eDataValidator (eDV): Validating Clinical Data for Confident Submissions

Abstract

Regulatory submissions for clinical and nonclinical studies increasingly require rigorous data quality, traceability, and compliance across multiple dataset versions. Although established processes exist for the development of SDTM, ADaM, and SEND datasets, validation is often conducted as a late-stage, standalone activity, which can introduce inefficiencies and increase regulatory risk. This paper describes a workflow-driven, continuous validation framework implemented using eDataValidator (eDV), a unified platform that integrates regulatory and CDISC validation into the dataset development lifecycle. By incorporating validation into routine workflows, the framework facilitates early identification of data quality and compliance issues, improves transparency across dataset versions, reduces manual rework, and supports automated generation of submission artifacts such as Define.xml and Study Data Reviewer's Guides. The approach aligns with regulatory expectations from agencies including the U.S. Food and Drug Administration (FDA) and Japan's Pharmaceuticals and Medical Devices Agency (PMDA), while improving operational efficiency in biometrics functions.

1. Introduction

Regulatory authorities such as the U.S. Food and Drug Administration (FDA) and the Pharmaceuticals and Medical Devices Agency (PMDA) rely on standardized clinical and nonclinical datasets to enable consistent and efficient regulatory review. Data standards developed by the Clinical Data Interchange Standards Consortium (CDISC), including the Study Data Tabulation Model (SDTM), Analysis Data Model (ADaM), Standard for Exchange of Nonclinical Data (SEND), and Define.xml, constitute the foundation of contemporary regulatory submissions.

Despite the widespread adoption of these standards, validation activities are frequently deferred until datasets are considered final. This late-stage validation paradigm often results in delayed identification of issues, repeated rework across dataset iterations, limited visibility into data quality trends, and increased risk during regulatory submission. In parallel, regulatory expectations increasingly emphasize consistency across dataset versions, traceability between datasets and submission documentation, and ongoing compliance throughout the data lifecycle.

This paper proposes a workflow-driven validation model that integrates continuous regulatory and CDISC validation throughout dataset development, from early-stage programming to final submission.

2. Limitations of Traditional Validation Practices

Traditional dataset validation practices typically involve executing validation tools a limited number of times near the conclusion of dataset development. While this approach may address fundamental compliance requirements, it presents several inherent limitations:

- Identification of validation issues occurs late in the development process, increasing the cost and effort required for resolution
- Dataset updates necessitate repeated manual revalidation and remediation
- Limited capability to track recurring issues across dataset versions
- Increased reliance on late-stage corrective actions, which may affect data consistency
- Elevated risk of regulatory questions or technical rejection

As study designs grow more complex and development timelines continue to compress, these limitations underscore the need for validation strategies that are more integrated, proactive, and scalable.

3. eDataValidator Platform Overview

eDataValidator is a unified validation platform designed to support continuous validation of clinical and nonclinical datasets intended for regulatory submissions, including Investigational New Drug (IND), New Drug Application (NDA), and Biologics License Application (BLA) submissions. The platform consolidates multiple regulatory and standards-based rule sets within a single validation workflow.

Validation support includes:

- FDA validation rules for SDTM and SEND
- PMDA validation rules for SDTM, ADaM, and Define.xml
- CDISC conformance rules for SDTM, ADaM, SEND, and Define.xml
- CDISC Open Rules Engine (CORE) rules
- Extended PointCross rules derived from FDA business rules and the FDA Technical Conformance Guide

Consistent application of these rule sets across dataset versions enables systematic monitoring of regulatory compliance and data quality throughout the development lifecycle.

4. Workflow-Driven Continuous Validation Framework

The proposed validation framework begins with dataset ingestion and extends through validation execution, issue resolution, quality monitoring, and documentation generation. The platform supports multiple input formats, including SAS XPORT, SAS7BDAT, Dataset-XML, CDISC Dataset-JSON (versions 1.0 and 1.1), CSV, and Excel.

Once datasets are ingested, all applicable validation rules are executed in a single automated process. Validation may be performed:

- At any stage of dataset development
- Across multiple dataset iterations
- Across multiple studies

This approach enables early detection of compliance and data quality issues, thereby reducing downstream remediation efforts and improving development efficiency.

5. Quality Monitoring and Issue Management

The framework incorporates integrated quality dashboards that provide visibility into validation results across domains and dataset versions. These tools support:

- Comparison of validation outcomes across successive dataset versions
- Identification and tracking of new, resolved, and recurring issues
- Drill-down analysis to individual records associated with validation findings
- Capture of study-specific annotations that may be incorporated into reviewer documentation
- Assignment and tracking of issue resolution activities

Such capabilities facilitate proactive issue management and enhance collaboration across programming, biostatistics, and biometrics teams, independent of the underlying statistical programming environment.

6. Automated Generation of Submission Documentation

In addition to dataset validation, the framework supports automated generation of key regulatory submission artifacts, including:

- Define.xml
- Clinical Study Data Reviewer's Guides (cSDRG)
- Nonclinical Study Data Reviewer's Guides (nSDRG)
- Analysis Data Reviewer's Guides (ADRG)

These documents are generated directly from validated datasets and associated validation metadata, ensuring internal consistency and traceability. Automation reduces manual documentation effort and minimizes the risk of discrepancies between datasets and supporting documentation.

7. Deployment Considerations and Operational Flexibility

The platform is available both as a cloud-based Software-as-a-Service (SaaS) implementation and as a locally installed desktop application. This deployment flexibility supports a range of organizational requirements, including collaborative validation workflows, offline processing, and varying data security constraints. The framework also allows for the incorporation of custom validation rules aligned with internal standards and organizational practices.

8. Practical Application and Observed Outcomes

The continuous validation framework has been applied to the validation of more than 10,000 regulatory submission datasets across SEND, SDTM, ADaM, and Define.xml. Use of the framework has been associated with the absence of technical rejections and nonconformance findings in supported submissions, illustrating the potential benefits of continuous, workflow-driven validation in regulatory data preparation.

9. Conclusion

A workflow-driven, continuous validation framework provides an effective approach for maintaining data quality and regulatory compliance throughout the clinical and nonclinical data lifecycle. By integrating regulatory and CDISC validation, quality monitoring, and automated generation of submission documentation within a unified environment, the proposed approach supports more efficient, transparent, and reliable preparation of regulatory submissions.