Paper AS02

# Transforming Study-Specific Reporting Trends Through an Innovative Data Pipeline Architecture Approach

Archana Nayak, Johnson & Johnson R&D, Great Britain
Jill Ophoff, J&J R&D, Belgium
Van Yper, Stefan, Associate, Luxemburg
Simon Walsh, J&J R&D, USA
Bhargava Reddy, J&J R&D, PA, USA

## ABSTRACT

In Analytical Risk-Based Monitoring, centralized real-time data reviews are vital for effective trial oversight. At J&J Innovative Medicine, dashboards with standard Key Risk Indicators are supplemented with customized Study Specific Reports tailored to the unique needs of each trial. Generating these reports promptly upon initial subject data receipt remains challenging.

To address this, we developed a re-purposeful data pipeline architecture that incorporates study-specific configuration files separately from the main pipeline, balancing ease of set up, flexibility and efficiency. This design employs a standard data structure, enabling near real-time mapping of raw data from external vendors and EDC systems.

We are currently testing this approach with a report on pharmacokinetic sampling compliance, utilizing a Biospecimen Events SDTM-like domain to detect potential sample collection errors. Preliminary results from a large Immunology Phase 3 trial are promising. This innovative pipeline enhances data analysis efficiency and ensures the timely availability of Study Specific Reports, driving better trial outcomes.

## INTRODUCTION

As clinical trials scale in complexity and global reach, timely identification of site-level data inconsistencies becomes critical. Study Specific Reports (SSRs), an internal J&J capability developed for Analytical Risk-Based Monitoring (ARBM), utilizes advanced statistical models and dashboards to monitor quality signals across data domains. These capabilities depend on reliable ingestion and transformation of data process historically requiring manual, study-specific customization.

### BACKGROUND AND MOTIVATION

Manual SSR creation workflows often include redundant mapping of domains, inconsistent test code definitions, and ad hoc report generation—all of which delay insights. Recognizing these inefficiencies, we proposed and developed a system where key metadata (e.g., test code identifiers, visit structures, and domain characteristics) are extracted via automated pre-processing scripts. These are compiled into study-specific configuration files used by downstream analysis pipelines.

### METHODOLOGY: DATA PIPELINE AND AUTOMATION

The pipeline includes components for eligibility filtering, metadata extraction, test code validation, and domain categorization. Python-based utilities extract and evaluate .xpt files from SDTM folders, converting them into human-readable tabular formats. Configuration files are reviewed during pre-run meetings, ensuring that clinical team input is encoded before analysis execution.

Key processing steps include distinguishing numeric vs. categorical variables, filtering test codes with inadequate visit data, appending differentiators to duplicated codes, and organizing domains into functional clusters (e.g., lab, questionnaire). Key processing steps include modelling to distinguish expected versus unexpected sampling, taking into account inadequate visit data and group and timepoint specific requirements. Due to higher computational challenges the processing is done in R, with resulting outputs being visualized via interactive dashboards, hence enabling efficient verification and traceability.

Any knowledge gained is used to correct expectations in a controlled feedback loop.

#### RESULTS

Across 8 pilot studies, automation reduced SSR setup time by over 60%. The reusable pipeline also improved the consistency of anomaly detection and provided traceable configuration files for every run. The pipeline supported dynamic rescheduling based on updated SDTM refreshes, minimizing the burden on programming and monitoring teams.

#### INTEGRATING AI INTO THE SSR DESIGN WORKFLOW

To enhance user experience and accelerate setup in the future, we eventually want to integrate a Machine learning model trained on historical trial data to pre-fill configuration suggestions. Based on characteristics like therapeutic area, region, and study type, the model predicts optimal parameter values including oversampling, under-sampling trends.

#### DISCUSSION AND FUTURE DIRECTIONS

By automating study-specific data configurations, we transform SSR from a reactive review process to a proactive surveillance model. The configuration files serve not only as templates but also as documentation artifacts, providing transparency and audit-readiness. Planned enhancements include integration with Databricks, visual feedback within the Data4You platform, and expansion into protocol amendment awareness. We also plan to integrate machine learning pipeline capabilities to summarize site-level risks from downstream outputs.

Study specific report (SSR), part of the Analytical Risk-Based Monitoring (ARBM) framework at Johnson & Johnson Innovative Medicine, enhances data quality oversight by identifying outlier signals across clinical trial sites. Historically, the generation of Study Specific Reports (SSRs) has required manual, non-scalable efforts, dependent on trial-specific customizations. To address this, we developed a modular, automated data pipeline architecture incorporating reusable study-specific configuration files. These configuration files streamline data ingestion from SDTM-like domains, automate the identification of relevant test codes, and enable dynamic scheduling of data refreshes for statistical surveillance. Leveraging cloud infrastructure and version-controlled codebases, the platform reduces cycle time, enhances traceability, and minimizes manual intervention, thereby improving operational efficiency and accelerating insights for centralized monitoring.

## CONCLUSION

The implementation of automated configuration files and metadata-driven data pipelines has led to measurable efficiencies in SSR reporting. By reducing manual effort, enabling reproducibility, and introducing AI-driven design recommendations, we have laid the foundation for scalable, intelligent data review workflows aligned with risk-based quality management strategies.

## REFERENCES


1. CSS in Clinical Trials. Lauwers et al., PHUSE 2018
2. Practical Implementation of SSR in Clinical Trials. Lo Regio et al., PHUSE US 2019
3. ICH E6(R2): Good Clinical Practice. https://www.ich.org/
4. Domino Data Lab Documentation. https://docs.dominodatalab.com/en/4.5.1/


## ACKNOWLEDGMENTS


Extending our sincere thanks to the excellent leadership team and all members of Advanced insights, Data science and Digital health department, BGIA, for the dynamic SSR application development activities support from Data4You team members and ongoing collaborations.

## CONTACT INFORMATION

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

Archana Nayak, PhD
DSDH, Johnson & Johnson R&D
Great Britain
Email: anayak32@its.jnj.com

Jill Ophoff
DMCM, Johnson & Johnson R&D
Belgium
Email: JOPHOFF@its.jnj.com