Paper SM10

Let's Talk Future: Automated Solution for Standard-Compliant Data, Mock Shells and TFL Generation Using Open-Source Platforms with AI

Abhijeet Arora, GenInvo, Toronto, Canada
Inderjeet Arora, GenInvo, Toronto, Canada

ABSTRACT

The traditional approach to standard-compliant data transformation, such as CDISC-SDTM and ADaM, as well as generating mock shells and Tables, Figures, and Listings (TFLs), heavily relies on manual, rule-based processes. This method is time-consuming, prone to errors, and demands specialized technical expertise. In contrast, automated solutions using open-source platforms like R and Python streamline the generation of standard data, mock shells, and TFLs. The automated process offers consistency, faster production, and reproducibility across studies, including an audit trail that enhances efficiency. Additionally, by incorporating artificial intelligence (AI) with a Human-in-the-Loop (HIL) approach, data transformation processes and TFL generation are further optimized, providing a more efficient, automated solution. This paper explores how automating standards-compliant data, mock shells, and TFL generation reduces manual effort, mitigates human error, and significantly enhances overall efficiency. Additionally, it highlights the advantages of automated mapping for studies with extensive datasets or projects involving multiple data refreshes.

INTRODUCTION

In clinical research, the transformation of raw data into standardized formats and the generation of Tables, Figures, and Listings (TFLs) are critical steps. Traditionally, these processes have been manual, labor-intensive, and susceptible to human error. The Clinical Data Interchange Standards Consortium (CDISC) has established standards like the Study Data Tabulation Model (SDTM) and the Analysis Data Model (ADaM) to ensure uniformity and facilitate data submission to regulatory authorities. However, adhering to these standards through manual processes can be cumbersome and inefficient.

The advent of open-source platforms such as R and Python has paved the way for automating these processes. Automation not only accelerates data transformation and TFL generation but also enhances consistency and reproducibility. Furthermore, integrating Artificial Intelligence (AI) with a Human-in-the-Loop (HIL) approach offers a promising avenue to optimize these processes, balancing machine efficiency with human oversight.

BACKGROUND

The traditional approach to clinical data standardization and reporting follows a structured flow from raw data to the final TFLs, but it remains a highly manual and time-intensive process. Raw clinical trial data, often sourced from various collection systems, must first be mapped and transformed into the CDISC SDTM format. However, since actual trial data can take months to become available, synthetic data is frequently used to build and test SDTM mappings in advance, reducing project delays. The transformation from raw data to SDTM requires extensive manual effort, including variable mapping, data derivations, and compliance checks to ensure adherence to regulatory standards. Once SDTM datasets are finalized, they serve as the foundation for ADaM datasets, where additional statistical derivations, baseline calculations, and complex dataset linking occur, all performed manually to maintain traceability. The transition from SDTM to ADaM is particularly labor-intensive, involving significant data reshaping and validation to ensure consistency with analysis requirements. Simultaneously, mock shells—templates defining the structure of TFL outputs—are manually created in Word or Excel, which not only increases the risk of inconsistencies but also requires frequent revisions due to protocol amendments. Finally, the production of TFLs from ADaM datasets involves writing extensive statistical programming code in tools like SAS or R, requiring repetitive manual coding, thorough validation, and meticulous quality control. Without automation, this entire workflow is prone to errors, inconsistencies, and prolonged development timelines, making the case for streamlined, AI-driven solutions ever more compelling.

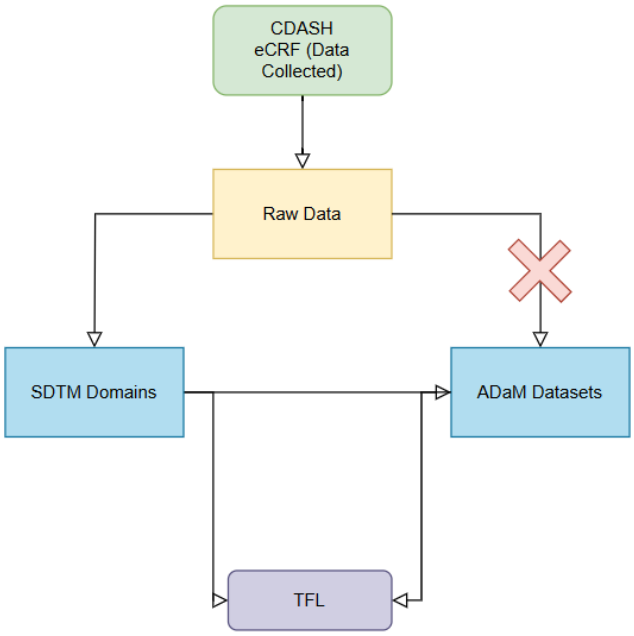


Figure 1: Raw Data to TFL Workflow

TRADITIONAL APPROACHES AND CHALLENGES

Historically, the process of transforming clinical trial data into CDISC-compliant formats and generating TFLs has been manual. This involves meticulous data mapping, programming, and validation to ensure compliance with SDTM and ADaM standards. The creation of mock shells for TFLs requires detailed planning and programming to accurately represent the data.

These manual processes are not only time-consuming but also prone to errors. The reliance on specialized technical expertise further adds to the complexity, making it challenging to maintain consistency across studies and increasing the risk of discrepancies.

CHALLENGES OF TRADITIONAL METHODS

The following are key challenges faced when using traditional methods for standard data transformation, mock shell generation, and TFL automation:

CHALLENGES IN STANDARDS-COMPLIANT DATA TRANSFORMATION

Manual SDTM mapping is a complex process that involves handling diverse data structures from multiple clinical trials, ensuring data quality despite inconsistencies, interpreting ambiguous CRF designs, and managing missing data. The process requires significant expertise to accurately map data elements to SDTM standards, making it prone to errors and inefficiencies. Additionally, complex transformations and a lack of standardized rules increase variability across studies, leading to delays and regulatory compliance challenges.

Large datasets require extensive effort for manual conversion into SDTM and ADaM formats. Each transformation step involves significant programming, validation, and review cycles, making the process inefficient and difficult to scale. The dependency on multiple teams also causes bottlenecks, delaying study timelines and increasing the risk of errors.

Manual processes struggle to keep up with frequent data refreshes and large-scale studies. With evolving study requirements, updates to datasets often require repetitive coding efforts, making it difficult to accommodate real-time data changes. The inability to efficiently scale manual transformations across studies results in wasted resources and delayed submissions.

Ensuring adherence to CDISC standards manually increases the likelihood of regulatory non-compliance. Manual verification of data standards is prone to oversight, leading to inconsistencies in dataset structure and metadata. Lack of automated checks may result in submission rejections, requiring extensive rework and additional time for

corrections.

CHALLENGES IN MOCK SHELL GENERATION

Manual creation of mock shells varies between teams, leading to inconsistencies across studies. Differences in formatting, structure, and reporting styles increase the risk of misinterpretation, leading to delays in report finalization. Without predefined templates, maintaining uniformity across multiple clinical trials remains a challenge.

Protocol amendments require extensive modifications to mock shells, consuming time and effort. Each change in study design necessitates updates to shell layouts, footnotes, and metadata, requiring multiple review cycles. Manual updates introduce delays, increasing the burden on study teams and prolonging finalization timelines.

Managing different versions of mock shells manually creates confusion and inefficiencies. With multiple stakeholders working on different document versions, tracking modifications and ensuring alignment across teams becomes difficult. The absence of an automated versioning system leads to data discrepancies and inconsistent report structures.

Traditional approaches lack interactive platforms for real-time updates and feedback. Team members working in silos struggle with information sharing, slowing down the mock shell approval process. Without a centralized system for collaboration, integrating feedback efficiently remains a significant challenge.

CHALLENGES IN TFL AUTOMATION

Manually coding tables, figures, and listings requires extensive statistical programming expertise. Creating customized output based on different study designs increases programming effort, leading to inconsistencies in output formatting. Additionally, manual TFL generation requires specialized knowledge of data handling, making it difficult for non-programmers to contribute effectively.

Lack of standardized metadata-driven approaches leads to inconsistent outputs. With manual methods, minor variations in code implementation across studies lead to discrepancies in TFL outputs. The inability to systematically reproduce output increases the risk of errors and inefficiencies in report generation.

Manual TFL generation relies on experienced statistical programmers, creating bottlenecks. Due to the complexity of statistical programming, a small pool of skilled professionals is often burdened with repetitive tasks, reducing overall productivity. The lack of automation also means that knowledge transfer between team members is inefficient, increasing risks during personnel transitions.

Errors in statistical table generation can lead to incorrect results and rework, delaying study timelines. Even minor formatting mistakes or incorrect calculations in manually created TFLs can require significant time for revalidation. Without automation, identifying and correcting errors in complex statistical reports remains a cumbersome process.

AUTOMATION USING OPEN-SOURCE PLATFORMS

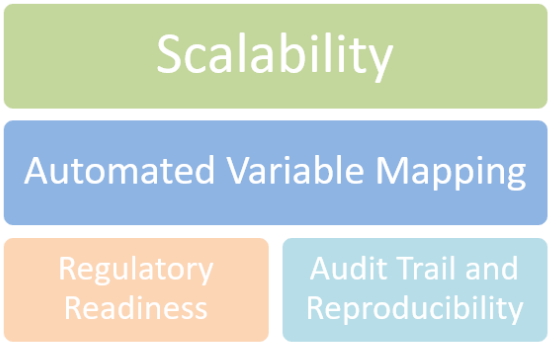
The emergence of open-source platforms like R and Python has revolutionized the automation of data transformation and TFL generation. Tools such as the SDTM Data Transformation Engine provide frameworks for modular programming of SDTM in R, facilitating the automation of SDTM dataset creation based on metadata-driven approaches.

Similarly, the TFL Designer is an innovative solution that automates the creation of TFL shells and provides machine-readable metadata, streamlining the programming of TFLs. It digitizes analysis results, ensuring alignment with CDISC Analysis Results Standards (ARS), and offers a central repository for TFL standards, templates, conventions, and metadata.

These tools enhance efficiency by reducing manual effort, ensuring consistency, and providing reproducibility across studies. They also offer flexibility, supporting various statistical software solutions and accommodating different study requirements.

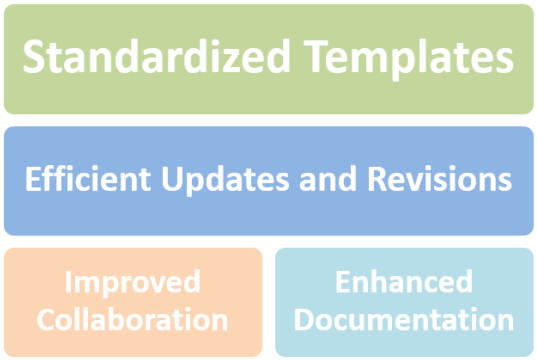
ADVANTAGES OF AUTOMATING STANDARDS-COMPLIANT DATA TRANSFORMATION

- Handles large datasets efficiently, accommodating frequent data refreshes.
- AI-driven NLP techniques enhance the accuracy of SDTM variable mapping.
- Built-in validation ensures compliance with SDTM and ADaM standards.
- Provides traceability and consistency across datasets.



ADVANTAGES OF AUTOMATING MOCK SHELL GENERATION

- Automation enforces consistency in mock shell structures.
- Changes to protocol or analysis plans are dynamically incorporated.
- Cloud-based mock shell tools enable multiple stakeholders to review and approve in real time.
- Automated annotations and metadata tracking improve audit trails.



ADVANTAGES OF AUTOMATING TFL GENERATION

- Automated processes significantly reduce the time required to generate TFLs.
- Metadata-driven automation ensures uniform TFL formatting across studies.
- Reduces human errors associated with manual table programming.
- Ensures adherence to CDISC and regulatory requirements with built-in validation mechanisms.

INCORPORATING AI WITH HUMAN-IN-THE-LOOP (HIL) APPROACH

Integrating AI into the automation process can further optimize data transformation and TFL generation. A Human-in-the-Loop (HIL) approach ensures that while AI handles routine tasks, human oversight is maintained to manage complex decision-making and ensure accuracy.

For instance, large language models (LLMs) have been explored for automating the generation of TFLs through prompt engineering and few-shot transfer learning. Using public clinical trial data in ADaM format, studies have demonstrated that LLMs can efficiently generate TFLs with prompt instructions, showcasing their potential in this domain.

This combination of AI and human oversight ensures that the automation process is both efficient and reliable, leveraging the strengths of both machine intelligence and human judgment.

ARCHITECTURAL FLOW OF AUTOMATION FOR SDTM STANDARDIZATION, MOCK SHELL GENERATION, AND TFL AUTOMATION

A well-structured automation architecture enhances efficiency and ensures compliance. The high-level workflow consists of:

SDTM STANDARDIZATION WORKFLOW

Raw Data Ingestion => AI/NLP-Based Variable Mapping => Metadata-Driven Transformation => Validation and Compliance Checks => SDTM Dataset Output

- Collect raw clinical data from different sources (EDC, external labs, etc.).
- Leverage AI to suggest SDTM mappings for raw datasets.
- Apply predefined transformation rules to convert raw data into SDTM format.
- Ensure CDISC compliance using automated validation tools.
- Generate submission-ready SDTM datasets.

MOCK SHELL GENERATION WORKFLOW

- **Metadata Extraction**: Identify study-specific reporting requirements.
- **Template Selection**: Choose predefined mock shell templates.
- **AI-Driven Customization**: AI generates annotations, table headers, and footnotes.
- **Review and Approval**: Interactive review mechanism for stakeholders.
- **Finalized Mock Shells**: Output mock shells in standardized formats.

TFL AUTOMATION WORKFLOW

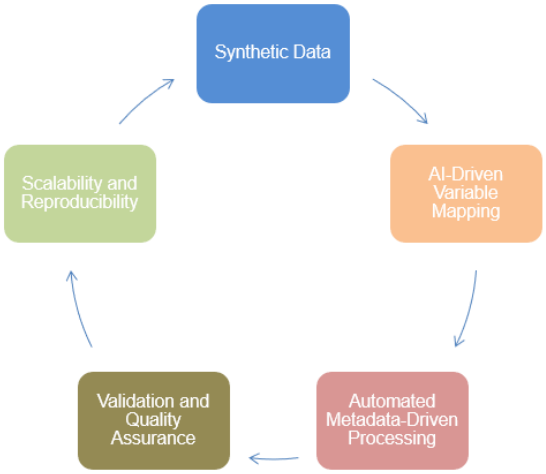
- **Define Analysis Requirements**: Capture statistical analysis plans.
- **Metadata-Driven Table Generation**: Use structured metadata to automate TFL creation.
- **AI-Assisted Layout Optimization**: AI optimizes table structures.
- **Integration with Statistical Tools**: Automate execution using R/SAS.
- **Finalized TFLs**: Generate submission-ready TFL outputs.

This architecture ensures seamless data standardization, efficient mock shell creation, and automated TFL generation, ultimately reducing manual effort and enhancing compliance.

SOLUTION

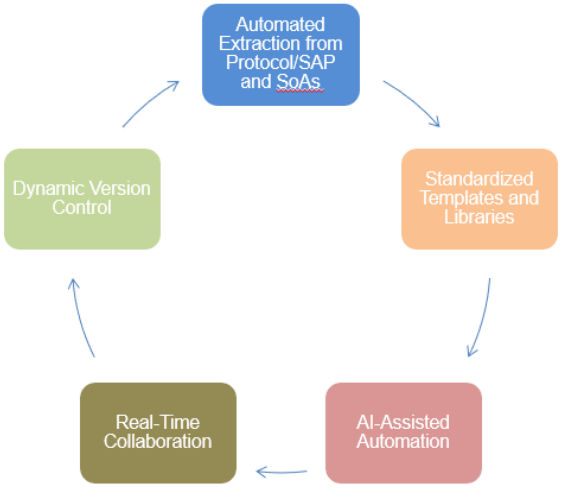
SDTM STANDARDIZATION SOLUTION

- Since the real clinical trial data takes considerable time to become available, synthetic data can be leveraged for early-stage SDTM preparation. This allows teams to develop, test, and validate data transformation processes well in advance, ensuring a seamless transition when actual data is available.
- Machine learning algorithms can be implemented to intelligently map raw variables to their corresponding SDTM domains, reducing manual effort and improving accuracy. This approach minimizes inconsistencies and ensures compliance with CDISC standards.
- By using metadata-driven frameworks, transformation processes can be streamlined and made more adaptable to study-specific requirements. This ensures that data conversions are consistent and reproducible across different clinical trials.
- Automated validation mechanisms can be integrated into the workflow to perform compliance checks against CDISC standards. These checks help to identify discrepancies early in the process, reducing the risk of regulatory non-compliance and the need for manual interventions.
- The automation framework should be designed to scale efficiently across multiple clinical studies with varying complexities. By ensuring that automation tools are modular and reusable, organizations can maximize efficiency and maintain uniformity in data transformation processes.



MOCK SHELL GENERATION SOLUTION

- Leverage AI and natural language processing (NLP) to scrape protocols and extract key study-specific metadata, including details on visits, treatment groups, and essential variables. This approach eliminates manual effort and ensures completeness in extracted information.
- Develop a well-structured repository of standardized templates that accommodate various study designs. These templates provide a strong foundation for generating consistent and regulatory-compliant mock shells across different trials.
- Deploy AI-powered models to intelligently generate table layouts, annotations, and footnotes. These AI-driven processes improve accuracy, streamline workflows, and reduce the likelihood of human errors in table formatting.
- Implement a cloud-based interactive platform that allows multiple stakeholders to review, provide feedback, and approve mock shells in real-time. This approach ensures seamless communication and faster iterations for study teams and regulatory reviewers.
- Maintain an automated version control system that tracks all modifications and updates in mock shell designs. This feature enhances traceability, ensures compliance with evolving study protocols, and minimizes inconsistencies across different versions of the mock shells.

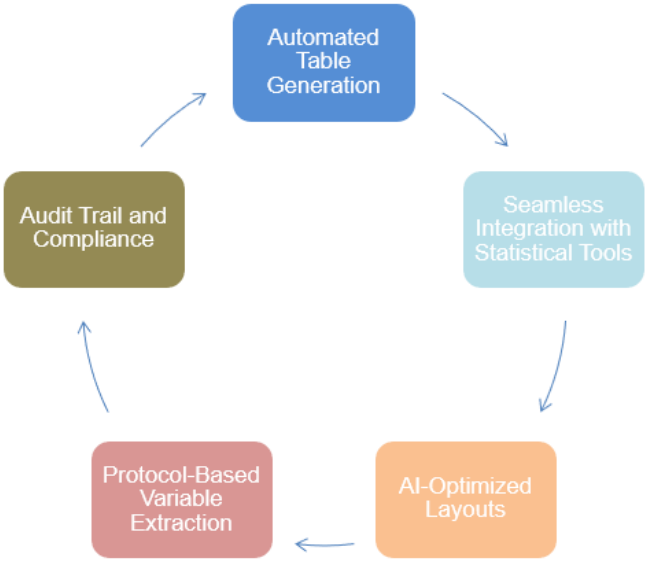


TFL AUTOMATION SOLUTION

- Utilize metadata-driven frameworks to define transformation rules that enable automatic creation of tables, figures, and listings. By leveraging predefined specifications, tables are generated in a consistent, structured manner, reducing the need for extensive manual intervention while ensuring standardization across studies.
- Direct integration with widely used statistical programming languages such as R and SAS enables smooth execution of automated processes. This connection ensures that statistical analyses align with regulatory standards while reducing redundancy in programming efforts.
- Implement AI-driven techniques to refine table structures, optimize spacing, and adjust formatting dynamically. AI models can analyze different output layouts and make real-time adjustments, enhancing

readability and overall presentation of the TFL outputs while adhering to predefined study requirements.

- Automate the identification and extraction of key variables from study protocols, statistical analysis plans, and related documentation. By using natural language processing (NLP) and AI-based parsing methods, critical study parameters such as treatment arms, endpoints, and timepoints can be extracted efficiently, reducing the manual effort required in defining TFL components.
- Ensure a transparent, traceable workflow by implementing an automated audit trail that logs all changes and modifications throughout the data transformation and TFL generation process. This enhances regulatory compliance, allowing for detailed tracking of changes, data lineage, and adherence to industry standards, ultimately supporting validation and submission requirements.



This solution approach ensures a seamless transition from raw data to SDTM, then to ADaM, and finally to TFLs, minimizing manual intervention, enhancing efficiency, and improving compliance with industry standards.

CONCLUSION

Automating the transformation of clinical trial data, mock shell generation, and TFL creation presents a paradigm shift in the way regulatory reporting is conducted. By leveraging open-source platforms, AI, and metadata-driven processes, organizations can streamline workflows, reduce manual effort, and enhance compliance with CDISC standards. Automation not only improves efficiency and scalability but also ensures data consistency and reproducibility across studies. Furthermore, integrating AI with a Human-in-the-Loop approach allows for intelligent automation while maintaining the necessary human oversight to handle complex cases. As clinical trials continue to evolve with increasing data complexity and regulatory requirements, embracing automation will be crucial in improving data integrity, accelerating timelines, and ultimately advancing medical research. Organizations that adopt these automated solutions will be better equipped to handle large-scale studies, regulatory submissions, and data refresh cycles with greater precision and efficiency.

REFERENCES

Clinical Data Interchange Standards Consortium (CDISC)

- CDISC. (2023). *Study Data Tabulation Model (SDTM) Implementation Guide.* Retrieved from https://www.cdisc.org/standards/foundational/sdtm
- CDISC. (2023). *Analysis Data Model (ADaM) Implementation Guide.* Retrieved from https://www.cdisc.org/standards/foundational/adam

U.S. Food and Drug Administration (FDA)

- FDA. (2021). *Study Data Standards Resources.* Retrieved from https://www.fda.gov/industry/fda-resources-data-standards

Pharmaceutical Users Software Exchange (PHUSE)

- PHUSE. (2023). *Open Source Technologies for Data Transformation and Analysis in Clinical Research.* Retrieved from https://phuse.global

PharmaSUG

- Various authors. (2024). *Automation in Clinical Data and TFL Generation Using Open-Source Platforms.* Retrieved from https://www.lexjansen.com/pharmasug/

TLF Automation

- GenInvo. (2023). Importance of “Table, Listing and Figures” Automation in Clinical Trials. Retrieved from https://geninvo.com/importance-of-table-listing-and-figures-automation-in-clinical-trials/
- GenInvo. (2023). CDISC Standards and Data Transformation in Clinical Trial. Retrieved from CDISC Standards and Data Transformation in Clinical Trial. - GenInvo

RECOMMENDED READING

Importance of “Table, Listing and Figures” Automation in Clinical Trials: Importance of TLFs Automation in Clinical Trials

Synthetic Data Vs Real Data: https://geninvo.com/synthetic-data-vs-real-data/

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author at:
Author Name: Abhijeet Arora
Company: GenInvo
Email: abhijeet.216@geninvo.com
Website: https://www.geninvo.com