Paper DS01

Embracing the Semantic Shift: SDTMIG 4.0 and the Rise of Intelligent Clinical Data Standards

Binitha Sathyabhama, Pfizer Healthcare, Chennai, India
Venkateswarlu Layam, Pfizer Healthcare, Mumbai, India

ABSTRACT

SDTMIG 4.0 represents a significant shift in clinical data standards, moving from traditional dataset structures to a concept-driven, metadata-powered framework. Built on Biomedical Concepts and CDISC Library integration, this next-generation paradigm improves traceability, automation, and interoperability-aligning with the broader CDISC 360 vision of metadata-driven automation. This presentation delves into what distinguishes SDTMIG4.0 from previous versions, including the shift toward semantic metadata, machine-readable specifications, and real-time regulatory readiness. We will explain Biomedical Concepts, how they drive uniformity across SDTM and define.xml, and how businesses can start aligning their own libraries and tools with this new paradigm. The discussion will emphasize the shift's benefits and challenges, as well as practical suggestions on how standards teams may prepare for implementation. Whether you're managing global specs or looking into metadata-driven transformations, this lecture will explain why SDTMIG4.0 is more than just a technical improvement; it's a strategic leap into the future of clinical data.

INTRODUCTION

SDTMIG has been the bedrock for decades in tabulating and submitting clinical trial data. Clarity and review-readiness of datasets, regulatory compliance, and consistency are what SDTMIG has offered right from its first release. But changing regulatory requirements, rising data complexity, and advancements in technology have emphasized shortcomings in the conventional dataset-focused model.

When we talk about the essential transition in clinical data guidelines with the shift towards SDTMIG 4.0, it is a significant change. What's new? What's the big change?

In previous releases, upgrading from version to version on a routine basis-meant incremental upgrades — adding more domains, eliminating a few variables, or adding some. But SDTMIG 4.0 introduces a distinct shift in approach.

This release represents a shift from the historical dataset model to a concept-based, meta-driven architecture. It fits with newer trends such as CDISC 360i and biomedical concepts, which seek to introduce greater standardization, interoperability, and efficiency in clinical data.

EVOLUTION OF SDTMIG

Prior Versions (3.x series)

- Dataset-focused: Oriented to fixed structures (domains such as DM, AE, LB).
- SUPPQUAL methodology: Leaning on supplemental qualifiers for non-standard variables, generating inefficiencies.
- Human interpretation: Specs and metadata were primarily human-readable PDFs.
- Limited automation: Reliant on significant use of programming effort to support compliance.

SDTMIG 4.0 Evolution

- REPLACES SUPPQUAL with Non-Standard Variables (NSV) datasets.
- Metadata is machine-readable by way of the CDISC Library.
- Biomedical Concepts are reusable, semantic building blocks.
- Framework aligned with regulatory readiness and automation.

CORE CONCEPTS IN SDTMIG 4.0

CDISC 360I VISION:

The vision of CDISC beyond 360 is to build a fully interoperable and automated clinical research ecosystem by enriching foundational standards with biomedical concepts, ensuring complete data meaning and relationships, and enabling machine-executable data flows from study design to analysis. It focuses on reducing variability in implementation, simplifying standards for end-users, and publishing them from a single trusted source. By aligning with global initiatives like HL7 FHIR, ICH M11, and Digital Data Flow, and advancing modern exchange formats such as Dataset-JSON and ODM v2.0, CDISC aims to integrate real-world, digital health, and device data seamlessly while maintaining regulatory traceability and reproducibility, thereby making clinical research more efficient, consistent, and future-ready.

At its core is linking the study protocol's Schedule of Activities (SoA) to Biomedical Concepts (BCs) that propels automated generation of SDTM specifications and datasets. By building integrated metadata packages, 360i puts executable standards in place that not only automate the creation of SDTM datasets but also improve data traceability and reusability.

BIOMEDICAL CONCEPTS (BCS):

A pragmatic, iterative development of biomedical concepts with an emphasis on giving tangible value to the CDISC community.

Simultaneously, Biomedical Concepts provide a semantic and normalized abstraction layer (Conceptual Layer) for clinical data elements like "Body Temperature" or "Date of Diagnosis" so that there is consistent interpretation in studies and sources. Providing an interphase between abstract concept and its SDTM realization (Implementation Layer), BCs allow dataset specializations to be constructed and enhance the data fidelity through the binding of conceptual definitions to their value-level metadata for SDTM. This rich semantic metadata (Logical data model) then drives automated processes, such as the creation of CDISC Define-XML files and the configuration of study systems.

Advantage: Promotes consistency between SDTM domains, ADaM data sets, and define.xml.

Example: A "Liver Function Test" BC can consistently map to LB domain variables, linked codelists, and metadata in define.xml.

Cumulatively, CDISC 360i and BCs optimize for performance by reducing redundancy, enhance quality through uniform representation, and enable cross-study aggregation and analysis through conceptually grouping data rather than merely by domain. They also enable the creation of FAIR (Findable, Accessible, Interoperable, Reusable) data products that enable sponsors, researchers, and regulators to work with interoperable, machine-readable data sets that speed research and regulatory review.

CDISC LIBRARY INTEGRATION

- Provides API access to metadata.
- Allows tools to take standards in machine-readable form directly.
- Minimizes manual mapping errors.

KEY DIFFERENCES: SDTMIG 3.4 VS. 4.0

1. REPLACEMENT OF SUPP DATASETS WITH NSV DATASETS

One of the most revolutionary improvements in SDTMIG 4.0, is to non-standard variable handling, from the "SUPP--" (Supplementary Qualifiers) domain structure to a new one. The change is being made from SUPP-- to NS-- (Non-Standard Variables (dataset)). They are implementing "horizontal" Non-standard dataset structure which allows for simpler merging with parent domains.

Every supplementary qualifier dataset is allocated a parent domain, according to available supplementary data structure. All of the main variables like STUDYID, RDOMAIN , USUBJID and IDVAR/IDVARVAL are used in referencing the record variable QVAL (data value) with parent and other qualifier variables as well.

This constrained the functionality in depicting numerical or structured values in an effective way and mostly contributed to added complexity in datasets' integration.

NSV datasets overcome these constraints by direct appending into parent domains without the need for transposition.

NSVs would be ordered after the standard variables.

Metadata at variable levels such as data type, length, and other qualifiers are stored so that numeric and structured data can be appropriately represented. This enhances the interpretability and usability of the data considerably.

The metadata role for NSV are determined as per SDTM variable role as below listed.

- Non-Standard Identifier
- Non-Standard Qualifier
- Non-Standard Timing

For instance: Hospitalization data with non-standard variables.

In addition to the standard HO variable, there are the following non-standard variables mapped in NSHO.

HOAERPFL	AE Reported This Episode?
HOMEDSFL	Meds Prescribed?

ho.xpt: Healthcare Encounters

Row	STUDYID	DOMAIN	USUBJID	HOSEQ	HOTERM	HOSTDTC	HOENDTC	HODUR	HOAERPFL	HOMEDSFL
1	1999001	HO	0001	1	Hospital	2004-01-05	2004-01-12	P1W	Y	Y
2	1999001	HO	0001	2	Hospital	2004-01-23	2004-02-07	P15D	Y	Y
3	1999001	HO	0002	1	Hospital	2004-01-21	2004-01-22	P1D	Y	N

Advantages of NSV:

- Numeric Data: Numeric NSVs can be established as a numeric data type rather than a character type.
- Variable Lengths: Character lengths suitable to each NSV can be established, instead of being limited by the QVAL default.
- Metadata at Variable Level: Metadata, such as Controlled Terminology, can be applied to NSVs directly, providing more control.
- Simplified Merging: NSV model employs a shared IDVAR/IDVARVLN specification across different domain types, reducing merging complexity over the previous SUPPQUAL structure. To a reviewer, usability is greatly enhanced with the advent of NSVs. Data can now back join with parent domains without ambiguity or constraint of SUPP structures, enabling quicker and more precise review workflows.

Since industry stakeholders themselves would be required to modify tools, validation rules, and workflows to accommodate NSVs, such an investment is bound to offer long-term efficiencies. This shift has the potential to bring regulatory expectation and industry practice in line, allowing clinical trial data submissions and reviews to be more streamlined.

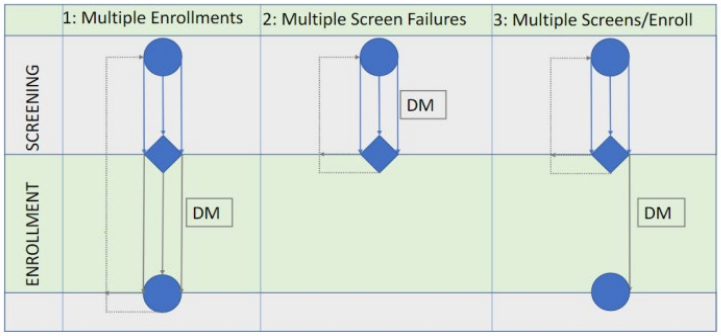
2. STANDARDIZED HANDLING OF MULTIPLE SUBJECT INSTANCES

Another major enhancement in SDTM-IG v4.0 addresses the representation of subjects who multiple screens/re-screens, multiple screen failures and multiple enrolments.

To address this gap, v4.0 introduces expanded guidance and a new domain - Demographics for Multiple Participations (DC) proposed by CDISC SDS (Submission Data Standards) team dedicated to managing multiple subject instances, with a set of essential standard variables that describe each participation of a subject in a clinical trial. This domain provides a standardized, traceable mechanism for distinguishing between separate participations or screening events for the same subject. By doing so, it ensures clarity in subject-level records, supports accurate analysis, and minimizes the risk of reviewer misinterpretation.

As per SDTM rule DM domain each subject should have only one single record per study. FDA clarifies the recommendation on multiple screenings/enrollments as below: -

- For subjects with multiple enrollments within a single study, the primary enrollment should be submitted in DM. Additional enrollments should be included in DC domain with a similar structure to DM.
- For subjects with multiple screenings and no subsequent enrollment, include the primary screening in DM with additional screenings in DC domain with a structure like DM.
- For subjects with multiple screenings and subsequent enrollment, include the enrollment in DM with screenings in DC domain with a structure like DM.



Here are multiple scenarios from different studies -

- Previous participation, subject re-screening, rescreening multiple times.
- Subject enrolled and randomized but no drug received, and same subject enrolled in another site, and received the drug.
- Subject enrolled and received the drug, went to another site and randomized but no drug received.
- Subject enrolled and drug received and went to another site randomized and drug received.
- Subjects are completing the treatment and enrolled to same study again to take another treatment as per protocol.

Subjects taking treatment from cohort and enrolled in other cohort within the same study leads to multiple

enrolments.

3. METADATA RESTRUCTURING

In SDTM IG 4.0 ,the metadata is restructured to align with SDTM 3.0. There are few updates associated with SDTM 3.0 version from previous versions

The updates made to the structure and presentation of domain metadata do not change the variable metadata required for regulatory submission. The intent of these changes is to make the documentation more informative and to help users better understand the purpose and appropriate use of each variable. Key elements such as variable name, type, core status, and most variable labels remain unchanged. Controlled terms, codelists, and formats retain their original definitions but have been reorganized into separate columns for improved clarity.

Some aspects of variable metadata have been modified to enhance interpretability. The “Role” metadata has been simplified by removing the qualifiers associated with the “Qualifier” role, as they were found to provide limited value. The conceptual function they served is now handled more effectively through the new Variable Relationships, which offer clearer and more comprehensive descriptions of how variables relate to one another.

A significant portion of metadata has been redistributed for improved structure. Information previously combined in a single column—such as codelists, formats, and controlled terms—has been separated into individual columns titled “Codelist,” “Format,” and “Allowed Controlled Terms.” Additionally, content formerly housed in CDISC Notes has been reorganized. Definition-like text has been moved to a new Variable Definition column, assumption-related content has been placed in general or domain-specific assumptions with references captured in a new Notes column, examples have been relocated to an Examples column, and valid values have been incorporated into the “Allowed Controlled Terms” column.

SDTMIG v4.0 also introduces several expanded metadata fields to improve guidance for implementers. These include Variable Definition and Variable C-code (as in SDTM), Variable Group—which groups variables with similar functions for informational purposes—and Variable Relationships, which replace the earlier limited roles of “Synonym Qualifier” and “Variable Qualifier.” The Notes column now captures assumption-related guidance previously embedded in CDISC Notes, and the Examples column provides practical illustrations of variable use, also migrated from the earlier CDISC Notes section. Together, these additions create a more structured, transparent, and user-friendly metadata framework.

A row in relationships table can be read by concatenating the subject variable, the linking phrase, and the object variable (e.g., "–TEST decodes the value in --TESTCD").

Subject Root Variable	Subject Root Variable Label	Linking Phrase	Predicate Term	Object Root Variable	Object Root Variable Label
--TEST	Name of Measurement, Test, or Exam	decodes the value in	DECODES	--TESTCD	Short Name of Measurement, Test, or Exam
--TESTCD	Short Name of Measurement, Test, or Exam	Is the code for the value in	IS_DECODED_BY	--TEST	Name of Measurement, Test, or Exam

4. ADDED NEW COLUMN TO SPECIFICATION TABLES – “VARIABLE GROUPS”

As part of the restructuring of metadata, new "Variable Groups" column in specification for grouping variables, making standard terms explicit such as "visit variable", "coding variable".

For instance, rather than describing a variable as being "qualifying" another, the specification will say the variable as "being the unit for the value in" the other variable.

5. UPDATED PROTOCOL DEVIATIONS DOMAIN TO SUPPORT NEW REGULATORY GUIDANCE

Enhances the DV domain to accommodate new regulatory guide documents such as ICH E6 R3. Introduces one new variable, Classification of Protocol Deviation (DVCLASI) which may have values e.g. Important vs Non-Important. DVCLASI can enable use of DVCAT/DVSCAT for other purposes. Enhances SDTM convergence with regulatory guide document and enables BIMO submissions.

6. NEW DOMAIN

Enhances the DV domain to accommodate new regulatory guide documents such as ICH E6 R3. Introduces one new variable, Classification of Protocol Deviation (DVCLASI) which may have values e.g. Important vs Non-Important. DVCLASI can enable use of DVCAT/DVSCAT for other purposes. Enhances SDTM convergence with regulatory guide document and enables BIMO submissions.

GASTROINTESTINAL SYSTEM FINDINGS (GI) DOMAIN

The Gastrointestinal System Findings (GI) domain is a findings domain introduced to capture physiological and morphological observations related to the gastrointestinal system. The GI tract is composed of the oral cavity,

pharynx, esophagus, stomach, small intestine, large intestine, and anal canal, along with accessory organs such as the teeth, tongue, salivary glands, liver, gallbladder, and pancreas.

Within this domain, physiological findings represent functional changes or symptoms like indigestion, heartburn, nausea, vomiting, or lactose intolerance. On the other hand, morphological findings represent structural abnormalities, such as gastritis (inflammation of the stomach lining) or ulcers. By distinguishing between functional and structural , the GI domain offers a perfect model to organize gastrointestinal-related information for clinical trials.

GI – Assumptions

1. This domain is used to represent results and findings of gastrointestinal diagnostic procedures. Information about the conduct of the procedure(s), if collected, is submitted in the Procedure Domain (PR).
2. Any Identifiers, Timing variables, or Findings general observation class qualifiers may be added to the GI domain, but the following Qualifiers would generally not be used in GI: --MODIFY, --BODSYS, --FAST, --ORNRLO, --ORNRHI, --TNRLO, --STNRHI, and --LOINC.

EVENT ADJUDICATION ASSESSMENTS(EA)

The Event Adjudication (EA) domain is one of the new Findings-About domain developed to facilitate standardization of adjudication data submission in clinical trials. Adjudication is a highly important process by which an independent committee examines and authenticates discrete clinical events—e.g., endpoints, diagnoses, or outcomes—so they are objective and consistent. The EA domain offers a structure to record these adjudication findings and link them back to the original events under observation.

In order to use this domain, sponsors will need to create an EA dataset and perhaps also be mandated to change or alter their existing adjudication forms to be in line with CDISC standards. Importantly, sponsors will also need to apply the proper Controlled Terminology (CT) in order that adjudication results, reasons, and criteria can be expressed in a standardized manner.

The EA domain extends and enhances PHUSE-recommended practice in event adjudication and supports regulatory advice, in particular the method outlined by the FDA's Technical Specification for Submitting Clinical Trial Data Sets for Treatment of Noncirrhotic Nonalcoholic Steatohepatitis (NASH). By establishing a standard, precise domain, EA increases consistency, supports traceability back to source events to adjudication output, and minimizes variability in presentation of adjudication data between studies. This ultimately facilitates more effective regulatory review and more comparable adjudication outcomes among sponsors and among therapeutic classes.

BENEFITS OF SDTMIG 4.0

- Improved Traceability: End-to-end linkage from protocol concepts → SDTM → define.xml.
- Automation: Machine-readable standards reduce manual programming and validation.
- Regulatory Confidence: Real-time readiness for FDA/EMA submissions.
- Interoperability: Aligns across multiple CDISC standards (SDTM, ADaM, SEND).
- Consistency: Biomedical Concepts ensure uniform implementation across studies.

CHALLENGES IN TRANSITION

- Tooling Upgrades: Existing validation and transformation tools may need reengineering.
- Learning Curve: Teams must understand Biomedical Concepts and metadata workflows.
- Change Management: Adoption involves coordination among data management, programming, and biostatistics groups.
- Legacy Studies: Plans need to be drawn up for active or retired studies.

ORGANIZATION IMPLEMENTATION ROADMAP

1. CDISC Library APIs – Incorporate machine-readable metadata into pipelines.
2. Biomedical Concepts – Begin with frequently used concepts (e.g., vitals, labs).
3. Tools & Standards Repositories – Make sure internal libraries mirror SDTMIG 4.0 schema.
4. Pilot Studies – Conduct a limited series of studies with 4.0 to pilot and optimize processes.
5. Train Standards Teams – Spend money on cross-functional training in order to be acquainted with semantic metadata.
6. Engage with Regulators – Secure early cooperation with regulatory agencies to ensure compliant submissions.

CONCLUSION

SDTMIG 4.0 is a strategic step forward in clinical data standard evolution. By moving beyond dataset-convention to a metadata-oriented framework based on Biomedical Concepts and CDISC Library integration, the standard supports unmatched automation, traceability, and interoperability. It is challenging to implement from a tooling and change management perspective, yet ultimately there are fundamental long-term benefits for data quality, regulatory compliance, and efficiency. Companies that push beyond SDTMIG 4.0 early will be very well situated for the future of metadata-enabled clinical research.

REFERENCES

www.cdisc.org
www.wiki.cdisc.org
Study Data Tabulation Model Implementation Guide (SDTMIG): Human Clinical Trials, Version 3.4

ACKNOWLEDGMENTS

I sincerely thank all my colleagues who reviewed the paper and provided valuable comments and suggestions

CONTACT INFORMATION

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

Author Name: Binitha Sathyabhama
Company: Pfizer Healthcare India Pvt. Ltd.
Address: Anna Salai, Chennai – 600 006, India.
Work Phone: +91-44-61568000
Email: binitha.s@pfizer.com
Website: www.pfizer.co.in

Author Name: Venkateswarlu Layam
Company: Pfizer Healthcare India Pvt. Ltd.
Address: Anna Salai, Chennai – 600 006, India.
Work Phone: +91-44-61568000
Email: Venkateswarlu.Layam@pfizer.com
Website: www.pfizer.co.in