

Managing your metadata efficiently - a structured way to organise and frontload your analysis and submission data

Kirsten Walther Langendorf, Novo Nordisk A/S, Copenhagen, Denmark
Mikkel Traun, Novo Nordisk A/S, Copenhagen, Denmark

ABSTRACT

Novo Nordisk has implemented a Metadata Management Application (MMA) together with NNIT as part of our Clinical Data Warehouse System (CDW). This application is used by the statistical programmers and statisticians to define the planned activities for a trial. This information includes protocol and statistical information that is being used during data review, reporting and eventually for defining submission data. It helps the users structure the collection of data, structures the statistical decisions taken for the different assessments, and facilitates the ability to 'google' your trial descriptions on a detailed level. By entering trial metadata into a repository it enables structuring and automation in all process steps: Trial setup, trial conduct, trial reporting and creation of registration application data. The metadata not only covers the trial metadata required by CDISC but is extended to facilitate the full work process and is designed to support use of standardised terminology. A prerequisite for managing the use of metadata efficiently is a well defined and managed repository for definition of the metadata. We call it a clinical metadata repository.

INTRODUCTION

The benefits that we strive for with the use of the Clinical Data Warehouse system are:



A key element in obtaining these benefits is standardisation and reuse. Optimising your analysis effort requires reuse of programs and hence building up a well-tested set of macros that provides you with output of good quality. The goals of your programming are

1. Build confidence that results and conclusion about safety and efficacy are complete and correct
2. Ensure high quality in trial specification, collected trial data and statistical programming code

One of the prerequisites for obtaining quality in the programs is to understand the requirements carefully and to understand your data. Understanding and managing our metadata is an efficient step in that direction:

- By entering our trial metadata into a database forces you to extract 'requirements' from the protocol and analysis plans into an operational format (no paper but datasets).

PhUSE 2009

- It makes you think carefully about the trial design and what you want to conclude upon before you start looking at data and programming.
- Using metadata in the programming minimises error in code and helps you identify observed observations that falls outside the requirements, e.g. outliers, missing observations.

Once you have built up good quality macros to support your analysis you 'only' have to focus on the 'requirements for the trial', i.e. the trial metadata.

Not only can trial metadata be used for optimising trial reporting as described above, but also provide an overview of the trial data for a full clinical project. This is needed when planning for a registration application or when exploiting your data for new hypotheses.

Trial metadata is often only considered in the context of the trial, but in many cases the trial data will and can be used in other contexts, such as a registration application. To ensure that data can be pooled without too much effort you also need to control the trial metadata and how it is used. This is done via the clinical metadata.

This paper gives a brief overview of how Novo Nordisk A/S uses metadata in the clinical development area, what we define as clinical and trial metadata and how we handle the data in our systems. A prerequisite for obtaining the benefits of the trial metadata is standardised data and a way to manage them. This is controlled via Clinical Metadata and a supporting process. The paper will focus on the clinical metadata; exemplify how it is implemented and describe the process for maintaining them.

Firstly, we describe how we define the metadata, the distinction between clinical and trial metadata and the usage of the metadata in the business process. The implementation of the clinical metadata in the metadata management application is exemplified and finally the process on how to manage the metadata is described. Since this article focuses on the clinical metadata only a minor part of the application will be shown.

NOVO NORDISK'S DEFINITION OF METADATA

The concept of metadata at Novo Nordisk A/S is based on the CDISC SDTM trial design but is extended with more information in order to support our internal business processes. We operate with two types of metadata: Clinical Metadata and Trial Metadata see Figure 1.

Metadata in CDW

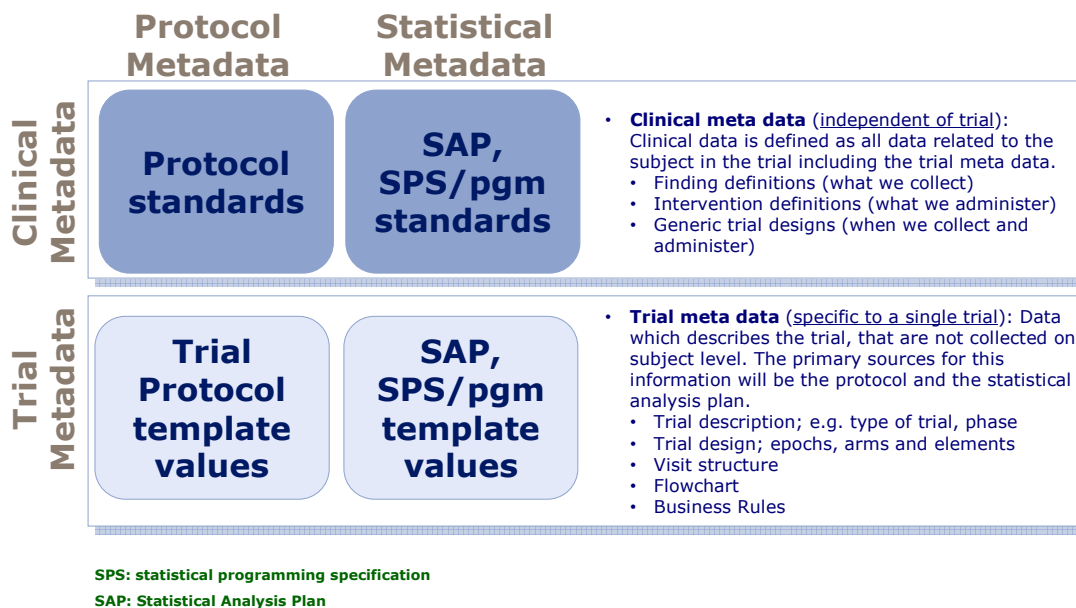


Figure 1 Types of Meta in CDW

CLINICAL METADATA

The Clinical Metadata is independent of the trial and is the definitions of templates for trial designs (generic trial design) and definitions of all the data we collect, e.g. topic code, labels, units, SAS® display format. These definitions are used to ensure consistency in the data that we collect, and to ensure consistency in the trial metadata. But it is

also more than controlled terminology see Figure 2. The clinical metadata describes relationship between the codes in the repository, traceability to source data (internal standard codes) and submitted data (CDISC codes) and additional attributes such as relation ship to other code lists.

What is Clinical Metadata?

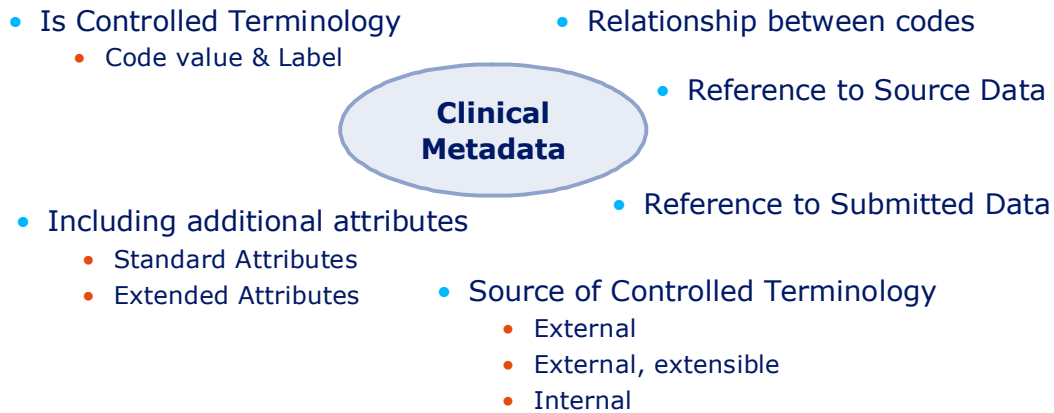


Figure 2 Definition of Clinical Metadata at Novo Nordisk A/S

Storing the clinical metadata in a repository enables you to standardise the terminology used across your trials and manage the terminology used in the different trials. In the pharmaceutical industry, trial data exists in many years and data and standards evolve over time, so the systems used to handle and store this type of data needs to be able to handle this and there need to be a business process in place.

Standardisation is a prerequisite for leveraging your clinical data across, e.g. for registration applications and cross-trial analyses without spending considerable amount of man hours to integrate data, see Figure 3. Also, it helps your organisation to compare terminology used in different clinical projects and to standardise, if needed.

Key changes

– Standardised terminology and data storage

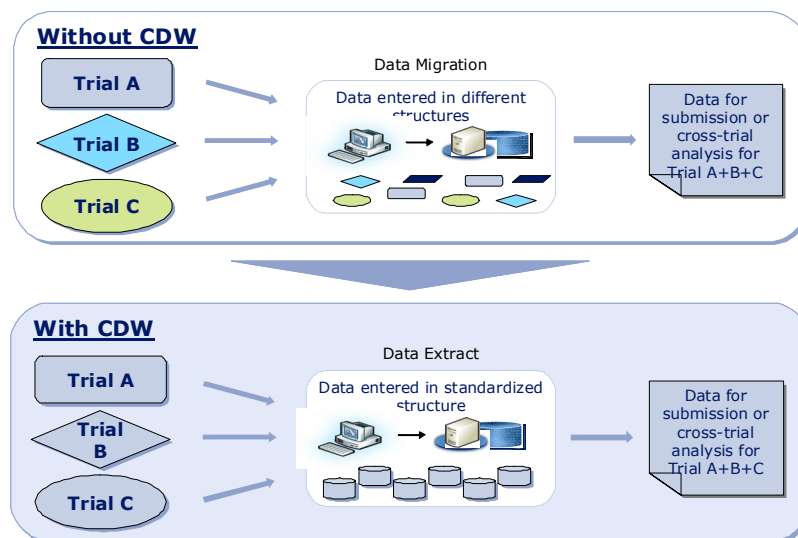


Figure 3 Clinical metadata repository enables standardised terminology and data storage. You can have trials with different design (illustrated by different shape) but if you use terminology inconsistently (illustrated by different colours in the trial) your data content will be difficult to search. Also, if you do not store the data in a consistent manner integration of data will be inefficient.

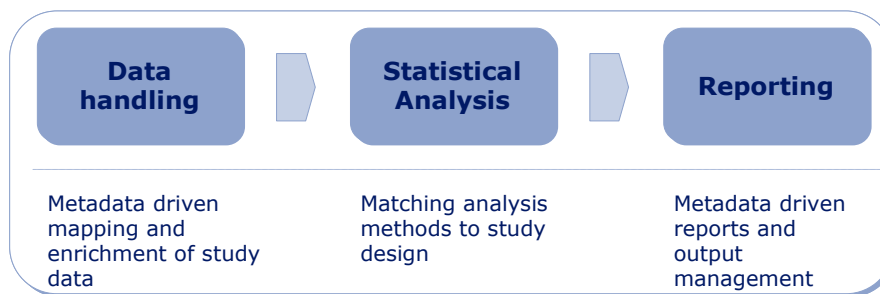
TRIAL METADATA

Trial metadata is a description of a particular trial: basic description, trial design (based on a generic trial design), the visits and a full description of the planned assessments, which we call the flowchart. Also, we define statistical business rules associated to visits and assessments, e.g. selection of different standard rules for missing observations, definition of baseline visit, treatment emergent adverse event rules.

This metadata information is stored in our clinical data model and used throughout the statistical business process, see Figure 4.

The trial metadata - used throughout the statistical business process

In order to make use of the data in the CDW, the trial metadata must be standardised



This enables:

- Programming code reusability
- Integration of data into standardised structure
- Data preparation & reporting optimisation
- Frontloading

Figure 4 Metadata used throughout the statistical business process

In the CDISC SDTM Implementation guide v. 3.1.2, Section 7.1.1., the purpose of specifying your trial metadata (The Trial Design Model) is:

The Trial Design Model in the SDTM provides a standardized way to describe those aspects of the planned conduct of a clinical trial shown in the study design diagrams of these examples. The standard Trial Design Datasets will allow reviewers to:

- *clearly and quickly grasp the design of a clinical trial*
- *compare the designs of different trials*
- *search a data warehouse for clinical trials with certain features*
- *compare planned and actual treatments and visits for subjects in a clinical trial.*

Modelling a clinical trial in this standardized way requires the explicit statement of certain decision rules that may not be addressed or may be vague or ambiguous in the usual prose protocol document. Prospective modelling of the design of a clinical trial should lead to a clearer, better protocol. Retrospective modelling of the design of a clinical trial should ensure a clear description of how the trial protocol was interpreted by the sponsor.

Not only do you provide the regulatory reviewer with a quick overview of your trial you also help yourself while analysing the trial by structuring the trial metadata, ref. page 1. So, at Novo Nordisk we go beyond these benefits. We also use the trial metadata in the data processing and analysis and reporting process.

AN APPLICATION TO MANAGE THE METADATA EFFICIENTLY

Since most of the clinical and trial metadata is in paper format like trial protocols, statistical analysis plans and submissions standards, we designed together with NNIT, an entry application to support that this information can be used in a more automated, consistent and structured way. The application is a Java™ based application where most of the trial metadata is entered using drop down lists, see Figure 5. The drop-down lists are controlled by the definitions in the Clinical Metadata, see Figure 6, and maintained by dedicated functional support.

Novo Nordisk A/S, CDW Operations - Microsoft Internet Explorer

File Edit View Favorites Tools Help Links »

CDW Operations
Metadata Management

Metadata Trial Manage Logout

Define Trial Definition Attributes
Define Protocol, Statistical and Cross-Trial Attributes for Trial Definition

Trial ID: BIAsp-1526
Trial Def ID: Clinical Data Migration
Trial Metadata Version: 2
Trial Definition Status: Draft

Protocol Attributes

EUDRACT Number: 0
Trial Phase: Phase III
ICH Category: Therapeutic Confirmatory
CDP Trial: True
Open Label: ☐
Blinding Level: Double Blind
Double Dummy Blinding: ☐
Generic Trial Design Class: Crossover
Trial Title: A Stratified, Randomised, Double-Blind, Two-Period Crossover Trial Comparing the Pharmacodynamics and Pharmacokinetics of BIAsp 50 and BIAsp 70 in Non-obese and Obese Subjects with Type 2 Diabetes
Primary Objective Text: To compare the pharmacodynamics (AUCGLU (0-24h)) of thrice daily BIAsp 50 with that of thrice daily BIAsp 70, both after 4 weeks of treatment, in both non-obese (BMI 23-28 Kg/m2) and obese (BMI 30-35 Kg/m2) subjects with type 2 diabetes.

Done Local Intranet

Figure 5 Example of Trial Metadata entry screen. E.g. the ICH Category is controlled by special users via definition of the Clinical Metadata, see Figure 6.

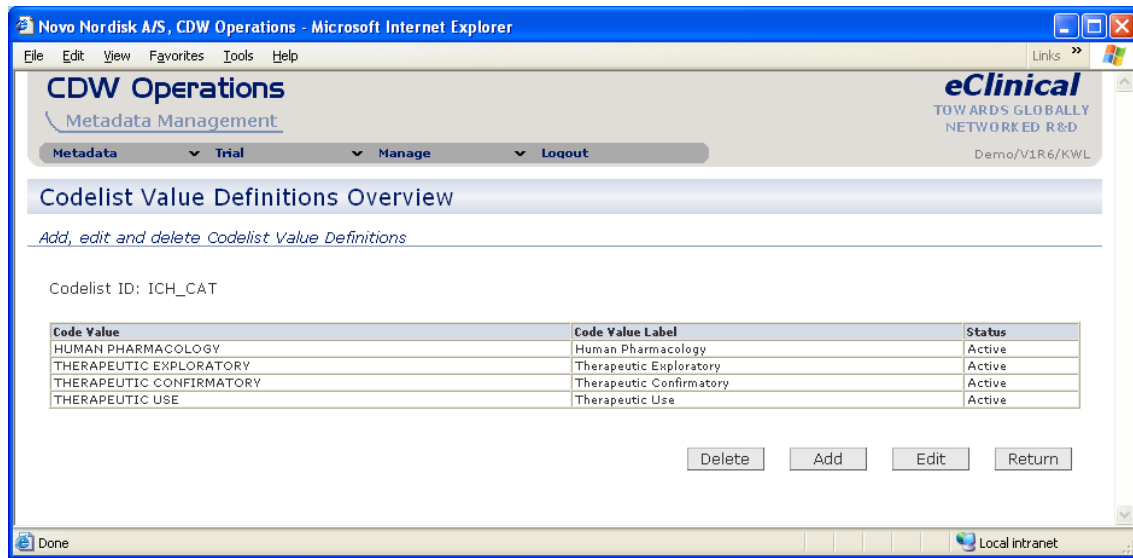


Figure 6 Example of Clinical Metadata entry screen

THE DATAMODEL BEHIND THE SCENES

In some systems controlled terminology is plain code lists without any interdependency. This has the disadvantage that codes are taken out of a context and often the semantics is lost. Therefore, in order to manage and use the metadata efficiently it is important to have a clear picture of what they should be used for and hence design a data model that supports these needs. For the data model behind the metadata we considered:

- Semantics
- Common definition of concepts
- Common language
- Structure on how to represent data
- Different data structures for different tasks

The conceptual data model the clinical metadata behind the application is described in Figure 7.

CDW Conceptual Data Model - Clinical Metadata data model

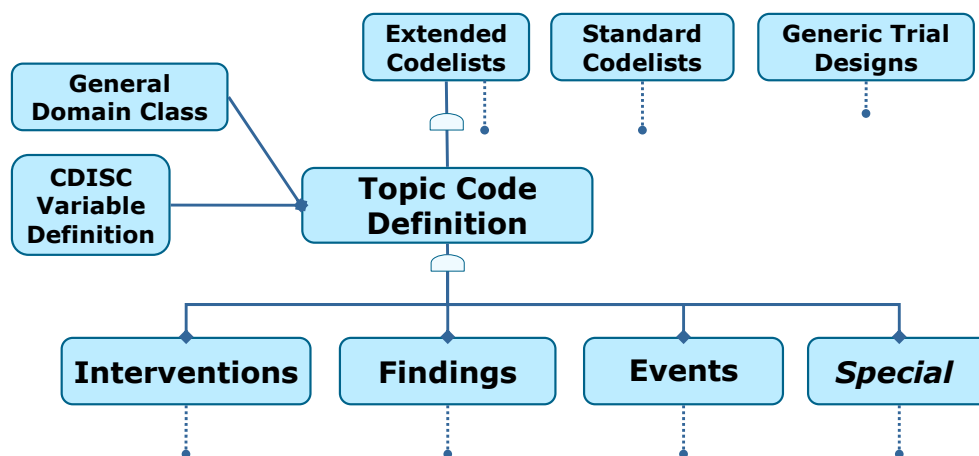


Figure 7 Conceptual Data Model of Clinical Metadata

The main elements of the clinical metadata model are:

- Submission data definition
 - The definition of the *CDISC Variable Definition* which are defined in SDTM model.
- General definition of the collected data
 - The definition of the *General Domain Class* which is the CDISC definition of the major data classes: Finding, Intervention, Events, Special Purpose and Trial design domains. For each of the domains the definition of a topic code and its attributes is central. Hence the topic code definitions are a library of the assessments collected in any trial.

A Topic Code defines a record in the datasets that are used for analysis. The codes can be of different types: Intervention, Findings, Events or Special codes. These codes are not just simple codes but have several attributes, see Figure 8 for an example of a Finding Topic code data model.

CDW Clinical Metadata data model - Finding Definition

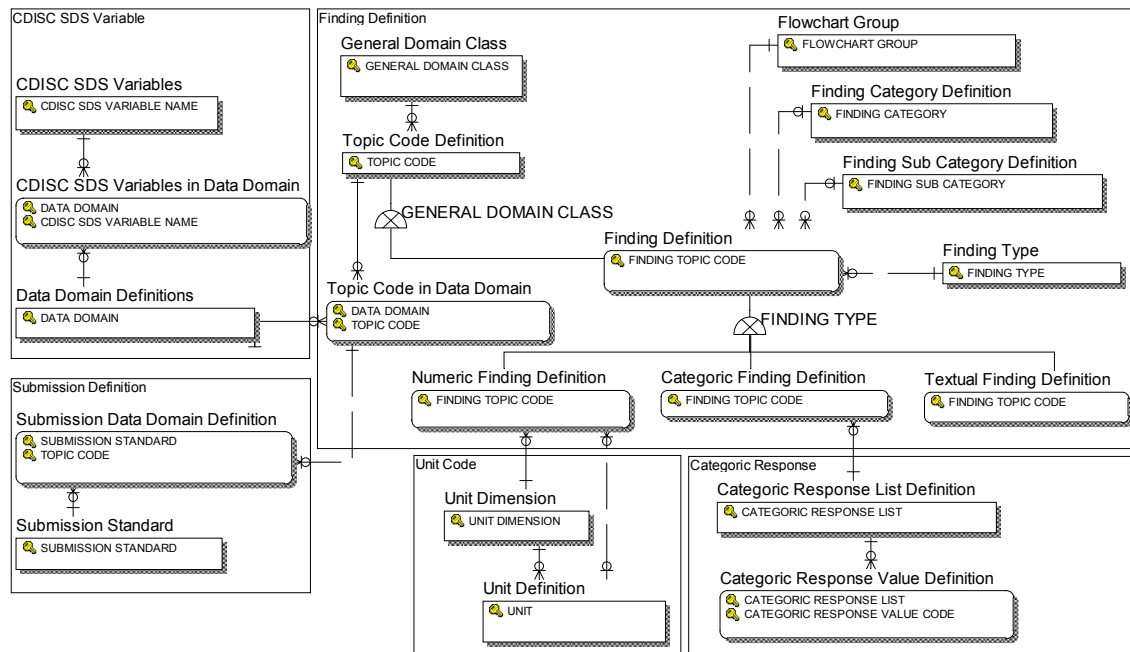


Figure 8 Data Model of a Finding Topic code

In the clinical metadata model we distinguish between numeric, categorical and textual findings. Depending of the type the code can have different attributes, e.g. numeric have a unit whereas categorical have response lists.

THE IMPLEMENTATION OF THE CLINICAL METADATA REPOSITORY

In Figure 9 an example of a definition of a numeric topic code is shown

Novo Nordisk A/S, CDW Operations - Microsoft Internet Explorer

File Edit View Favorites Tools Help Links »

CDW Operations
Metadata Management

Metadata Trial Manage Logout Production/V1R6/KWL

Maintain Finding Definition

Define common Finding Definition attributes

Finding Topic Code: TOTAL_PROTEIN_SERUM
 Short Finding Topic Code: PROT
 Finding Label: Serum total proteins
 Finding Short Label: Total proteins
 Description: CDISC submission value: PROT
 CDISC code: C64858
 CDISC synonym: Protein
 CDISC: A measurement of a group of complex organic macromolecules composed of one or more alpha-L-amino acid chains in a biological specimen.
 Sort Sequence: 2050
 Status: Active
 Language:
 Standard: ☒
 External Terminology: ☒
 Finding Category: Biochemistry
 Finding Sub Category:
 Finding Type: Numeric Finding
 Derived Variable: ☐
 SAS Display Format: 5.3
 Default Flowchart Group: Safety

Save Cancel

Done Local intranet

Novo Nordisk A/S, CDW Operations - Microsoft Internet Explorer

File Edit View Favorites Tools Help Links »

CDW Operations
Metadata Management

Metadata Trial Manage Logout Production/V1R6/KWL

Define Numeric Finding Definition attributes

Define attributes specific for a Numeric Finding Definition

Finding Topic Code: TOTAL_PROTEIN_SERUM
 Status: Active
 Unit Dimension: Concentration
 Standard Unit Code: g/dL
 Lower Possible Range: 0.0
 Upper Possible Range: 100.0
 Molecular Weight:

Return Save Cancel

Done Local intranet

Figure 9 Example of definition of numeric topic code Total Protein and its attributes

The topic code is defined only this place and used through out the system. This ensures a common understanding of the assessment and a consistent definition across the trials it is used in. Changes to the definition will only have to be made one place and subsequently be effective in all trial using the topic code. This, however, requires that a process is in place to handle legacy trials using the 'old' definition, if a code is changed. See section *A Business process to manage the metadata effectively* on page 11.

In order to align to the standardisation induced by CDISC and standardisation internally at Novo Nordisk A/S we have made a convention regarding the content of fields in the screens above, see Figure 10. Here, we map out where the information is coming from and used.

Example of a Clinical Metadata Topic Code (Findings) Total Protein

Code Value	<i>TOTAL_PROTEIN_SERUM - May relate to OC DVGs</i>	SAS Display Format	<i>5.3 Value Display in EOT</i>
Short Code Value*	<i>PROT - May relate to CDISC code value</i>	Default Flowchart Group	<i>Safety - Default Flowchart Group</i>
Label*	<i>Serum total proteins - As in Protocol and EOT</i>	For Categorical Findings only:	
Short Label*	<i>Total proteins - As short values in EOT</i>	Categoric Response List	<i>Link to std. code list</i>
Description	<i>CDISC submission value: PROT CDISC code: C64858 CDISC synonym: Protein CDISC: A measurement of a group of complex organic macromolecules composed of one or more alpha-L-amino acid chains in a biological specimen.</i> <i>Include CDISC terminology</i>	For Textual Findings only:	
Sort Sequence*	<i>Internal sorting number</i>	Maximum text length	
		Split text in supplemental qualifier	<i>In submission standards</i>
External Terminology	<i>Y - Flag if defined externally</i>	For Numeric Findings only:	
		Unit Dimension*	<i>Concentration Unit conversion in DDM</i>
Finding Category	<i>Biochemistry -Default grouping in Flowchart + EOT</i>	Standard Unit Code	<i>g/dL In EOT</i>
Finding Sub Category	<i>-Default grouping in Flowchart + EOT</i>	Lower Possible Range	<i>0 Flagging in DDM</i>
Finding Type	<i>Numeric Finding</i>	Upper Possible Range	<i>100 Flagging in DDM</i>
Derived Variable	<i>Defined as Clinical Metadata but calculated in DDM</i>	Molecular Weight	<i>Unit conversion in DDM</i>

EOT refers to the report standards, if available otherwise as it would be used in an EOT, DDM are derived data

Figure 10 Overview of data sources for clinical metadata

We are sourcing our data from an Oracle® Clinical system in which we use DVG's (Discrete Value Groups, i.e. code list). In order to transfer the data from the source to the CDW system the codes need to match. Also, it gives us traceability to the source system. We refer to CDISC if the code exists in the terminology defined by CDISC. Otherwise internal codes are agreed upon. We specify different label used in the protocol and used in the statistical analysis output (EOT, End-of-Text), since shorter labels are need in tables and figures and long labels are used in Listings. As part of the Topic code definition we also specify in which category a topic code mostly belongs to – this is the Finding category, Finding subcategory and Default Flowchart Group. In relation to the output tables we define SAS display format, standard unit, upper and lower possible range. Within a unit dimension we can convert any unit to the standard unit or any other unit in the dimension.

The topic code definition comes into play when you select the trial metadata for your particular trial. In the application, the full overview of a trial's clinical planned events is defined as specified in the protocol. This is called the trial flowchart, see Figure 11. This is used in particular in the data handling process, comparing planned against the observed. Each assessment in the flowchart corresponds to a topic code and the grouping is defined as part of the attributes, see Figure 10.

Trial Flowchart
Define Trial Flowchart Attributes

Trial ID: BIAsp-1526
 Trial Definition ID: Clinical Data Migration
 Trial Metadata Version: 2
 Trial Definition Status: Draft

Assessment	Visit Short Label	V1 Screening visit	V2 Baseline visit	V3 Treatment visit	V4 Treatment visit	V5 Treatment visit
EFFICACY						
Blood Glucose		x				
Blood HbA1c						
PROFILES						
Serum C-peptide						
Serum Glucose						
Insulin Aspart Specific Antibodies						
Self-measured plasma glucose profile						
Self-Measured Plasma Glucose level						
EXCLUSION CRITERIA						
INCLUSION CRITERIA						
Inclusion Criteria						
SOP						
Inf cons before trial-related activities		x				
DEVICE EFFICACY						
OTHER ASSESSMENTS						
End of trial						
Subject Discontinued						
REMINDERS						
SAFETY						
Adverse event						
Adverse event			x	x	x	x
Hypoglycaemic episodes			x	x	x	x
Hypoglycaemic episodes						
Biochemistry						
Serum ALAT		x				
Serum creatinine		x				

Figure 11 Example of a flowchart for a trial - all clinical planned events are described and controlled via the clinical metadata – the topic code definitions.

The application does not have a report module, but all the clinical metadata (and trial metadata) is transferred to the SAS® DD, which constitutes the statistical computing environment of the CDW system. All the clinical metadata is stored in SAS® data sets in a dedicated folder available to all users. This enables the user to look up any definition of the terminology used and use it in the statistical programming.

As specified in the data model in Figure 8 the application also provides a specification of CDISC Submissions standards. Each SDTM domain is defined, the SDTM variables are defined, and topic codes to be part of a domain can be selected from the list of Topic codes, see Figure 12 and Figure 13.

CDW Operations
 Metadata Management
 Metadata Trial Manage Logout
 eClinical
 TUWARIUS GLOBALLY NETWORKED R&D
 Production V1R.6/9/WL

Maintain Data Domain Definition
Define Data Domain Definition attributes

Data Domain: VS
 Short Data Domain: VS
 Label: Vital Signs
 Short Label: Vital Signs
 Description: CDISC standard data domain for Vital Signs
 Sort Sequence: 1
 Status: Active
 Language: English (U.K.)
 Standard: ☒
 External Terminology: ☒
 Data Domain Structure: One record per vital sign measurement per time point per visit per subject

Topic Codes Variables Save Cancel

Figure 12 CDISC Domain definition

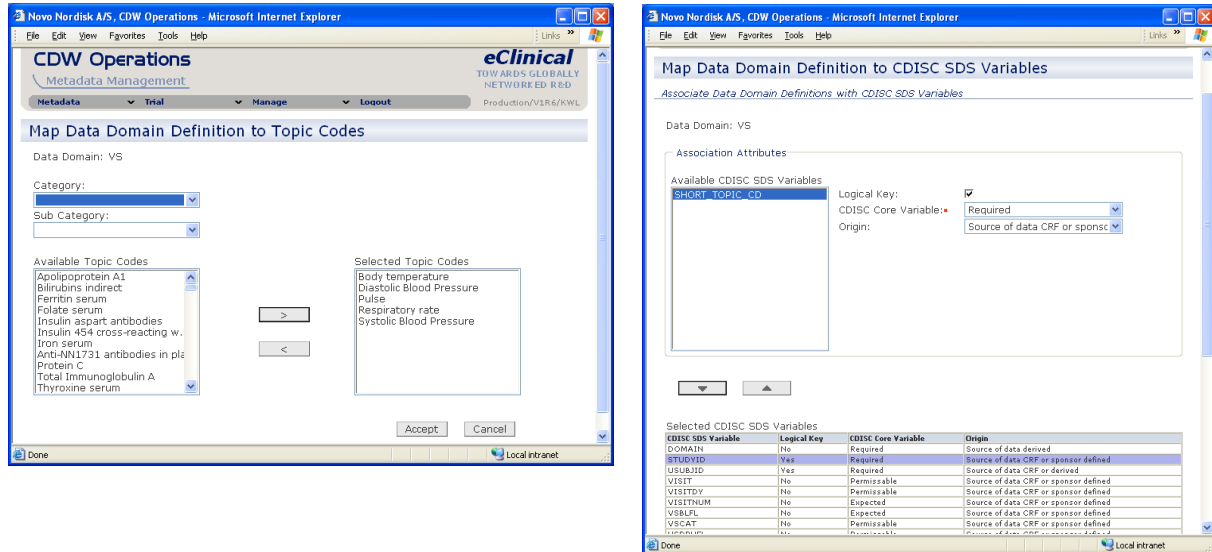


Figure 13 Selection of Topic codes (rows) and CDISC SDS variables (Columns) in the domain

This information is stored in datasets, so by joining these, it is possible to create CDISC specification as an empty table. The actual mapping of the CDW data variables to CDISC SDS variables will be done in a SAS® program. The terminology is controlled centrally and can be used for all trials ensuring a consistent interpretation of the submissions standards.

A BUSINESS PROCESS TO MANAGE THE METADATA EFFICIENTLY

First step in managing you metadata efficiently is to have an application or at least stored in a database, e.g. as SAS® data sets. Next step, which is the more difficult part, is to institute a process and allocate dedicated resources to maintain the repository.

At Novo Nordisk A/S we have established three standards groups: Protocol standard, CRF standards and Report standards groups. The purpose of the groups is to describe and manage the standard terminology used, CRF pages, Tables, listings, Figures and business rules to be used in the trial data processing. Each group has a chair person and representatives from several functional groups involved in the data collection and analysis and reporting process. These groups manage what we call the Business Standards. Change requests to a standard are coming from the users and managed by these groups.

In order to facilitate that the standards are operational, personnel have been allocated to maintain the clinical metadata repository. In general, all codes represented in the Business Standards are also in the repository. Change requests to the Business Standards are communicated to the operational group and implemented into the system. However, the Business Standards are defined to a certain level of granularity. E.g. specific responses to questionnaires and other detailed code lists is not managed through the Business Standards. These codes are requested by the users through an issue list and managed by the operational group.

In order to ensure consistency, regular code review meetings are held. This is particularly important in the build up phase of the clinical metadata repository. Also, to ensure maintainability in the long run, a strategy and process should be in place to handle changes to existing codes and what to do with trial data affected by a code change.

CONCLUSION AND LEARNINGS

In this article we have tried to give the reader a taste of the metadata repository that we have at Novo Nordisk A/S. The system went live June 2008, so we have not seen the full benefit of it yet. However, in the year that we have used the system we can draw the following conclusions and learnings:

- Clinical metadata should not be simple code lists but designed around a data model accounting for the context, semantics and relationships. Same assessments are actually called the same, defined in the same way and controlled centrally
- A clinical metadata repository helps you drive standardisation. The data used in analysis is largely driven by clinical metadata selected for a trial – trial metadata
- An application like the Meta Data Management System helps you structure our data and facilitate

PhUSE 2009

communication about what we collect and how

- It takes long time to build up a clinical metadata repository – it is not done in weeks!
- Ensure dedicated personnel to maintain the repository
- Ensure that a process is in place to maintain the repository and in particular changes to it. Standards are not static.
- At first the process with requests seems like a heavy process, but it makes the requester think about the data and the approval process will ensure some level of consistency and reuse.

ACKNOWLEDGMENTS

I will like to acknowledge my colleagues Mikkel Traun (co-author of this article) for taking time to challenge and inspire within the field of CDISC and clinical metadata. For letting me join him in the pioneer days when the CDW system took its first design steps. It is always inspiring to have innovative colleagues you can share ideas with.

RECOMMENDED READING

CDISC SDTM implementation guide

CDISC Terminology standards

CONTACT INFORMATION

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

Kirsten Walther Langendorf

Novo Nordisk A/S

Vandtaarnsvej 114

Soeborg 2860

Work Phone: +45 30757464

Fax: +45 4442 4600

Email: kwl@novonordisk.com

Web: www.novonordisk.com

Brand and product names are trademarks of their respective companies.