Implementing CDISC at Boehringer Ingelheim

Michael Knoessl, Boehringer Ingelheim, Ingelheim, Germany
Peter Leister, IBM, Hamburg, Germany

ABSTRACT

The Study Data Tabulation Model (SDTM), as designed by the Clinical Data Interchange Standards Consortium (CDISC) has evolved into a global standard data structure for submitting clinical data to the U.S. Food and Drug Administration (FDA). Concurrently, it also became an accepted interchange and collaboration standard across the pharmaceutical industry, suppliers and software vendors. Boehringer Ingelheim (BI) has taken the decision to design a corporate wide clinical data model for building efficient pooled databases (PDB) on either project or substance level. This medical data model will be founded on the CDISC data standards. This paper illustrates the Boehringer Ingelheim approach to implementing and using the CDISC standards.

INTRODUCTION

The Study Data Tabulation Model (SDTM), as designed by the Clinical Data Interchange Standards Consortium (CDISC) has evolved into a global standard data structure for submitting clinical data to the U.S. Food and Drug Administration (FDA). Concurrently, it also became an accepted interchange and collaboration standard across the pharmaceutical industry, including suppliers and software vendors. With the advent of the standardized Analysis Dataset Model (ADaM) and the Clinical Data Acquisition Standards Harmonization (CDASH), a common format for clinical data from the acquisition all the way down to reporting becomes tangible in the near future. As of 2004-07-21, the FDA has accepted the SDTM structure as a submission standard format for case report tabulations (CRT).

Currently, the pharmaceutical branch is adopting the CDISC standards not only at the back end when preparing the deliverables for submission, but many companies rather adopt them as an integral part of the business processes following the flow of clinical data from data entry, via a data base management system (DBMS), analysis and reporting until submission. Driven by internal and external demands CRO's and laboratories seem to move faster - while the industry (carrying vast libraries of historic data, programs and systems) takes more time to thoroughly evaluate the path to go and adopts slower. Though very appealing long-term benefits can be gained from a globally harmonized structure for clinical data, the transition from currently used, company specific standards to the CDISC standards is a major transitional effort.

The long-term target is to fully exploit data across systems and company boundaries, across formats, semantics and data currencies. So what we strive to achieve is full semantic-interoperability. A submission package is compiled with the results of more than one clinical trial. Those trials are tabulated and analyzed in parallel and/or collectively, i.e. from a pooled database. Beyond the direct usage for submission purposes, pooled databases form the source for meta-analyses and for an increasing number of company internal requests (such as ad hoc analyses) from various functions, including statisticians, physicians, marketing and pricing.

Boehringer Ingelheim (BI) has taken the decision to design a corporate wide clinical data model that will be founded on the principles of the CDISC data standards, mainly the SDTM, ADaM, ODM, and Controlled Terminology standards. One major impulse for this decision was the increasing demand for building efficient pooled databases (PDB) on either project or substance level.

APPROACH AND PROJECT MANAGEMENT

Adopting the global CDISC standard as an integral part of the in-house clinical data process (operational data structure), and not just at the back-end when preparing the submission deliverables, is a major change to the organization, a challenge in itself. Not only the structure in which clinical data are kept, analyzed and submitted is affected, it concomitantly has a significant impact on processes, systems and tool of various functions involved in working with clinical data; on the long-term, it may even impact CRF design and data entry.

Given the complexity of the CDISC transit, assuring a successful, sustained and efficient result is not a low hanging fruit. Achieving this requires a strong management mandate, a well designed result, a clearly planned transition and an adequate management.

The ultimate target at BI is the creation of a corporate wide, harmonized clinical data model that is effectual for single clinical trials as well as pooled databases (PDB). A clinical data model fundamentally needs to support in-house business requirements and to comply with regulatory requirements.

BI is involving the full breadth of functions that deal with clinical data at a global scale. To manage the beast the initiative was cut into four main slices:

1) Assessment phase: Assessing the current clinical data flow, related aspects and the magnitude of change at BI; initial definition of an implementation approach.
2) Design Example phase: Evaluating the SDTM structure for being suitable as an operational data structure.
3) Design phase: Complete design of a clinical data model (based on the principles of the CDISC data standards) and of the new working processes.
4) Transition phase: Implementation of the clinical data model, business- & IT-processes, and system changes.

Currently, we are moving into the full design (third phase) and have successfully completed the first two phases. The following sections give a description of these two phases, the results obtained, and some lessons learned.

ASSESSMENT PHASE

The target of the assessment phase was to develop an effective implementation scenario and to assess its impact on the organization. To achieve this target, we drafted various implementation scenarios and performed assessments (with regards to risks, benefits, impact) of these scenarios against:
  - processes and clinical data flow,
  - systems and tools,
  - external data processes,
  - compliance (functions, SOP's, validation, training),
  - structures and standards (standards, O*C GLIB, macros, automation add-ons),
  - resources & commitment (other initiatives, transition, effort, stakeholders),
as all of these criteria are currently being lived and effective at BI.

Based on the assessment results obtained, in an iterative manner the drafted implementation scenarios were further developed, optimized, weighed and fine-tuned until we finally agreed on and gained management approval for one of the scenarios.

DESIGN EXAMPLE PHASE

The design example phase primarily aimed at evaluating the CDISC SDTM structure for its operational quality. An operational data structure should efficiently support at least the following:
  - performing data quality checks
  - creation of analysis data sets
  - performing ad hoc statistical analysis
  - generation of pooled data bases

During the design example phase, a selected set of domains, namely the Adverse Event, the Exposure and the Efficacy domains were used. On them a mapping and a transformation of clinical data from our in-house O*C view standard structure to the plain CDISC SDTM structure as well as to a modified / enhanced SDTM structure (regarded to as SDTM plus or SDTM+) was performed. In addition, mapping and transformation of data from both the plain SDTM as well as the enhanced SDTM+ structure to analysis datasets (ADS) were done.

Selection of the domains followed the desire to dive into the more advanced areas of SDTM and to make best use of the design example.

To perform the mappings and transformations we used clinical data from two completed, recently submitted trials. Since already submitted, plain SDTM as well as analysis datasets were available for both legacy trials. All mappings and transformations (i.e. O*C views to plain SDTM to ADS and O*C views to SDTM+ to ADS) were first notionally depicted and secondly tested for feasibility by SAS® programming. It was essentially the requirements imposed by analysis and implicit in the structure and content of the ADS as well as the SAS® programming feasibility, that identified so called "plus-elements", which in return went into the example design of the SDTM+ data structure. Both results and lessons learned taken from this design example gave interesting insights into SDTM.

DESIGN EXAMPLE RESULTS

We tagged the findings in three main areas, where plain CDISC SDTM needs be enhanced in order to promote it to an operational data structure:
  a) mapping from existing data structure to enhanced SDTM+
  b) additional elements (e.g. additional variables) to enhance the SDTM structure
  c) key principles to facilitate the efficient implementation of CDISC standards and to support pooling of clinical data

Mapping issues

Mapping of clinical data from a source to a target structure (where both can consist of one or more datasets) essentially addresses the aspects of selection and joining of data. This is valid for both data records as well as data variables. Mapping data from our O*C database management system, i.e. from an O*C view structure to the SDTM/SDTM+ structures for the adverse event, exposure and efficacy domains used in the design example, elucidated that there are two types of complexities:

For domains, where the clinical content is clearly defined and rather confined (e.g. adverse event) the degree of selecting and merging was rather low. It mainly referred to the parent domain (AE) / supplemental qualifier domain (SUPPAE) issue. However, for the exposure data, a number of source O*C views and a number of target SDTM

domains (e.g. EX, TA, TE, SE) were identified, that capture data relevant for exposure information. A similarly high degree of complex splitting, selecting and joining of data was found for the efficacy domains. Complexity to mapping is added to efficacy by the fact, that there are no SDTM domains suggested in the SDTM Implementation Guide (IG) for efficacy data. In addition, some BI specific derived variables were identified, which the SDTM structure does not distinctly accommodate for.

The design example essentially showed, that there is no simple 1:1 relation between the BI O*C view and the SDTM domains and that SDTM does not clearly provide place for all clinical data stored in the BI DBMS.

<u>Lessons learned</u>: Despite the definitions of domains in the SDTM and the SDTM IG, while not being all-embracing and comprehensive, and due to the complexity of mapping from the current BI O*C view structure to an to-be-designed SDTM+ structure, when designing a clinical data model the content of clinical information need be distinctly and unambiguously defined for each domain. Successively, this definition will guide the definition of the mapping rules.

SDTM Plus Elements

While mapping and transforming the clinical data from the BI O*C-view structure to the plain SDTM and further to analysis datasets led to the identification of particular, so called "plus elements", which adding to the plain SDTM structure would beneficially enhance this structure towards operability.

- **Data type:** SDTM variables, irrespective of their clinical content, largely are character variables. Variables that contain data that are analyzed numerically should be stored in numerical variables – either in addition to or substituting their SDTM character counterpart. Keeping those data in numeric type would alleviate from back and forth num-char conversion. For example: SDTM --ORRES, SDTM+ "--ORRESN"

- **Date/Time:** All date/time information is given in ISO8601 format as character value. For analysis purposes, numeric date/times are required. Again, for circumventing frequent num-char conversions, storing date/time values as character and numeric values in parallel would be beneficial.
  Moreover, ISO8601 is very suitable for date/time reporting, since it provides a standard way for even tabulating incomplete date/time values. Effective imputation rules for incompletely collected date/time data, which should be consistent at least within a domain, would grow to a prerequisite for keeping complete date/time in numeric type as well as for pooling clinical trials. For example, the imputation rules for e.g. adverse event, concomitant medication, and medical history may be particular, but should be consistent within each domain and across trials. The imputation rules per date/time variable need additionally be stored as plus elements, according to ADaM. So, the reported date/time, the imputed date/time and the imputation rule per date/time information should be stored.

- **Coded variables:** Due to SDTM being character based, of information having a code list attached, only the decodes are stored in plain SDTM. Again, creation of ADS and analysis itself is greatly supported by keeping the code in a separate variable in addition to the decode for each categorical variable. Moreover, in a normalized structure (e.g. the findings class) clinical information and variables are not related in a 1:1 fashion, i.e. more than one tests (in general: any categories) are defined in one variable (e.g. --TESTCD) and their respective results are collected in yet another variable (e.g. --ORRES). Thus, when transposing these data from a rather de-normalized structure (e.g. one test per variable with the results as variable values; and each variable with its own code list/format attached), the values of the normalized variables might become composed of codes or decodes defined in more than one code lists. In order to enable tracing back data values to the original code lists, an additional variable, having the code list names as variable values, would be required. So, a code variable (num), a decode variable (char) and a variable containing the name of the associated SAS®-formats per code/decode value (i.e. per record), should be stored.

- **Origin:** The CRF origin of clinical data stored in SDTM is a meta-information captured in define.xml, but is not included in SDTM itself. When programming on clinical data, for practical reasons, it would be beneficial to store the origin (e.g. CRF name, or O*C Data Collection Module name), where a variable or data record originated from, together with the respective data. This would make the connection between data and CRF readily available.

- **Missing SDTM definitions:** Some clinical information that are contained in the BI O*C views have been identified, be it collected or derived information, for which SDTM does not hold definitions for. Since these information is topic related rather than record qualifying, they are not suited for being sent into the respective supplemental qualifier datasets. These data are required for analysis but are not subject to being submitted in tabulation datasets. Additional variables in an enhanced SDTM structure would be required for keeping these clinical data available for analysis.

- **Extension/Follow-up trial information:** Extension trials (follow-up trials) are set up as separate, individual trials in O*C, each having its own trial number and an own subject number assignment applies. In order to match clinical data for patients that continue from a previous trial to the respective extension trial, the relatedness between previous to extension trial and subject numbers need be constructed. This could be done in an own relation (i.e. a separate dataset) or possibly could be incorporated in the enhanced subject characteristics (SC) domain. Including this relation in the data model will facilitate pooling of trials and creation of a unique subject identifier for a given submission.

- **Supplemental Qualifier domains:** The SDTM allows creation of supplemental qualifier domains (SUPPQUAL) for each parent domain for clinical data that do not fit into the parent domains as specified in the SDTM IG. Incorporating these clinical data into modified, enhanced parent domains would reduce the scattering of data with

a topic specific commonality across datasets, and thus would reduce the effort of joining data that clinically belong together when using and analyzing these data. The enhanced parent domain should include a means of indicating variables and records that are destined to the respective SUPPQUAL, to facilitate mapping of data from the enhanced SDTM to the plain SDTM + SUPPQUAL at the time when creating the submission deliverables.

- **Key variables:** The SDTM identifier variables and additional key variable are covered in the next section.

Lessons learned: SDTM as defined requires additional variables and data records and other modifications to render it operational, i.e. to create an SDTM based data structure that is suitable for addressing data cleaning, creation of ADS, data analysis, trial data pooling, and eventually mapping to SDTM as defined. These modifications at least encompass additional variables for e.g. numeric variables, code/decode information, for clinical information not acknowledged by the SDTM standard. For these additional variables a consistent naming convention need be defined.

The "plus elements" identified so far, apply to particular clinical information or particular types of variables.

Key Principles

The mapping and transformation example performed, in addition to the "plus elements" described above, also yielded, as we called it, "key principles", which apply to the fundamental components of a data structure that is intended to enable operability, consistency and data pooling for single trials and across trials. These key principles pertain across all domains.

- **Key variables:** The SDTM defines a total of seven identifier variables (STUDYID, DOMAIN, USUBJID, --SEQ, --GRPID, --REFID, --SPID). These identifiers provide uniqueness of records within a domain, and, but only, allow relating data between a parent domain to the respective supplemental qualifier domain (SUPPQUAL), to the comments domain (CO) and to the RELREC domain. They, however, do not allow relating data across parent domains – at least not without major effort. Aside STUDYID, DOMAIN and USUBJID, the SDTM identifiers are surrogate keys that do not convey a clinically meaningful information. They primarily allow unique identification of records, but not unique identification of clinical information in a pragmatic way. Feasible selection of clinical data and merging clinical data across domains is provided with natural keys, which are key variables that convey existing attributes of an object about which data are collectively stored together in a dataset. For example, the patient number, visit number, and the laboratory parameter code would suit as natural keys in a simple laboratory dataset, and they do convey clinically meaningful information. A major enhancement towards making the CDISC data structure operational would be the definition of a key variable concept, which is a set of key variables together with a controlled terminology for valid key variable values. This set of key variables need be defined consistently across the datasets and across clinical trials. A defined set of key variables and clinically meaningful key variable values will render the unequivocal identification of data (selection) and the unequivocal combination of data (merging) possible. They thus pragmatically would facilitate the two major functions of keys/identifiers: Selection of data and joining (merging) of data. SDTM provides a number of qualifier variables that can be modified to promote them to natural keys. This set of key variables need not be defined completely extra to the SDTM variable definitions, rather plain SDTM does offer a number of variables that, with some modification, could be promoted to natural keys. For example: The findings class variable --LOC (Location Used for the Measurement) could be promoted by renaming it "LOC" for the decode, creating a "LOCCD" variable for a short location code and defining a controlled terminology for locations (for which currently no CDISC terminology is in production). VISITNUM (Visit Number), for example, could be used as a natural key as such.

- **Controlled Terminology:** The majority of SDTM variables of the interventions, events, and findings class are defined such as to expect a discrete set of values (controlled terminology, CT), being either sponsor defined or taken from an external, published source. The current CDISC controlled terminology in production encompasses 42 code lists; it thus does not provide sufficient cover yet. At BI, for current and legacy trials, over the years, a number of CTs have been created and were or are still in use. The existing CTs would need to be compiled, consolidated and assessed for their compatibility with the CDISC SDTM variables and the CDISC code lists in production. To some extent, new CTs need be defined, such that there is one and only one code list for each SDTM variable, for which a CT is expected. In addition, for each CT the area of application is to be defined; a code list to be applied to all trials across the whole company, to all trials of a particular substance, a particular indication, a particular project or, in the case, even for only a single trial. Here the aim should be to define CTs on a level as high as possible, i.e. on a corporate or a substance level, to facilitate pooling of trial data. The defined CTs surely should be based on the ones existing in a company. Nevertheless, mapping from existing CTs to SDTM compliant CTs will need to be performed. The magnitude of CT mapping is expected to be lower for sponsor defined CTs, and higher for mapping to the CDISC terminology.

- **Transformation & Mapping Rules**: Mapping clinical data from a non-CDISC data structure to either the plain SDTM or an enhanced SDTM+ data structure covers at least the following topics: mapping of variables, mapping of variable values (controlled terminology; see above "Controlled Terminology"), transformation of variable values (imputation rules, derivation rules), and mapping of data to parent domains plus supplemental qualifier domains plus the comments domain (see above " Supplemental Qualifier domains ").

Lessons learned: Enhancing the CDISC SDTM structure towards an operational data model that facilitates data evaluation and analysis not only means addition of variables to particular domains. It also covers modifications that need be introduced consistently across the SDTM domains. In addition, a subsequent use of those key principles will

establish a higher level of standardization, enhance analysis, increase the re-use of standard programs and facilitate the compilation of the submission package (incl. the define.XML). An interesting lessons learned for the last point was that having the right structure defined and the key principles applied is as important as having established effective and well designed processes.

Conclusion

Many pharmaceutical companies currently are adopting the CDISC standards as an integral part of the business processes of the flow of clinical data. Boehringer Ingelheim (BI) has initiated an initiative to design a corporate wide clinical data model for building efficient pooled databases (PDB), which will be founded on the principles of the CDISC data standards. The assessment phase helped us to understand the magnitude of change this initiative has with respect to our business processes, our systems & tools, the functions and internal regulations that are affected. It helped us to involve the right resources and functions early in the process. And it helped to us to well define and agree the clear target - as well as the approach to reach the same.

The design example phase, gave us cognition of aspects, where the CDISC SDTM standard need be modified and enhanced regarding mapping and transformation and additional elements. The results of these two project phases furnished us with a solid base for the next step, the design of an operational clinical data model.

ACKNOWLEDGMENTS

CONTACT INFORMATION

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

Author Name:	Dr. Michael Knoessl
Company:	Boehringer Ingelheim Pharma GmbH & Co. KG
Address:	Binger Strasse 173
City / Postcode:	55216 Ingelheim am Rhein
Work Phone:	+49-6132-77-97055
Fax:	+49-6132-72-97055
Email:	michael.knoessl@boehringer-ingelheim.com
Web:	www.boehringer-ingelheim.com

Brand and product names are trademarks of their respective companies.