Paper DS04

AI-Powered Digital Standards Library for Accelerated Regulatory Submissions

Anshu Aman, Hexaware Technologies, Pune, India
Varun S, Hexaware Technologies, Chennai, India

ABSTRACT

Clinical trial programming suffers from fragmented workflows and decentralized specifications, leading to manual rework, data inconsistencies, and delayed regulatory submissions despite existing standards like CDISC. The absence of real-time data traceability and non-digitized metadata management further hinder efficiency and regulatory confidence.

This solution proposes an AI-enabled, cloud-native digital metadata repository (MDR) integrated with statistical programming automation. It leverages Machine Learning (ML) agents and API-based data ingestion to enable continuous data quality checks, real-time reconciliation, and automated generation of statistical outputs. The system dynamically interprets Statistical Analysis Plans (SAPs) into executable SAS/Python/R code, ensuring traceable and version-controlled standard implementation across SDTM, ADaM, and TLF generation.

This innovation significantly enhances GxP compliance, accelerates submission timelines, and increases confidence in regulatory deliverables. It transforms statistical programming into a more autonomous, quality-driven, and efficient process, supporting the industry's digital transformation towards resilience and innovation

INTRODUCTION

The global life sciences industry is undergoing a fundamental structural transition from a document-centric paradigm to a data-centric model of drug development and regulatory interaction. Historically, the process of bringing a novel therapeutic to market has required the manual generation and review of thousands of unstructured documents, a method increasingly incompatible with the complexity of modern medicine and the urgent need for rapid patient access to treatments. Currently, an estimated 90% of clinical trial protocols are still created and managed as unstructured text files, often exceeding one hundred pages. In the traditional Clinical Trial Information Flow, the documents are manually parsed by disparate teams to configure downstream systems such as Electronic Data Capture (EDC), Clinical Trial Management Systems (CTMS), and Trial Master Files (TMF). This manual transfer of data is not only prone to human error but also creates significant redundancy and a lack of interoperability, resulting in an average lag of four months between protocol approval and study start-up.

To address these systemic inefficiencies, the Digital Data Flow (DDF) initiative- a collaboration between the Clinical Data Interchange Standards Consortium (CDISC) and TransCelerate BioPharma, aims to modernize clinical trials by enabling a fully digital workflow through protocol digitalization. The core objective of DDF is to move away from traditional document models toward standardized, machine-readable representations of study protocols. This transformation is built upon the Unified Study Definitions Model (USDM), which provides a standardized logical data model (LDM) that defines the relationships between critical entities such as study objectives, endpoints, and the schedule of activities. By utilizing the USDM, organizations can establish a machine-executable "single source of truth" that allows protocol data to flow automatically and in parallel between systems without the need for custom transformations.

The implementation of a cloud-native digital metadata repository (MDR) or Study Definition Repository (SDR) is critical to this architecture, as it serves as a centralized middle layer between study definitions and study execution. An integrated MDR ensures that consistent standardized data definitions are applied across the entire trial lifecycle, from data collection to tabulation. Such a repository enables the automated creation of study assets, such as draft eCRF specifications and SDTM mapping metadata, which can reduce trial delays by up to 50% and lower integration and compliance costs by up to 40%. Furthermore, the MDR maintains immutable audit logs and persistent lineage identifiers, ensuring total traceability from the digitized protocol to submission-ready deliverables.

Artificial Intelligence (AI) and Machine Learning (ML) are paramount in enhancing the capabilities of these digital standards libraries by automating high-volume, knowledge-heavy tasks. AI-powered systems utilize Natural Language Processing (NLP) agents to parse digitized protocols and Statistical Analysis Plans (SAPs), mapping complex endpoints and derivations directly to CDISC standards. These libraries leverage NLP-driven field

interpretation to understand the underlying meaning of Case Report Form (CRF) labels even when phrased inconsistently, resulting in a 60-80% reduction in CDASH mapping time. Advanced organizations have achieved up to 95% automation in SDTM generation and 80% automation in ADaM variable creation by employing ML algorithms that generate executable SAS, R, or Python code. By integrating AI into the metadata governance framework, clinical data teams can transition from manual testing to designing AI-powered validation frameworks, effectively accelerating the journey from research to market while maintaining rigorous GxP compliance

This paper proposes an AI-enabled, cloud-native digital metadata repository (MDR) integrated with statistical programming automation. At the core of which is an AI-powered digital standards library, an integrated ecosystem designed to digitize clinical protocol data and automate the mapping of trial outputs to industry standards. By establishing a translational bridge between model-informed drug development (MIDD) and clinical implementation, these libraries turn regulatory evidence into actionable insights throughout the product lifecycle. This solution utilizes an AI-enabled, cloud-native digital metadata repository (MDR) integrated with statistical programming automation to enable continuous data quality checks, real-time reconciliation, and the automated generation of statistical outputs. The implementation of such a library aims to achieve 40-60% faster regulatory submissions while reducing compliance costs by replacing point-to-point custom integrations with scalable, cloud-native interoperability frameworks.

TECHNICAL ARCHITECTURE OF PROTOCOL DIGITALIZATION

The Unified Study Definitions Model (USDM) and DDF

The cornerstone of a modernized clinical trial is the digitalization of the study protocol using the Unified Study Definitions Model (USDM). Developed as part of the Digital Data Flow (DDF) initiative, the USDM provides a standardized logical data model (LDM) that utilizes Unified Modeling Language (UML) class diagrams to define critical relationships between study objectives, endpoints, and the schedule of activities. The USDM serves as the "single source of truth," ensuring that protocol data is interoperable across different systems without custom transformations, thereby reducing cycle times and improving data reliability. This architecture supports a vast range of clinical study components, including study designs, arms, epochs, detailed study logic, encounters, estimands, and inclusion/exclusion criteria. The USDM is designed to be implementable in various physical data stores, such as relational databases, property graph databases, and RDF triple stores, to support a uniform understanding of trial data elements.

The USDM framework is comprised of five key artifacts that ensure technical and regulatory alignment:

- **Logical Data Model (LDM):** A UML class diagram depicting data entities and their relationships, functioning as the foundational source of truth for study definition technologies.
- **API Specification:** A standardized RESTful API, specified using the OpenAPI Specification (OAS) version 3.0, which enables the high-velocity exchange of study definitions between machines.
- **Controlled Terminology:** Standardized code lists for protocol elements (e.g., sex, race, lab tests) developed by CDISC to ensure semantic consistency across the entire trial lifecycle.
- **Implementation Guide (IG):** Comprehensive documentation providing solution architects with reusable designs, industry best practices, and conformance rules.
- **Study Definition Repository (SDR):** A centralized middle-layer component built on the USDM architecture to facilitate automated study asset creation and system configuration.

The Study Definition Repository (SDR) as a Functional Middle Layer

The Study Definition Repository (SDR) serves as a centralized middle layer between study definitions/protocols and study execution systems. Built on the USDM architecture, the SDR enables high-velocity data flow from the digitized protocol to downstream systems via RESTful API specifications. By mapping USDM elements to the Operational Data Model (ODM), the SDR can automatically generate draft electronic Case Report Form (eCRF) specifications, which significantly reduces the manual effort and time required for trial builds. The SDR is cloud-agnostic and open-source, serving as a functioning template to connect systems that produce protocol information with those that consume it for clinical operations. It supports multiple media types, including JSON and XML, and provides versioning and immutable audit trails to maintain data provenance throughout the trial.

Semantic Integration through Biomedical Concepts (BCs)

A critical feature of the SDR architecture is the integration of Biomedical Concepts (BCs). These concepts provide a standard way to define data elements for clinical data collection, ensuring that the same definitions are applied

consistently from initial data capture in the EDC to final tabulation in SDTM. The model includes a biomedical concept layer designed to aid in CRF automation, which creates a seamless lineage that supports the scientific validity and ethical integrity of the trial. By utilizing BCs, organizations can ensure that SDR knows exactly which concepts are referenced from specific forms and domains, allowing for near-total automation in the generation of mapping metadata.

Global Standards Alignment and ICH M11 Integration

The technical architecture of the digital standards library has evolved to support global harmonization efforts, specifically through alignment with the Clinical Electronic Structured Harmonized Protocol (CeSHarP) M11 initiative. The release of USDM version 4 in 2025 marked a milestone by achieving alignment with the ICH M11 technical specification, which governs the interoperable electronic exchange of protocol content across global regulatory jurisdictions. This alignment allows the entire protocol document to be held within the USDM, enabling the protocol to be generated directly from the model while accommodating both structured and unstructured elements. This transition from a document-centric to a data-centric paradigm allows for the reuse of document contents and sections across the research community, providing a foundation for advanced analytics and AI-driven efficiencies.

METHODOLOGY FOR AI-DRIVEN ORCHESTRATION AND INGESTION

Data Ingestion and Normalization The methodology integrates high-velocity API connections to aggregate heterogeneous sources, including EHRs, lab feeds, eCRFs, and registries like ClinicalTrials.gov. Advanced implementations utilize a data mesh architecture where ingestion patterns are grouped—for example, treating all file-based sources with similar ingestion logic to ensure scalability. Each incoming dataset is normalized and assigned a persistent lineage identifier to maintain traceability from raw data to submission-ready deliverables.

AI Intelligence Core and NLP Agents The solution utilizes Natural Language Processing (NLP) agents to parse SAPs and digitized protocols, mapping endpoints and derivations to CDISC standards. Hybrid engines, combining machine learning (ML) with rule-based logic, generate executable code in SAS, R, or Python. To maintain quality, the system employs a "human-in-the-loop" model where confidence scores determine whether outputs are automatically accepted or routed for expert review. This approach allows the AI to learn from previous study conventions and adapt to organization-specific therapeutic area nuances over time.

STANDARDS-DRIVEN PIPELINE FOR AUTOMATED DELIVERABLES

Automating CDASH, SDTM, and ADaM Mapping The library enforces a standards-driven pipeline where each stage-CDASH to SDTM to ADaM to Define-XML is automated through metadata-driven mappings. NLP-driven field interpretation allows the system to understand the meaning of CRF labels even when phrased inconsistently, resulting in a 60-80% reduction in mapping time for CDASH. For SDTM, organizations have achieved up to 95% automation in domain generation, while ADaM variable creation can be 80% automated, ensuring a clear lineage from analysis back to raw data.

Real-Time Quality Control and GxP Compliance Validation is embedded throughout the workflow, with automated Pinnacle21 checks and logged remediation loops. Machine learning algorithms detect programming errors and inconsistencies faster than manual review, enhancing the quality of regulatory dossiers. Technical controls, including immutable audit logs, electronic signatures, and version control, ensure compliance with 21 CFR Part 11 and EU Annex 11. This shift transforms clinical data teams from manual testers to designers of AI-powered validation frameworks.

REGULATORY ALIGNMENT AND FUTURE-PROOFING

The FDA 7-Step Credibility Framework To ensure regulatory confidence, the implementation aligns with the FDA’s risk-based credibility assessment framework. This 7-step process requires sponsors to define the specific regulatory question, characterize the context of use (COU), assess the model risk, and develop a plan to establish credibility through rigorous testing. Higher-risk AI applications, such as those used for patient risk stratification, require more extensive evidence and continuous lifecycle monitoring to address potential data drift.

Transition to eCTD v4.0 The library facilitates the transition to eCTD v4.0, which moves the submission paradigm from a document-centric to a data-centric model. AI plays a critical role in precisely tagging documents according to v4.0 metadata standards and managing unique Object Identifiers (OIDs), which is unfeasible to perform manually. Version 4.0 supports bidirectional messaging and document reusability across different applications, reducing the risk of technical rejections and accelerating interactions with health authorities.

CONCLUSION

The implementation of an AI-powered digital standards library represents a transformative shift in clinical trial data management and regulatory submissions. By digitizing protocols through the USDM and leveraging cloud-native metadata repositories, organizations can eliminate the persistent lag between study design and execution while ensuring total data traceability. The synergy of NLP-driven mapping and automated code generation not only reduces manual rework and costs but also enhances the scientific coherence and accuracy of regulatory dossiers. Successful adoption requires organizations to embrace data-centricity, prioritize high-velocity automation, and align with risk-based regulatory frameworks like the FDA's 7-step credibility assessment. Ultimately, this solution positions the life sciences industry to deliver life-saving therapies with greater efficiency, transparency, and clinical precision.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author at:
Author Name: Anshu Aman/Varun S.
Company: Hexaware Technologies
Address: Pune, Maharashtra
Work Phone: +91-9661927173
Email: anshua2@hexaware.com
Website: https://www.hexaware.com