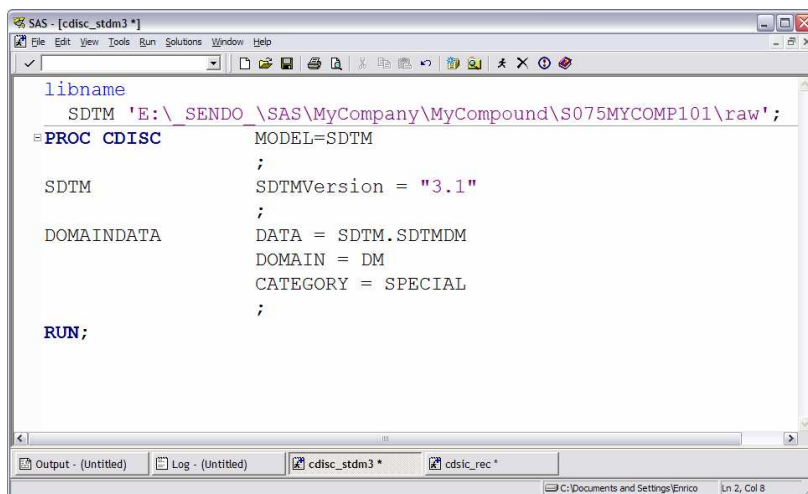- assigned dose-levels in "Trial Arms" table (see MOD 3.2.2 "Planned Arms in the trial").

Module-Specific issues (AE)

- changes in AE-condition in the same visit (e.g. severity grade – IG 3.1.2 "Findings CF module to be applied");
- multiple action taken and drug relationship for drug-combination trials;
- "Any AE occurred" is not allowed (see IG "When adverse events are collected with the recording of free text, some sponsors may enter a record into the data management system to indicate 'no adverse events,' for a specific subject. For these subjects, do not include a record to indicate that there were no events unless AEOCCUR is being used to represent records for non-occurring events. *Otherwise, records should be included in the submission dataset only for adverse events that have actually occurred*").

## VALIDATING SDTM IMPLEMENTATION

Even if you build a % compliant SDTM clinical database, how do you know it is a valid SDTM? You need for example to check that mandatory variables are present or that standard variables have the required characteristics. For "simple" control tasks, SDTM 1.1 data validation is available in SAS 9.1.3 (service pack 4 available for download at ftp.sas.com/techsup/download/hotfix/hotfix.html) trough the CDISC procedure [8]; the following options are available for SDTM validation:

- 15 SDTM modules supported;
- various checks on models / variables attributes (e.g. variables not defined in the domain);
- various checks on models / variables contents (e.g. e.g. missing values in required variables).



**Figure 4: Validate SDTM-DM domain implementation**

Figure 4 shows an example of proc CDISC use in validating a DM dataset and figure 5 shows the outcome of the validation (log file) where you can find the following issues:

- Expected variable warning: Variable AGEU is expected in this domain(DM), but is not present (1);
- Permissable variable note: Variable DMDY is permitted in this domain (DM), but is not present (2);

- ▪ <u>Data Error Incorrect ISO time-units</u>: Variable REFNDTC has incorrect content in observation 4, Incorrect data is 10-12-2008; A time or date component was outside expected range (<u>3</u>).

However, when you apply a "post" or a "hybrid" model, checking that the structure of the datasets are compliant with the SDTM models as described above is not enough. You need to add additional checks in principle verifying that your transformed data did not lose any importance piece (e.g. an entire study visit).
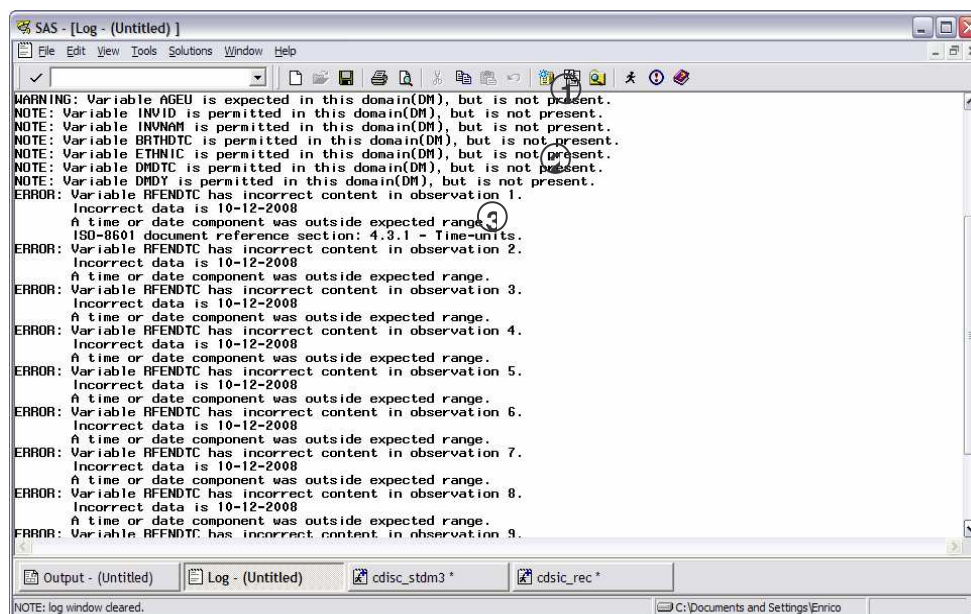


**Figure 5: Results from proc CDISC**

## CONCLUSIONS

The implementation of CDISC standards requires a lot of resources at least in the set-up step. However in the long-term, their implementation makes the process of clinical trials, from data-management to data-analysis, a much more efficient process (e.g. re-usability of tools, such as data-listings and patient profiles for clinical data review) .

SDTM is not just a model for submission, but it can be used to review or to create new company standards. We have demonstrated that SDTM models and more in general the SDTM approach can be adopted also by small organization such as ours, where we have to face different clients each one still adopting different methodologies (e.g. standard naming conventions); moreover, we suggest and we hope that standards such as the CDISC/SDTM will be soon considered also by the academic organisations where data(knowledge)-sharing has been the "leit-motiv" for ages (e.g. meta-analysis).

The CDISC SDS[7] Team encourages initiatives for developing SDTM models for specific therapeutic area, for example the SDS Oncology sub-team. The aim of such a team is to develop Oncology Implementation Guideline (IG) that eventually may be part of the SDTM IG as an appendix. For more information you can reach the team in the internet at the following address http://tech.groups.yahoo.com/group/DV_CDISC_UN.

We have of course encountered some problems and we want therefore to highlight some SDTM criticisms and alerts.

### SDTM CRITICISMS

- ▪ it is still SAS V5 Transport dataset dependence, which means there is a limitation on variable name length (8 characters) and on character variable length (200 characters);
- ▪ SDTM is not always suitable for data-entry and data-validation.

### SDTM ALERTS

- ▪ as already stated, applying SDTM is not just a variable renaming activity; it requires much more efforts and for this reason we suggest in the initial steps to use resources for training of internal personnel directly or indirectly involved (biostat group for example) ;
- ▪ we found the "hybrid" model a good compromise, as it allows small violations and the management of

---

[7] Submission Data Standards

**PhUSE 2008**

'diversity' among clients;
- try to be up-to-dated on model evolution and If possible be involved in the draft documents review process;
- apart from the guideline developed so far, you can find on the web a lot of additional resources; we found vey useful the CDISC public forum (figure 6), where we found a lot of solutions to our doubts/problems.



**Figure 6: CDISC Public Forum (www.cdisc.org/discussions/index.html)**

**REFERENCES**
1. Study Data Tabulation Model - Version 1.1[8]
2. CDISC SDTM Implementation Guide – Version 3.1.1[8]
3. Fisher I., Rahman M., Thompson D. All roads lead to SDTM? Which one shall we take? PhUSE 2007, Paper RA08
4. Boudraa Y. Creation of Submission and Analysis Datasets by following SDTM and ADAM requirements. PhUSE 2007, Paper RA05
5. Shostak J. Implementation of the CDSIC SDTM at the Duke Clinical Research Institute. PharmaSUG 2005, Paper FC01
6. Wood F., Guinter T. Evolution and Implementation of the CDISC Study Data Tabulation Model (SDTM). PharmaSUG 2006, Paper FC08
7. Therass P. et al. New Guidelines to Evaluate the Response to Treatment in Solid Tumors. J Natl Cancer Inst 2000; 92:205-16
8. SAS, The CDISC Procedure for SAS Software at http://www.cdisc.org/discussions/discussions.html; 2004

**CONTACT INFORMATION**
Your comments and questions are valued and encouraged. Contact the author at:
Angelo Tinazzi
SENDO Tech S.r.l.
Medical Informatics and Biometry Unit
Via Visconti di Modrone 12
Milan / 20122
Work Phone: 0039 02 76420426       Fax: 0039 02 76017484
Email: tinazzia@sendo-org.it       Web: http://www.sendofoundation.org
SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.

---

[8] SDTM Model 1.2 and SDTM Implementation Guide 3.1.2 are currently under development