Paper DS02

# CDISC 360i Demo: Automating from Study Design to Submission

Darren Moreland, CDISC, Boston, US

## ABSTRACT

Clinical studies face challenges throughout their lifecycle due to time-consuming manual processes, operational data silos, gaps across standards, and disparate systems. The CDISC 360i initiative aims to address issues in implementing data standards with the addition of a semantic layer to enable automation of the study lifecycle from digital protocol to submission datasets to analysis results. This session demonstrates automation of the clinical data lifecycle beginning with the Unified Study Definitions Model (USDM) through SDTM. Attendees can observe how the 360i initiative leverages open-source software and uses standards with interoperable metadata to enable automation, enhance data integrity and data quality, and accelerate innovation.

## INTRODUCTION

Clinical research continues to experience inefficiencies in the way study definitions, operational metadata, and clinical data are designed and managed throughout the study lifecycle. Although CDISC standards provide structured models for representing study design and clinical data, implementation practices often vary across organizations. Protocol information is frequently transcribed manually, metadata is duplicated across multiple systems, and interpretation of standards can be inconsistent. These issues introduce risks to data integrity and fidelity, limit end-to-end traceability, and increase the effort required to generate SDTM and ADaM datasets.

The Unified Study Definitions Model (USDM) was developed to address these challenges by offering a machine-readable, structured representation of study design. As a single authoritative source, USDM enables more reliable exchange of study definitions between systems and supports consistency in clinical planning and execution. However, USDM alone is not sufficient to achieve seamless end-to-end automation. To fully leverage structured study definitions, downstream processes must be capable of interpreting the metadata and transforming it into operational and submission-ready outputs.

The CDISC 360i initiative was launched to bridge this gap by operationalizing CDISC standards and enabling automation across the study lifecycle - from digital protocol through SDTM, ADaM, and analysis. A key element of the initiative is a semantic layer that links protocol-level concepts to downstream structures, creating a shared, interoperable metadata layer across USDM, Biomedical Concepts, Data Set Specializations, and SDTM structures. Volunteer-led development has produced utilities for interpreting structured protocol content, generating the Schedule of Activities, applying SDTM mappings, and validating outputs using existing conformance frameworks.

Phase 1 of 360i focused on demonstrating an end-to-end automated pipeline that transforms a USDM-compliant study design document into SDTM-ready outputs. Several independently developed open-source components were integrated to form a cohesive workflow, illustrating how standards-based metadata can drive automation and reduce manual effort. This paper introduces the integrated solution and demonstrates how CDISC standards, combined with open-source tooling, can streamline clinical data processes, strengthen traceability, and establish a foundation for extending automation into ADaM and analysis result sets in future phases. This paper reports on Phase 1 outcomes - automation from USDM to SDTM - and establishes a foundation for Phase 2 work extending into ADaM and analysis results.

## APPROACH

### Objective

Phase 1 of CDISC 360i demonstrated an automated, standards-driven workflow that transforms a USDM JSON study design into SDTM-ready outputs by integrating independently developed open-source components into a cohesive pipeline. The approach emphasizes the USDM JSON as the single source of truth for study design and as the primary enabler of downstream automation.

**USDM as the Blueprint**
The workflow begins with a machine-readable USDM JSON file capturing structured study elements - arms, visits, activities, interventions - and linkages to Biomedical Concepts (BCs) and Data Set Specializations (DSS). This authoritative representation eliminates repeated manual transcription and supports consistent interpretation across tools.

**Semantic Layer and Interoperable Metadata**
A semantic layer binds protocol-level concepts in USDM to downstream SDTM structures, enabling reuse of interoperable metadata across components rather than re-specification in each system. This underpins early asset generation (e.g., Trial Design domains, define.xml), CRF generation (including annotated CRFs), and SDTM shell dataset creation directly from standardized definitions.

**Metadata Integration**
Volunteer teams developed purpose-built tools (e.g., for USDM JSON generation, SDTM shell creation, conformance checks) and integrated them into a single sequence driven by the USDM JSON. The integration focused on aligning assumptions and exchanging metadata consistently so that each stage could pass complete, standards-conformant inputs to the next with minimal manual intervention.

**Operational Data Store (ODS) and Execution Environment**
Traceability and persistence were provided by an ODS comprising Google object storage and a SQLite relational database to store raw clinical data, the study definition, and all runtime outputs. Python and R were selected for the open-source implementation; R (via {sdtm.oak}) generated SDTM domains, and notebooks were used to make the automation executable and transparent for users (e.g., the initial demonstration in Google Colab).

**Conformance and Quality**
The CDISC Open Rules Engine (CORE) was embedded to validate both the USDM and SDTM outputs against the open conformance rules. Integrating CORE validation at multiple stages surfaced structural or metadata issues early in the transformation lifecycle when easiest to fix with the least impact on downstream artifacts, reinforcing the USDM JSON as the upstream quality gate.

**Outcome**
The integrated, metadata-driven approach shows that a properly structured USDM JSON can power an automated pathway from study design to SDTM with embedded validation, creating a scalable foundation for extending automation into ADaM and analysis result sets in future phases.

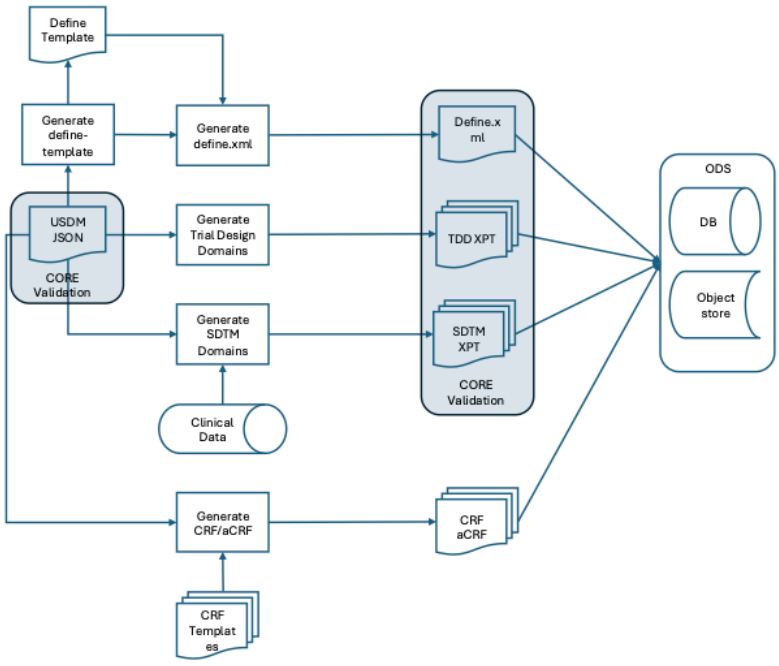


*Figure 1 – high-level overview of the data lineage from the USDM JSON through early assets and SDTM shells to validated SDTM datasets persisted in the ODS*

## METHODS

#### Author and Validate the Study Definition (USDM JSON)

The study design was authored in the Unified Study Definitions Model (USDM) using a machine-readable JSON representation that captured arms, visits, activities, interventions, and links to Biomedical Concepts (BCs) and Data Set Specializations (DSS). The USDM JSON served as the authoritative source for all downstream processes. Conformance checks were applied early using automated validation to ensure that the study definition was structurally sound and aligned with standard specifications before additional assets were generated. The Study Definition Workbench was used to author the USDM JSON for the CDISC LZZT pilot study, which then served as upstream input to all subsequent automation steps.

#### Derive Early Study Assets from USDM

Several early-phase outputs were generated directly from USDM metadata, reducing manual transcription and duplication:

- Schedule of Activities (SoA) and Trial Design domains were derived from the structured visit and activity elements.
- define.xml was generated by combining USDM metadata with standard codelists and template structures.
- CRFs and annotated CRFs were created using BCs and DSS metadata to align collection forms with SDTM variable expectations.
- SDTM empty shell datasets were generated to define dataset structures prior to the availability of clinical data.

#### Prepare Data and Metadata in the Operational Data Store (ODS)

An Operational Data Store composed of Google object storage and SQLite was used to persist all study assets and runtime outputs, including the USDM JSON, raw data, CRFs, SDTM shell datasets, define.xml, and final SDTM datasets. Storing all assets in a unified repository supported traceability and enabled a consistent metadata flow throughout the pipeline.

#### Generate SDTM Datasets via Orchestrated Open-Source Tools

Python scripts were used to orchestrate workflow execution, manage metadata exchange, and coordinate tool interactions. SDTM datasets were produced using R - specifically the {sdtm.oak} package - which applied transformation rules derived from BCs and DSS metadata. The pipeline was implemented and demonstrated using notebook-based execution to maintain transparency and reproducibility.

#### Validate Outputs and Iterate

The generated SDTM datasets were evaluated using CDISC CORE, which assessed conformance with CDISC standards. Validation findings were stored in the ODS, enabling iterative refinement of metadata and transformations. Structural or semantic inconsistencies, once identified, can be used to update upstream metadata in the USDM JSON or DSS libraries prior to pipeline re-execution.

#### Govern and Collaborate

Volunteer teams collaborated through GitHub using version control and issue tracking to coordinate development. Each component was built and maintained independently, allowing parallel progress while ensuring interoperability through adherence to shared metadata specifications. Technologies were selected based on openness, maturity, and long-term maintainability to support community adoption.

## RESULTS

Phase 1 of the 360i initiative delivered a working, automated pipeline that integrates independently developed components to transform a USDM study design and raw study data into SDTM-ready outputs with embedded conformance checking. The initial demonstration was implemented as a Google Colab notebook to provide a transparent, executable walkthrough of the end-to-end process.

When executed against the LZZT pilot study USDM JSON, the integrated workflow generated early study asset and SDTM deliverables, orchestrated via open-source Python/R tooling and validated using CORE. The execution confirmed that the USDM JSON can drive downstream asset creation with minimal manual intervention by passing interoperable metadata between tools in a consistent sequence.

Across a single execution, the pipeline produced the following artifacts and made them available in multiple formats to support review, reuse, and submission preparation:

- Trial Design Domains derived from structured study design and visit/activity metadata
- Define-XML documenting structures and codelists aligned to USDM-driven metadata
- Case Report Forms (CRFs) and annotated CRFs (aCRFs) generated from Biomedical Concepts and Data Set Specializations
- SDTM empty shell datasets to establish variable structures prior to population
- SDTM submission datasets produced from the raw clinical data using R-based transformations

To ensure traceability and downstream usability, the Operational Data Store (ODS) retained both intermediate and final assets in standard, inspection-ready formats:

- Raw clinical data in Parquet format
- CRFs/aCRFs in ODM-XML, HTML and Dataset-JSON
- SDTM datasets in CSV, XPT and Dataset-JSON formats
- Empty SDTM shells for each domain in Dataset-JSON format
- Conformance reports produced by embedded CORE validation
- Relational database tables (SQLite) storing raw clinical data and SDTM data for lineage and reproducibility

The use of open-source technologies enabled straight-forward integration of separately developed components into a cohesive pipeline. The resulting workflow automated tasks from structured protocol to SDTM outputs, demonstrating low manual intervention and reproducibility through notebook execution.

Embedding CDISC CORE within the pipeline provided immediate feedback on structural and semantic conformance for both the USDM study definition and the generated SDTM datasets. Findings were captured in the ODS, enabling rapid iteration on metadata prior to reruns and reinforcing USDM as the upstream quality gate.

#### Lessons Learned

Execution confirmed that a well-formed USDM JSON can drive early asset generation and SDTM transformation with minimal manual intervention, provided interoperable metadata are consistently applied across tools. Early, embedded CORE checks reduced downstream rework by surfacing structural and semantic issues prior to SDTM generation. These observations inform Phase 2 priorities, including expanded metadata mappings to support broader CRF derivation and downstream analyses.

By not prescribing a set of tools or technologies and focusing on solutions over standards, the 360i initiative ensured the projects could be developed independently. CDISC standards are tool-agnostic, so emphasizing the use of Open-Source Software made the projects more accessible to a wider audience encouraging collaboration and innovation.

## THE FUTURE

The needs discovered during Phase 1 informed the creation of a Schedule of Activities Workbench utility and highlighted additional metadata mapping requirements to support CRF generation from standards, which has led to the early drafting of the Define-JSON solution.

Phase 2 will extend the pipeline beyond SDTM to ADaM and analysis results (ARS), leveraging the metadata resources established in Phase 1. Biomedical Concepts will be augmented with Analysis Concepts and Derivation Concepts to support further downstream automation. Work will focus on broadening metadata coverage, defining downstream analysis concepts, and linking these assets into the existing workflow. The technology choices made in Phase 1 support continued development and integration for this expansion. In addition, using the USDM JSON study designs provides the opportunity to compare different versions of the study design to support creation of impact analyses for protocol amendments.

## CONCLUSION

Phase 1 of the CDISC 360i initiative demonstrated an automated pathway from a USDM-compliant digital protocol to SDTM-ready datasets by integrating independently developed open-source components into a cohesive workflow. This work showed that structured, interoperable metadata - anchored by the USDM JSON as the single source of truth - can drive generation of early study assets (e.g., Trial Design domains, define.xml, CRFs/annotated CRFs, SDTM shells) and enable reproducible transformation of raw data into SDTM submission datasets with minimal manual intervention.

Embedding CDISC CORE validation throughout the pipeline provided immediate feedback on structural and semantic conformance for both the USDM study definition and the resulting SDTM datasets. Findings persisted in the Operational Data Store supported iterative refinement and reinforced the role of the USDM JSON as the upstream quality gate for downstream automation.

The use of open-source technologies played an enabling role in Phase 1 by providing accessible, transparent tools that could be integrated and extended by volunteer contributors. Their inclusion supported reproducibility and collaboration without prescribing specific implementations, allowing the focus to remain on demonstrating how standards-based metadata can drive automation across the study lifecycle.

The outcomes of Phase 1 establish a practical foundation for extending automation beyond SDTM to ADaM and analysis results in future work. Priorities for the next phase include broadening metadata coverage, enhancing mappings to support CRF derivation and downstream analyses, and linking these assets into the existing standards-driven workflow. The technology approach and governance model adopted in Phase 1 are suitable for continued development and integration as these capabilities expand.

## REFERENCES

| **Utility** | **Repository URL** |
|---|---|
| CDISC CORE Rules Engine | https://github.com/cdisc-org/cdisc-rules-engine |
| CRF creation | https://github.com/lexjansen/cdisc360i-pocs |
| Define-XML creation | https://github.com/swhume/template2define |
| Define-XML template creation | https://github.com/dostiep/360i |
| Google Colab Notebooks | https://github.com/cdisc-org/cdisc-360i-notebooks |
| Open Study Builder | https://gitlab.com/Novo-Nordisk/nn-public/openstudybuilder/OpenStudyBuilder-Solution |
| Raw subject data | https://github.com/alidootson/UpdatedCDISCPilotData/tree/main/UpdatedCDISCPilotData/CDASH |

| | |
|---|---|
| Study Definition Workbench | https://github.com/data4knowledge/study_definitions_workbench |
| Study USDM documents | https://github.com/cdisc-org/360i/tree/main/data/protocol/LZZT/usdm |
| Trial Design Dataset creation | https://github.com/pendingintent/cdisc-usdm-utils |
| USDM validation utility | https://github.com/pendingintent/cdisc-json-validation |

**ACKNOWLEDGMENTS**

The underlying development work for this paper could not have been created without the contribution and support of CDISC developers, members and volunteers. Without their assistance and dedication to the CDISC 360i initiative, the goal of achieving automation through CDISC standards could never have been achieved.

**CONTACT INFORMATION**

Darren Moreland
CDISC
Director, Digital Transformation
dmoreland@cdisc.org
www.cdisc.org