Paper DS04

CDISC Biomedical Concepts
What are they and why should I care?

Stuart Malcolm, Head of Standards, Efficiency and Automation, Veramed
Stuart.Malcolm@Veramed.co.uk

ABSTRACT

The CDISC initiative to develop standard Biomedical Concepts (BC) that link all foundation standards offers exciting possibilities – but what exactly is a biomedical concept? What problems does it solve? How and why should we use them?

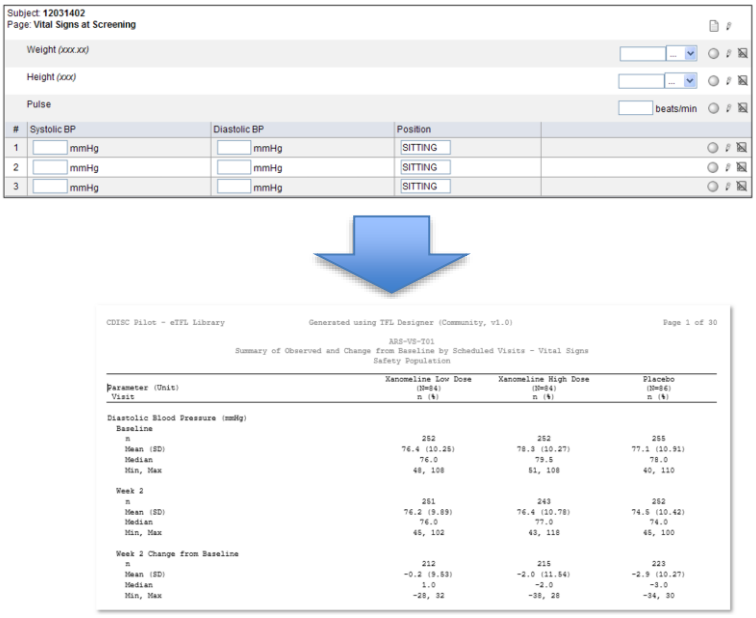
The CDISC Biomedical Concept initiative started in 2022 with an aim to create biomedical concepts and integrate them into all the Foundation Standards (CDASH, SDTM, ADaM), and during 2024 the initiative has completed its first public review of BCs with SDTM dataset specializations [1].

This presentation aims to offer a “Getting Started” guide to CDISC Biomedical Concepts initiative, introducing the concepts, walking through practical examples, and discussing possible use-cases and possible “quick-win” implementation scenarios.

WHAT PROBLEM ARE WE SOLVING?

Overall, CDISC Biomedical Concepts aim to improve the quality and efficiency of clinical trials by promoting consistency, automation and ease of implementation.

Basically – how to make it easier to go from source data to results:



Source Data
EDC, ePRO, RWD, etc.

Results
Dataset, TFL, Visualisation, etc.

Figure 1Biomedical Concepts improve efficiency of transforming source data to outputs

CDISC Biomedical Concepts (BCs) are designed to address several key challenges in clinical research data management [1]:

- **Reducing Variability in Standards Implementations:** By providing preconfigured value-level metadata, BCs help ensure consistent implementation of standards across different studies and organizations
- **Increasing Metadata-Driven Automation:** BCs facilitate automation in data collection, transformation, and reporting processes, reducing manual effort and the potential for errors

- **Reducing Barriers to Operational Implementation**: BCs streamline the process of configuring standards for data collection, making it easier for organizations to adopt and implement these standards.

CURRENT CDISC WORKFLOW

The current workflows across the research lifecycle is shaped by the need to work with source documentation (Clinical trial documentation such as Protocols, Data Management Plans, SAP, etc. as well as CDISC standards documentation) which are electronic but are not machine readable, in the sense that they do not contain the semantic metadata to allow the contents to be interpreted algorithmically.

To deal with this “source documentation is PDF” scenario, we have a workflow that is based on manual processes :

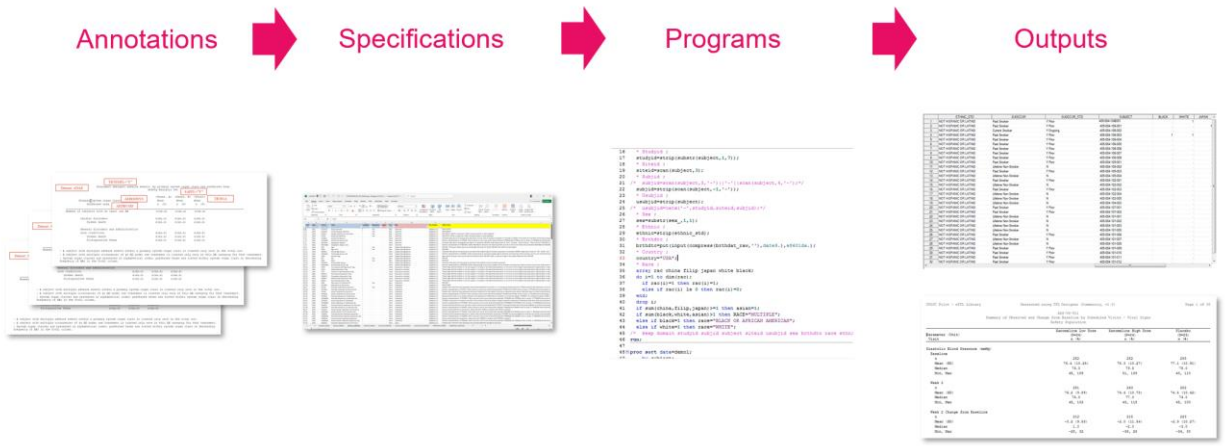


Figure 2 Current workflow involves multiple manual processes to transform data from source to output

The annotation process helps to ensure that all relevant parts of the source document are being ingested, also to document the traceability and to make it easier to verify the specification, which defines the requirements in sufficient detail to allow a program to be (usually manually) written. The programming is used to create the outputs in a reproducible manner which are validated with respect to the source data and documentation.

Note that the workflow described in Figure 2 above is not intended to describe the end-to-end workflow, but rather that end-to-end we use smaller ‘cycles’ based on variations of the annotate-spec-program-output workflow – so for example during the study-build process by the clinical data management team, through to statistical programming.

MOVING FROM METADATA TO MODELLING

CDISC Biomedical Concepts are a foundational component of the digital transformation of clinical trial biometrics. This is a shift from the current ‘electronic’ paradigm – where source documentation is in electronic format (e.g. PDF, Excel, etc.) – to a digital paradigm, where source documentation is machine readable – i.e. a digital model that describes the content, structure and semantics of the clinical trial.

A digital model is an abstraction of things in the real world that incorporates both behaviour and structure [2]

- It aims to express the meaning of terms and concepts used by domain experts to discuss the problem
- ..and to find the correct relationships between different concepts
- ..is described using defined notation

So, where we currently use the term “metadata” which is a static description of data we move to use “models” which contain both the static structure and the semantic relationships of the data.

This paper aims to describe how CDISC Biomedical Concepts meets this definition of a digital model, and how that can be used to drive automation and improved efficiency.

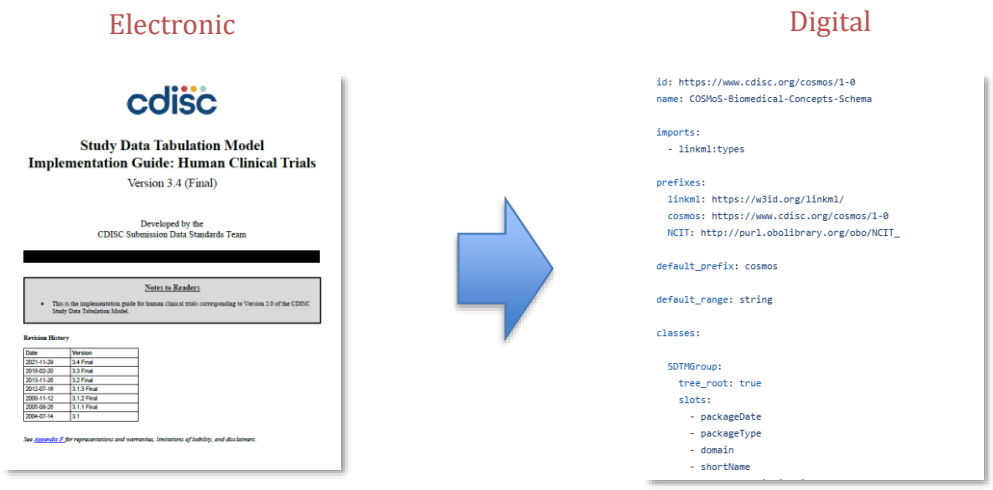


Figure 3 The move from electronic documents to digital models

One of the key benefits of digital models is that they can be implemented in software, allowing them to be easily manipulated by software tools to allow users to work at a higher level of abstraction which is closer to the domain-specific language used by experts (e.g. clinicians, investigators, statisticians, etc.) and eliminating the need for some of the manual processing steps/programming that we rely on today.

INTRODUCTION TO CDISC BIOMEDICAL CONCEPTS

Information on CDISC BC's are published on the CDISC website [1] and also on the projects GitHub repository [3] where there are introduction and training materials for different use-cases, from using existing BC through to the governance process for creation of new BC content.

CDISC Biomedical Concepts (BCs) have a two-layered architecture.

- **Conceptual/abstract layer** that provides standards-agnostic, unambiguous semantic definition largely based on NCIt concepts.
- **An implementation layer** consisting of CDASH and SDTM Dataset Specializations provides value level definition that facilitates metadata-driven automation.

The Conceptual/Abstract Layer is the 'core' set of biomedical concepts – There are two parts to each definition (see below), the Biomedical Concept (BC) and the Data Element Concept (DEC). Each Biomedical Concept has zero or more Data Element Concepts associated with it.

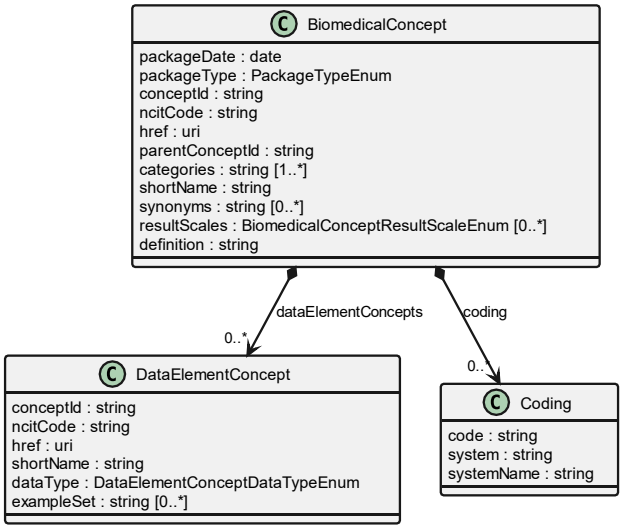


Figure 4 CDISC BC model showing relationship with Data Element Concepts and Coding system

The **Biomedical Concept** is the top-level definition and is the "unit of knowledge" that defines a real-world phenomena in clinical research. Each Biomedical Concept has a unique identifier, usually based on the NCI Thesaurus [4].

For example, a Ketone Measurement is defined as "A quantitative measurement of the amount of ketones present in a sample" has the unique BC identifier C64854.

The **Data Element Concept** provides a structured way to represent clinical data, ensuring consistency and clarity across different studies and datasets. A Biomedical Concept can contain zero or more Data Element Concepts (DEC) and each DEC has its own unique identifier. Data Elements can be shared/reused across many BC – so, for example there is one Laboratory Test Result (C36292) data element concept which is referenced by many Laboratory biomedical concepts.

For example, the Ketone Measurement concept (C64854) encapsulates the following Data Element Concepts: C36292 Laboratory Test Result, C83309 Laboratory Test Fasting Status, C48207 Unit of Concentration, and C70713 Biospecimen Type.

The Ketone Measurement Biomedical Concept (BC) and associated Data Element Concepts (DEC) is shown in the table below:

Biomedical Concept			Data Element Concept			
short_name	bc_id	definition	ncit_code	label	data_type	example_set
Ketone Measurement	C64854	A quantitative measurement of the amount of ketones present in a sample.				
Ketone Measurement	C64854		C36292	Laboratory Test Result	string	
Ketone Measurement	C64854		C83309	Laboratory Test Fasting Status	boolean	
Ketone Measurement	C64854		C48207	Unit of Concentration	string	mg/dL; mmol/L
Ketone Measurement	C64854		C70713	Biospecimen Type	string	Blood; Serum; Plasma; Urine; Vitreous Fluid

Figure 5 Ketone Biomedical Concept BC) and related Data Element Concepts (DEC) in tabular form

So, to summarise, the Conceptual/Abstract Layer consists of a model (aka schema) that defines both Biomedical Concepts (BC) each of which has a unique identifier, and associated Data Element Concepts (DEC) for each BC, which also have unique identifiers.

In addition to the Conceptual Layer, the CDISC initiative also includes an Implementation Layer which consists of a number of models which 'map' the Conceptual Layer onto CDISC data standards, such as CDASH or SDTM. These are effectively machine-readable versions of the CDISC Implementations guides (IG) which are currently published as PDF documents.

BIOMEDICAL CONCEPT IMPLEMENTATION LAYER – CDASH DATASET SPECIALIZATIONS

The implementation layer currently consists of CDASH and SDTM Dataset Specializations which provides value level definition that facilitates metadata-driven automation.

These can be through of as a digital model that describes the current CDISC Implementation Guides, so, for example, the Collection (CDASH) Dataset Specialization schema describes a Data Collection Group which contains multiple Data Collection Items, each of which contains multiple Code Lists, List Value, Pre-populated Value and SDTM Target (an annotation text, a list of SDTM variables, and a mapping string)

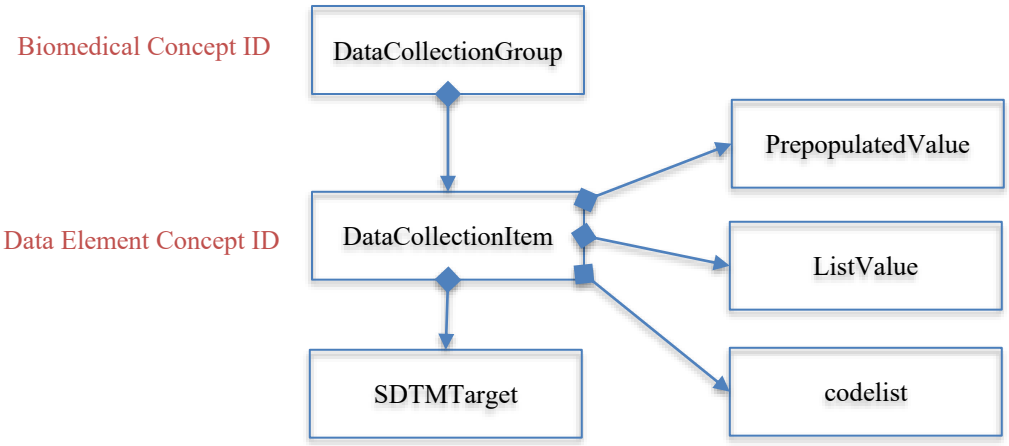


Figure 6 Collection (CDASH) Dataset Specialization Schema

As shown in figure 6 above, each Data Collection Group has a Biomedical Concept ID associated with it, and each Data Collection Item has a Data Element Concept ID.

This structure allows us to easily do two things:

1. Given a biomedical concept that needs to be collected (e.g. Body Height) this BC can be used to select all the Data Collection Groups with that BC identifier, and the associated Data Collection Items, etc. This can then be used to populate an EDC with CRF forms to collect the required data.
2. Given data collected for a given BC (e.g. Body Height CRF) then the SDTM target variables can be selected and used to automate the generation of an annotated CRF, and/or used to populate an SDTM mapping definition (or automate the generation of SDTM dataset using the SDTM Dataset specialization – see below)

BIOMEDICAL CONCEPT IMPLEMENTATION LAYER – SDTM DATASET SPECIALIZATIONS

The SDTM dataset specialization follows a similar pattern to the CDASH specialization with SDTM Group (aka Domain) which has a BC identifier, and contains SDTM Variables which has a DEC identifier along with attributes defined in the SDTM standard (role, datatype, length, format, etc.). Similarly, Assigned Term and CodeList are familiar from the SDTM standard.

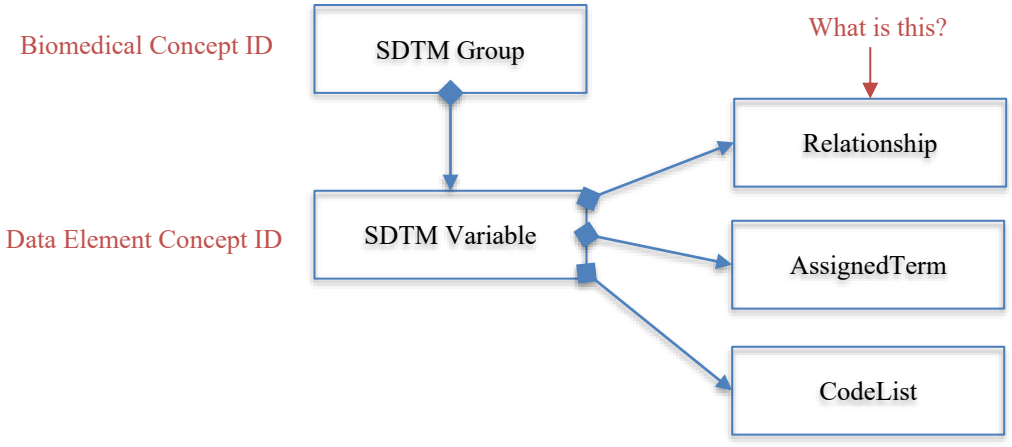


Figure 7 Simplified SDTM Dataset Specialization schema

However, what about the Relationship class? This defines the relationship between SDTM Variables. The Relationship class has 4 attributes (shown below). This is a subject-predicate-object tuple that defines the relationship using a linking phrase (human-readable decoded text) and a predicate term (coded version of the linking phrase)

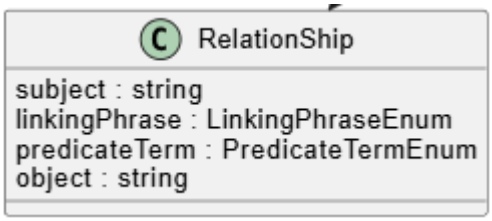


Figure 8 SDTM Dataset Specialization Relationship attributes

The linking phrases and predicate terms are defined as enumerations (aka codelists) within the Dataset Specialization – there is no mechanism for 'extending' these lists, however there is currently a long list of pre-defined relationships to use.

Subject	Linking Phrase	Predicate Term	Object
VSDTC	is the date of occurrence for	IS_TIMING_FOR	VSTESTCD
VSORRES	is the result of the test in	IS_RESULT_OF	VSTESTCD
VSORRESU	is the unit for the value in	IS_UNIT_FOR	VSORRES
VSSTRESC	is the result of the test in	IS_RESULT_OF	VSTESTCD
VSSTRESN	is the result of the test in	IS_RESULT_OF	VSTESTCD
VSSTRESU	is the unit for the value in	IS_UNIT_FOR	VSSTRESN
VSTEST	decodes the value in	DECODES	VSTESTCD
VSTESTCD	is the code for the value in	IS_DECODED_BY	VSTEST

Figure 9 Example of Relationship subject-predicate-object values from VS dataset specialization

NOTE that there is not a 1:1 relationship between Linking Phrase and Predicate Term. Currently the standard has 29 Predicate Terms (these are the coded predicates e.g. IS_TIMING_FOR, IS_RESULT_OF, etc.) and there are 81 unique Linking Phrases (these are the human readable text phrases, e.g. "is the date of occurrence for", etc.)

So, for example the SPECIFIES predicate term is used with all the following linking phrases:

- further describes the test in
- further specifies the anatomical location in
- identifies the reference used in the genomic test in
- specifies the anatomical location in
- is the severity of the toxicity in
- is the method of secondary analysis of results in
- is the method for the test in
- is the frequency of administration of the amount in
- is the administration anatomical location for the treatment in
- indicates the progression status of the previous irradiated tumor identified by
- indicates the previous irradiation status of the tumor identified by

In addition, some of the Predicate Terms appear to be synonyms, such as PERFORMED and PERFORMS with are both used with the same linking phrase:

- is the role of the assessor who performed the test in

Given the relatively early stage of development of the SDTM Dataset Specialization, it is perhaps understandable that the predicate terms and linking phrases are yet to be fully stabilized – however, given the importance of variable relationships this is clearly a priority focus for standardization in the future.

By adding these variable relationships to the SDTM Dataset specialization, it embeds the semantic information into the SDTM "metadata" which can be used to automate downstream processing of the data for e.g. reporting, creation of analysis datasets, etc.

CONCLUSION

CDISC Biomedical Concepts (BCs) are designed to address several key challenges in clinical research data management [1]:

- Reducing Variability in Standards Implementations
- Increasing Metadata-Driven Automation
- Reducing Barriers to Operational Implementation

CDISC Biomedical Concepts represent the next step in the evolution of CDISC Standards – from electronic metadata to digital models

A digital model is an abstraction of things in the real world that incorporates both behaviour and structure [2]

- It aims to express the meaning of terms and concepts used by domain experts to discuss the problem
- ..and to find the correct relationships between different concepts
- ..is described using defined notation

This enables data and metadata to be transformed from one standard to another end-to-end and also enables the creation of software tools for clinical trial biometrics to replace low-level manual processing.

CDISC Biomedical Concepts (BCs) have a two-layered architecture.

- Conceptual/abstract **layer** that provides standards-agnostic, unambiguous semantic definition largely based on NCIt concepts.
- An implementation layer consisting of CDASH and SDTM Dataset Specializations provides value level definition that facilitates metadata-driven automation.

While currently there are CDASH and ADaM Dataset Specializations in the implementation layer, it should be notes the CDISC roadmap to include all foundation standards in the implementation layer, and that including other non-CDISC standards is also possible.

REFERENCES

[1] CDISC Biomedical Concepts https://www.cdisc.org/cdisc-biomedical-concepts

[2] End-to-end automation using model-driven development. PHUSE EU Connect 2024. (AD09) https://phuse.s3.eu-central-1.amazonaws.com/Archive/2024/Connect/EU/Strasbourg/PRE_AD09.pdf

[3] CDISC Conceptual and Operational Standards Metadata Services (COSMoS). https://github.com/cdisc-org/COSMoS

[4] NCI Thesaurus https://evsexplore.semantics.cancer.gov/evsexplore/welcome?terminology=ncit

CONTACT INFORMATION

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

Author Name: Stuart Malcolm
Company: Veramed
Email: stuart.malcolm@veramed.co.uk
Website: https://www.linkedin.com/in/stuart-malcolm/