Paper AD11

E2E Clinical Metadata Journey:

'Single Source of Truth' to Accelerate Clinical Research – From Collection and Tabulation to Analysis and Reporting

Indumathi Narasetty, AstraZeneca, Bangalore, India

ABSTRACT

As clinical trials grow in complexity, robust metadata management is increasingly important; however, many organizations still encounter fragmented and inconsistent metadata practices, which undermine data discovery, hinder compliance with standards, and restrict automation in processes like study setup, data cleaning, and analysis.

This paper presents an **E2E clinical metadata journey** as a transformative approach to harmonizing the collection, management, and utilization of clinical metadata from the Protocol to Submission. At the heart of this journey is the **Clinical Data Standards Metadata Repository (CMDR)**, which serves as a centralized and trusted repository for both foundational standards and study-specific metadata. The paper details the implementation of CMDR, including the establishment of standards metadata repository and the creation/ongoing management of machine-readable study metadata that facilitates smooth integration to downstream systems. This comprehensive framework ensures that each phase of the clinical trial process remains transparent, reproducible, and fully compliant with regulatory requirements across the entire trial lifecycle.

INTRODUCTION

In the evolving landscape of clinical research, robust metadata management underpins rapid decision-making, preserves data integrity, and ensures compliance with global regulatory standards. However, the lack of machine-readable data specifications has impeded the automation of critical workflows such as study construction, data validation, and analysis. Compounding this, traditional metadata repositories often fall short in scalability and adaptability, leading to fragmented processes and complex integration hurdles. Recognizing these limitations, our organization identified the need for a streamlined **MetaData Repository MDR**, capable of automated, robust integration with multiple downstream systems.

This implementation advances metadata governance through automation, granular versioning, and robust compliance frameworks. We detail the architecture and rollout of a standards-based metadata repository with dynamic management of machine-readable study instance metadata spanning Collection (RAW), Tabulation (SDTM), and Analysis (ADaM & TLFs) domains. The cloud-native, GxP-validated MDR platform incorporates comprehensive audit trails and secure, rule-based electronic approvals. Its API-centric design enables deep integration with EDC, statistical programming, and reporting ecosystems, facilitating metadata exchange and validation. As a result, the MDR not only strengthens operational agility and traceability but also minimizes manual intervention and expedites regulatory submission workflows.

BACKGROUND

Several organizations across the industry are either in the process of creating MetaData repositories (MDRs) or have already deployed first-generation solutions. These initiatives reflect a common recognition: clinical data standards metadata - spanning protocol elements, CRFs, controlled terminology, SDTM/ADaM standards, derivations, and submission artifacts - must be governed and traceable to ensure quality, speed, and compliance across the clinical trial lifecycle. However, traditional metadata repositories and document-centric practices have struggled to deliver the required end-to-end visibility and operational agility, particularly as portfolios scale and change velocity increases.

Traditional MDRs typically exhibit several limitations that hinder traceability and automation:

Fragmentation across multiple systems:

Metadata is distributed among protocol authoring tools, EDC libraries, standards repositories, mapping/transformation platforms, programming codebases, and submission packaging solutions. Each system maintains its own schema, audit trail, and versioning, with minimal cross-system linkage. This results in duplicated definitions, local variants, and "shadow repositories".

Siloed stakeholder workflows: Clinical Operations, Data Management, Biostatistics, Statistical Programming, Regulatory, and Standards teams work in parallel with limited real-time visibility. Changes in one area (e.g., a CRF item or derivation rule) do not reliably propagate downstream, increasing the risk of misalignment and rework.

Opaque lineage and limited impact analysis: Traditional repositories often capture metadata as static documents or semi-structured records. Lineage from protocol to CRF to SDTM to ADaM to TLFs and submission datasets is incomplete or manual, making it hard to assess the impact of changes, justify deviations, or demonstrate inspection-ready traceability.

Integration constraints: Legacy tools offer heterogeneous APIs (or none), rely on batch exports, and lack eventing. Synchronization is brittle, leading to stale metadata, inconsistent states between environments, and limited automation opportunities.

In response to these challenges, our MetaData repository (MDR) is designed to address fragmentation, enable automation, and provide robust integrations with downstream systems, with an emphasis on end-to-end traceability and governed change control.

CLINICAL DATA STANDARDS MDR IMPLEMENTATION JOURNEY

1. VENDOR SELECTION

Category	Summary
Objective	Conduct market scan, evaluate options, and select an MDR solution.
Scope	Identify a Clinical Data Standards Metadata Repository (MDR) for standards management, study-level automation, and integrations.
Stakeholders	IT, IT architecture, Statistical Programming, Data Management, and the CDS team.
Pre-screening (RFI)	~15 vendors assessed for capabilities, roadmap maturity, interoperability (APIs/eventing), security/compliance, and TCO.
Shortlist	6 vendors invited for RFP and live demos.
Scoring	Standardized framework with numeric ratings plus qualitative comments.
Key Criteria	CDISC alignment, rule formalization, lineage/impact, versioning/approvals, integration (bi-directional APIs), identity/permissions, usability, scale performance, vendor viability, implementation risk.
Finalists	3 vendors progressed to sandbox evaluations.
Sandbox setup	Loaded CRF standards, SDTM/ADaM libraries, controlled terminology, and sample study assets to test fit, API behavior, and change orchestration.
Decision process	Consolidated scores and feedback for Programme Board review with pros/cons across capabilities, risks, and operating model fit.
Outcome	Selected a vendor and adopted a two-phased MDR implementation for dynamic metadata management.

Table 1.1 Vendor Selection Process

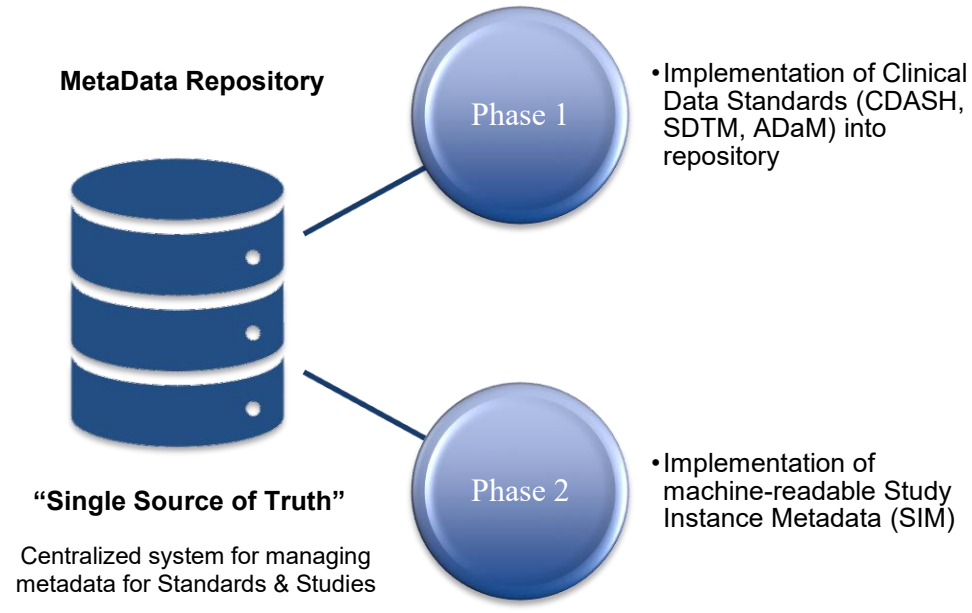


Figure 1.1 MetaData Repository (MDR), Phases

2. PHASE 1 - STREAMLINING STANDARDS

Phase 1 delivered as the product's major release v1.0 and Minimum Viable product (MVP) - centered on implementing and maintaining clinical data standards within the repository to establish consistent metadata governance and harmonize data collection and standardization.

It streamlines standards onboarding and governance (CRF components, SDTM/ADaM libraries, controlled terminology), enforces disciplined naming and global identifiers, and establishes automated conformance checks, hierarchical versioning, and lineage from standards to downstream assets.

Operationally, Phase 1 prioritized enhanced maintenance of standards metadata libraries for data collection (CDASH), tabulation (SDTM), and analysis (ADaM), including creation of a **Study Instance Metadata (SIM) Register (A SIM Bank)** comprising standards components for CDASH, SDTM, and ADaM. We have also implemented Foundational Standards metadata linkage: RAW to SDTM mapping functionalities.

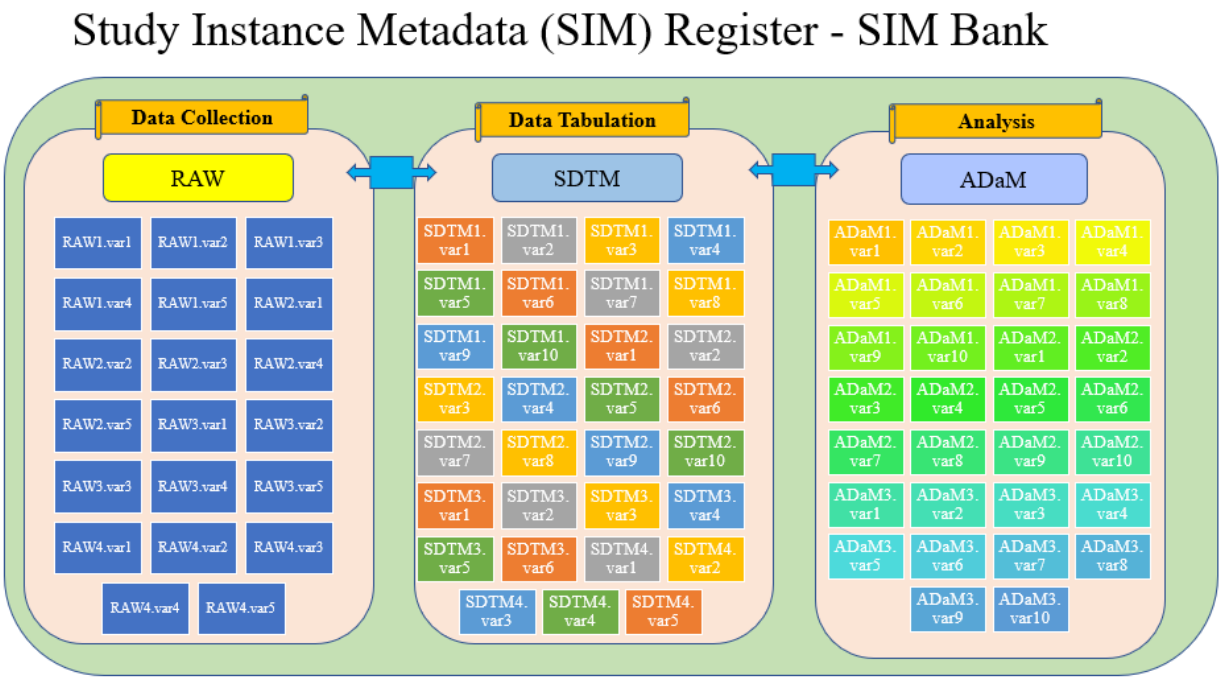


Figure 2.1 Study Instance Metadata (SIM) Register – SIM Bank

3. PHASE 2 - STUDY INSTANCE METADATA (SIM) IMPLEMENTATION

Phase 2 (SIM Study Register and study-level enablement):

- Phase 2 builds on the MDR's existing SIM metadata by introducing the SIM Study Register. This register creates and maintains machine-readable study metadata for data collection (CDASH), tabulation (SDTM), and analysis (ADaM).
- For each study, the register starts with metadata from the global standards libraries and then applies study-specific updates based on the study protocol & Statistical Analysis Plan (SAP). This enables study-level authoring and reuse of standards using clear, machine-readable rules.
- The study metadata defined in the SIM Study Register can be reused across multiple sister studies within the same compound. Teams can inherit common components and rules, make controlled adjustments where need, and maintain consistency and traceability across related studies.
- Phase 2 also delivers strong integrations. API-driven adapters (event-driven where available) connect the MDR to downstream systems such as EDC for CRF creation, SDTM/ADaM mapping services, SAS/R programming libraries, and analytics/TLF repositories.
- End-to-end traceability is maintained through unified audit trails, global identifiers, and cross-functional workflows. This covers the full lineage from protocol to CRF to SDTM to ADaM to TLF to submission, supports governed change control, and enables faster, more reliable impact analysis across the clinical trial lifecycle.

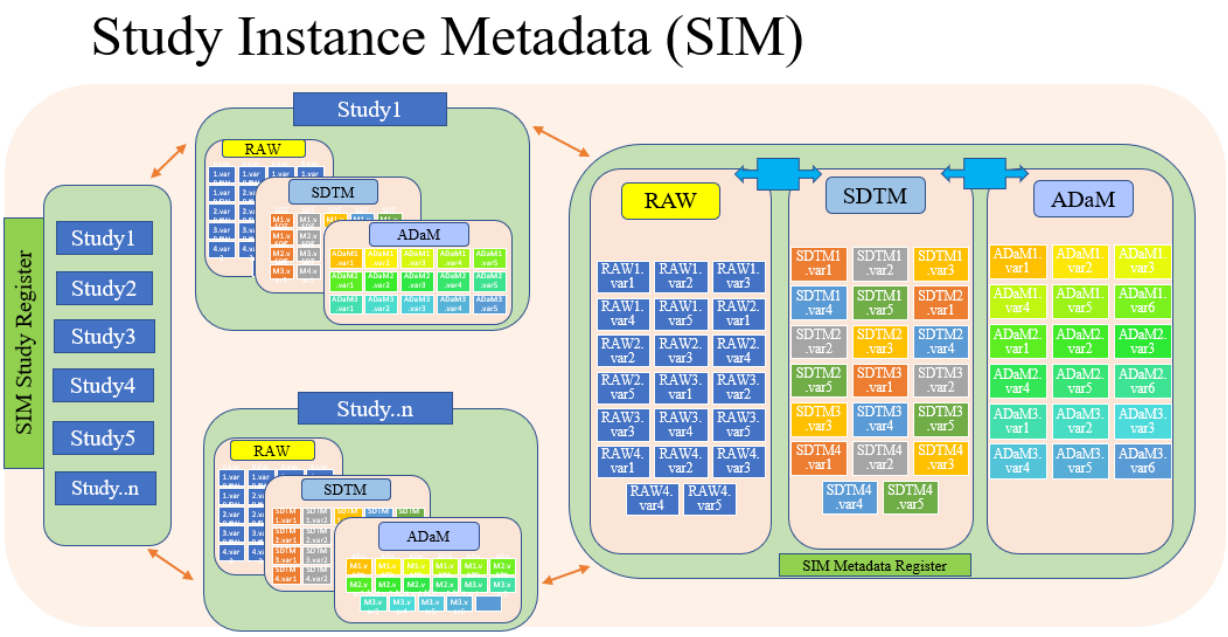
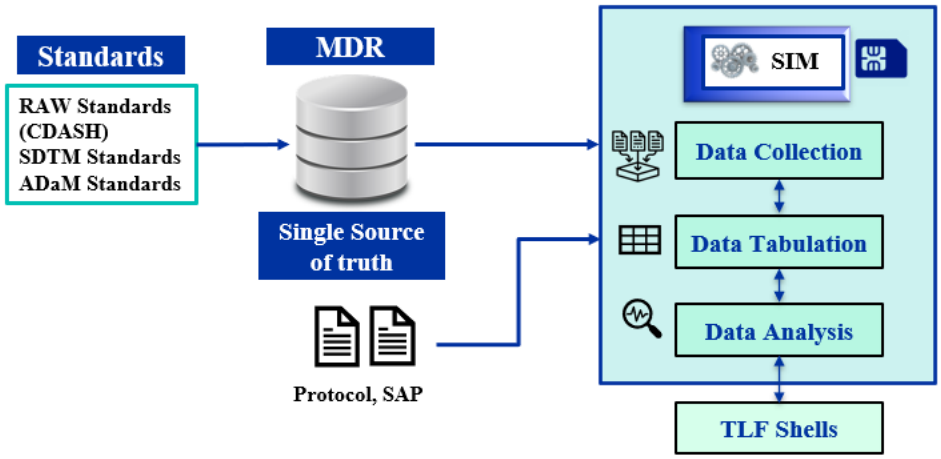


Figure 3.1 Study Instance Metadata (SIM)

4. WHAT IS STUDY INSTANCE METADATA (SIM)?

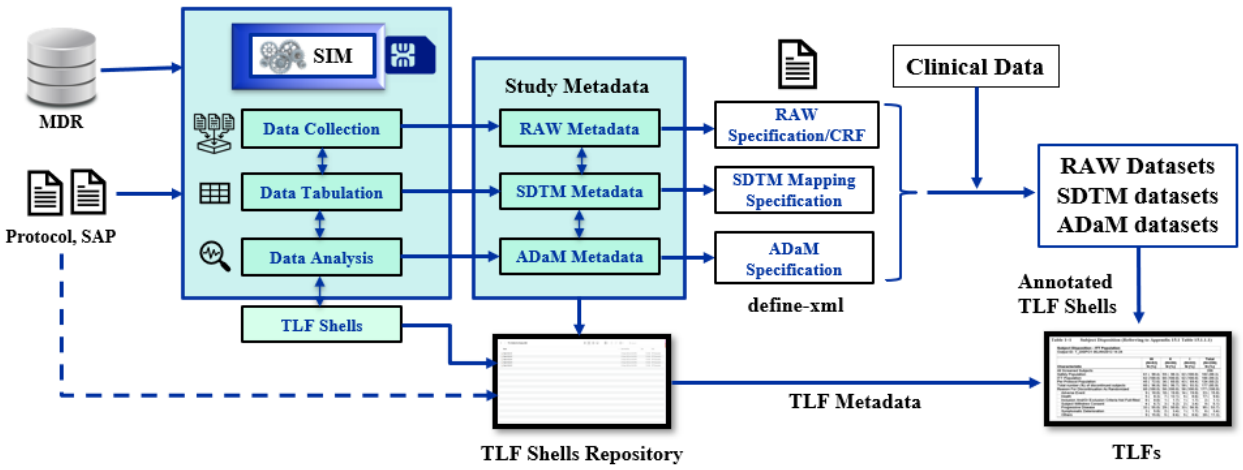
Study Instance Metadata (SIM): *The machine-readable study-specific metadata instantiated from enterprise standards* - capturing the selected forms, code lists, parameters, configurations, rules, mappings, and derivations that operationalize a particular trial from protocol through data collection, SDTM/ADaM, analysis, serving as the authoritative single source of truth for that study.

5. SIM FLOW



MDR – MetaData Repository; SAP – Statistical Analysis Plan; SIM – Study Instance Metadata; CDASH - Clinical Data Acquisition Standards Harmonization; TLF – Tables, Listings, Figures; SDTM – Study Data Tabulation Model; ADaM – Analysis Data Model;

Figure 5.1 SIM FLOW



MDR – MetaData Repository; SAP – Statistical Analysis Plan; SIM – Study Instance Metadata; CDASH - Clinical Data Acquisition Standards Harmonization; TLF – Tables, Listings, Figures; SDTM – Study Data Tabulation Model; ADaM – Analysis Data Model; CRF – Case Report Form;

Figure 5.2 SIM FLOW Contd...

Study Instance Metadata (SIM) is the study-specific instantiation of enterprise standards curated in a governed Metadata Repository (MDR). The MDR is the single source of truth for CDASH, SDTM, and ADaM standards, controlled terminology, and derivation rules. For each study, teams select the relevant standard packages and and then applies study-specific updates based on the study protocol & Statistical Analysis Plan (SAP), which produces machine-readable specifications for the three components: data collection, data tabulation, and data analysis, plus TLF shells.

From SIM, collection specifications configure CRF & non-CRF modules and code lists aligned to CDASH. Tabulation specifications define mappings from collected clinical data (including non-CRF sources) to SDTM domains and variables with value-level metadata and constraints. Analysis specifications describe ADaM dataset structures, derivations, analysis flags, and linkages to TLF shells. All specifications are versioned and traceable back to their originating MDR standard objects.

The MDR/SIM model generates define-XML-ready specifications for SDTM and ADaM, ensuring submission readiness. Integration services automatically publish these specifications to downstream systems: EDC libraries are provisioned for CRF build; transformation pipelines generate RAW and SDTM datasets; and statistical environments consume ADaM specifications. SIM identifiers and lineage are retained across systems to keep environments aligned and support inspection.

TLF production combines two governed metadata sources. ADaM metadata from SIM provides the analysis datasets and rules, while study TLF metadata from the TLF Shells Repository supplies display titles, footnotes, populations, sorting/grouping, and variable selections. Together, these drive automated generation of Tables, Listings, and Figures with programmatic links back to ADaM datasets, enabling end-to-end traceability from standards through submission artifacts.

6. SIM LIFE CYCLE & VERSIONING

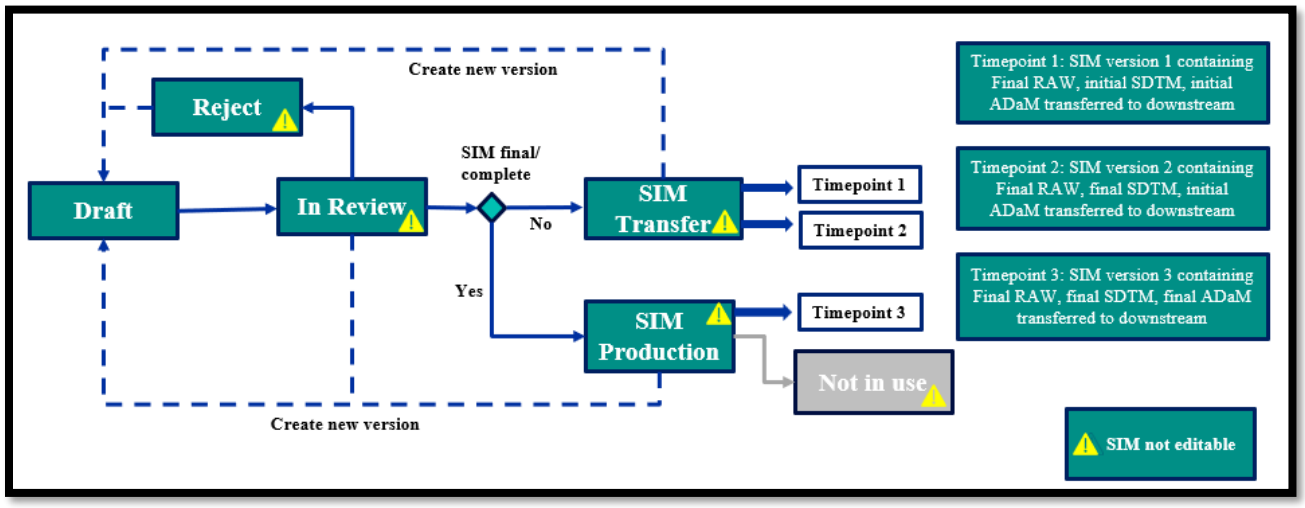


Figure 6.1 SIM Life Cycle

The above figure depicts the typical stages in the SIM life cycle. Each SIM version can only have one status at a time.

Below is a summary of the SIM version status definitions, whether the version is editable in each status, and the key benefits associated with each stage.

Status	Definition	SIM Version Editable ?	Benefits
Draft	1) Status upon initial creation of SIM, when it is editable 2) Status each time a new version of a SIM is created for updates	Yes	SIM is editable and RAW, SDTM, ADaM metadata can be added/updated
In Review	SIM can be reviewed by SIM support team/cross functional/external stakeholders	No	SIM is non-editable and can be reviewed Next SIM version can be created to implement updates
SIM Transfer	SIM in preliminary/initial version is transferred to downstream systems	No	Downstream systems can receive initial SIM metadata before final SIM is available
SIM Production	SIM in its final version is transferred to downstream systems	No	Downstream systems receive final SIM
Reject	SIM updates have been rejected	No	Incorrect/Outdated SIM versions can be transitioned to “Reject” to avoid accidental transition of SIM to “SIM Transfer” or “SIM Production” status triggering the export of the SIM
Not in use	Users should never transition any SIM version to this status “Not in use” for any study	No	Users should never transition any SIM version to this status “Not in use” for any study because reusability cannot be guaranteed

Table 6.1 SIM Life Cycle Definitions, Benefits

Initial status (Draft): SIM setup starts in **Draft**, enabling edits across all components (**RAW**, **SDTM**, **ADaM**).

Versioning for updates: Each set of changes creates a **new Draft version** to continue editing without altering prior versions.

Stakeholder review (In Review): When ready for review (e.g., **Data Managers** for collection; **Statistical Programmers/Statisticians** for SDTM/ADaM), SIM status will be pushed to **In Review**. The SIM is **locked**; if changes are needed, a new draft to be created

Change tracking: We use **Compare reports** to visualize differences between versions and rely on the **audit trail** (who/what/when) for full **traceability**.

Transfer at agreed timepoints (SIM Transfer): Post SIM review, the version to be pushed to **SIM Transfer** to automatically publish to EDC, SAS, and other downstream systems. Subsequent updates begin with a new Draft.

Finalization (SIM Production): When metadata content is final, we will push SIM to **SIM Production** to transfer the final version to downstream; any future milestone edits start as a new Draft.

Prevent accidental exports (Reject): Incorrect/Outdated SIM versions can be transitioned to "Reject" to avoid accidental transition of SIM to "SIM Transfer" or "SIM Production" status triggering the export of the SIM.

A new SIM version can be created from any SIM life cycle except **"Not in use"** status (i.e., In review, Reject, SIM Transfer, SIM Production).

Note: The timepoints and SIM versions shown in the figure are illustrative only. In practice, a study may progress through multiple versions before finalizing each SIM component, and there is no preset limit on the number of version creations, SIM transfers, or SIM production events.

7. HOW DID WE DO IT (IMPLEMENTATION OF MDR)?

We implemented the MDR using the Agile Scrum framework because requirements evolved during discovery. Scrum's iterative approach enabled short sprints, incremental delivery, continuous feedback, and rapid adaptation, reducing risk and improving efficiency.

Work was organized into time-boxed sprints and Product Increment (PI) planning cycles every 8 - 12 weeks. Each increment aligned cross-functional teams and stakeholders on objectives, user stories, dependencies, and delivery plans, resulting in committed PI objectives with clear timelines. Across the program, we delivered multiple GxP-validated releases and sub-releases, each adding functionality, enhancing performance, and expanding integrations with downstream systems.

Delivery was supported by defined roles:

- **Product Owner**: Sets product vision, prioritizes backlog, represents stakeholders, validates increments.
- **Scrum Master**: Facilitates, removes impediments, runs ceremonies (Daily standups, Sprint planning, reviews, retrospectives).
- **Developers** (engineering, QA/validation, data/integration, automation): Design, build, test, document, implement integrations, maintain automated checks, ensure **Definition of Done** and compliance.

This disciplined, feedback-rich approach consistently converted priorities into working, validated increments that met the Definition of Done, compliance obligations, and operational readiness.

Please see the detailed product release information below; all listed releases are GxP-validated.

Releases	Success Criteria	Dataset	Time taken for creation of Dataset (s)	Improvement
Release 1.0	**Minimum Viable Product (MVP) for the product**	**-**	**-**	MVP
Release 1.1	Bug Fixes, Integration Improvements	-	-	Testing Phase
Release 1.2	General Product Enhancements, Bug Fixes	-	-	Testing Phase
Release 2.0	RAW - SDTM Mapping functionality deployment	-	-	Testing Phase

Release 2.1	Mapping Bug fixes	AE ADAE	AE: 1.5 - 2 hours ADAE: 6 - 7 hours	-
Release 2.2	Plugin Updates, Bug fixes, Domain restructuring, Performance Improvement	AE ADAE	AE: 1.2 - 1.6 hours ADAE: 5 - 7 hours	~ 25%
Release 2.3	RAW Plugin Updates - Visualization enhancements, ADaM Plugin Updates	AE ADAE	AE: 1.2 - 1.6 hours ADAE: 5 - 7 hours	No performance improvement
Release 2.4	Mapper Plugin Updates related to mapping functionality enhancements	AE ADAE	AE: 1.2 - 1.6 hours ADAE: 5 - 7 hours	No performance improvement
Release 3.0	**SIM Technical Release, Performance Improvement, Integrations with few downstream systems**	AE ADAE	AE: 0.8 - 1.2 hour ADAE: 4 - 6 hours	~ 30% - 40%
Release 4.0	Performance Improvement, Integrations with few other downstream systems, Study Compare reports, Mapping enhancements, SDTM & ADaM Plugin updates for compatibility of define-xml submission readiness	AE ADAE	AE: 0.6 - 1 hour ADAE: 3 - 5 hours	~ 25%
Release 4.1	ADaM Plugin updates, ADaM Value Level Metadata content enhancement, Change Log integration enhancement etc.,	AE ADAE	AE: 0.6 - 1 hour ADAE: 3 - 5 hours	No performance improvements in this release

Table 7.1 Product Releases Information

KEY OUTCOMES

- Generation of RAW Data Specification
- define-xml ready Study Metadata Specifications - SDTM, ADaM, TLF Metadata
- RAW - SDTM Mapping Specification (Facilitating the Traceability from RAW - SDTM)
- Study Metadata - For RAW, SDTM, ADaM, TLF Metadata
- Generate compliant **define-xml** for submissions.

BENEFITS

- **Faster study builds:** Up to 30-40% reduction in setup time via reusable, standardized metadata
- **Stronger standards governance:** Improved maintenance of CDASH, SDTM, and ADaM libraries
- **Streamlined setup:** Efficient EDC build and smoother Analysis & Reporting (SDTM, ADaM specs, define-xml)
- **Machine-readable SIM:** Structured metadata enabling automation, consistency, and reuse
- **Accelerated sister-study setup:** Leverages the SIM study register to quickly replicate similar studies within the same compound
- **API-centric integration:** Seamless connectivity with downstream systems

- **End-to-end traceability:** Clear lineage from Protocol → eCRFs → SDTM → ADaM → Submission
- **Easier CDISC upgrades:** Faster adoption of newer CDISC versions

CONCLUSION

Our MDR journey demonstrates that a centralized, GxP-validated MDR can deliver robust automation while resolving integration challenges across the clinical data lifecycle. Automated governance, validation, and lineage tracking reduce manual effort, improve consistency, and accelerate workflows. An integration architecture interoperates with upstream standards libraries, study build tools, data management platforms, and analytics environments to enable reliable, scalable exchange of controlled metadata. Beyond automation and integration, the MDR addresses additional challenges described in the Background - harmonization of standards, versioning and change control, traceability, compliance and audit readiness, and cross-functional adoption - positioning the MDR as a resilient, reusable foundation for quality, shorter cycle times, and continuous improvement.

REFERENCES

- Transforming Biometrics: The Future of Statistical Programming Through AI-Powered Automation and Agile Scrum https://phuse.s3.eu-central-1.amazonaws.com/Archive/2025/SDE/APAC/Bengaluru/PRE_Bengaluru08.pdf
- Scrum.org

CONTACT INFORMATION

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

Author Name: Indumathi Narasetty
Company: AstraZeneca India Pvt Ltd
Address: Block D3, Manyata Embassy Business Park,
Outer Ring Road, Rachenahalli, Bengaluru, Karnataka, 560045
Work Phone: +91-9581229439
Email: indumathi.narasetty@astrazeneca.com
Website: https://www.astrazeneca.com