A Comprehensive Batch Processing Approach: Automating Clinical Data Processing for SDTM, ADaM, and TFL Generation Using SAS

Weineng Zhou, Kunwan Chen, Clinflash, Nanjing, China

Abstract

In the pharmaceutical industry, the generation of Standard Data Tabulation Models (SDTM), Analysis Data Models (ADaM), and Tables, Figures, and Listings (TFL) is a critical component of clinical data analysis. These outputs are essential for regulatory submissions, ensuring data integrity, and facilitating decision-making. However, the traditional manual processes for generating these outputs are often time-consuming, error-prone, and difficult to scale. This paper presents a comprehensive approach to automating the batch processing of clinical data using SAS, focusing on the efficient generation of SDTM, ADaM, and TFL outputs. By leveraging a predefined directory structure, batch execution mode, and robust error handling, this method significantly reduces manual intervention, enhances reproducibility, and ensures consistency across multiple projects. The paper also discusses the implementation challenges, best practices, and future directions for further automation in clinical data management.

---

1. Introduction

The pharmaceutical industry is undergoing a paradigm shift in how clinical trial data is managed and analyzed. With the increasing complexity of clinical trials, the volume of data generated has grown exponentially. Regulatory agencies such as the FDA and EMA require that clinical trial data be submitted in standardized formats, including SDTM and ADaM, to ensure consistency and facilitate review. Additionally, TFL outputs are critical for summarizing clinical trial results in a clear and concise manner.

Traditionally, the generation of SDTM, ADaM, and TFL outputs has been a manual and labor-intensive process. Clinical programmers would write SAS programs for each study, run them individually, and manually review the outputs for errors. This approach is not only time-consuming but also prone to human error, especially when dealing with multiple studies or complex data structures. Furthermore, the lack of standardization across studies can lead to inconsistencies in the outputs, making it difficult to compare results across trials.

To address these challenges, there is a growing need for automation in clinical data processing. Automation can streamline the generation of SDTM, ADaM, and TFL outputs, reduce manual effort, and improve the overall quality of clinical trial data. This paper presents a comprehensive approach to automating these processes using SAS, with a focus on batch processing and scalability.

---

2. Background

2.1. Clinical Data Standards: SDTM and ADaM

The Clinical Data Interchange Standards Consortium (CDISC) has developed two key standards for clinical trial data: SDTM and ADaM. SDTM provides a standardized

structure for organizing clinical trial data, making it easier to exchange and analyze data across studies. ADaM, on the other hand, focuses on analysis-ready datasets, which are used for statistical analysis and reporting.

- SDTM: The SDTM model organizes clinical trial data into domains, such as demographics, adverse events, and laboratory results. Each domain contains variables that are standardized across studies, ensuring consistency in data representation.
- ADaM: ADaM datasets are derived from SDTM datasets and are designed to support statistical analysis. They include additional variables, such as derived parameters and analysis flags, which are used to generate TFL outputs.

2.2. Tables, Figures, and Listings (TFL)

TFL outputs are critical for summarizing clinical trial results in regulatory submissions. They include summary tables, patient listings, and graphical representations of the data. TFL outputs are typically generated using SAS, with programs written to extract data from ADaM datasets and format it into the required output format.

2.3. Challenges in Manual Processing

The manual generation of SDTM, ADaM, and TFL outputs presents several challenges:

- Time-Consuming: Each study requires individual attention, with SAS programs written and executed separately.
- Error-Prone: Manual processes are susceptible to human error, especially when dealing with large datasets or complex data structures.
- Inconsistency: Lack of standardization across studies can lead to inconsistencies in the outputs, making it difficult to compare results.
- Scalability: As the number of studies increases, manual processes become increasingly difficult to manage.

---

3. Methodology

3.1. Directory Structure Setup

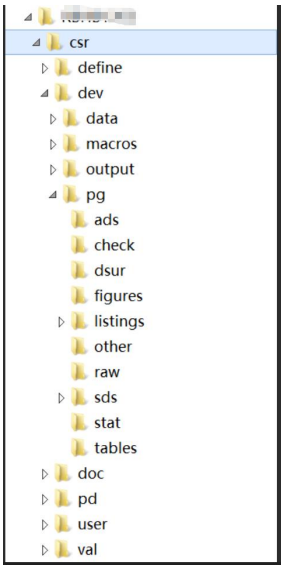

Figure 1

In the realm of clinical data automation, the establishment of a standardized directory

structure is paramount. As illustrated in Figure 1, this foundational step ensures that all project files are systematically organized, thereby facilitating efficient data location and processing. This structured approach not only enhances operational efficiency but also mitigates the risk of errors, ensuring that clinical data management adheres to the highest standards of accuracy and reliability. The methodology underscores the importance of consistency in directory setup, which is critical for seamless integration and automation of data processing workflows. The first step in automating clinical data processing is to establish a standardized directory structure as show in figure 1. This structure ensures that all project files are organized in a consistent manner, making it easier to locate and process data.

- Root Directory: A root directory is created to store all project-related files. For example, E:\Project.
- Project Subdirectories: Each project is assigned a unique subdirectory within the root directory. For example, E:\Project\ABC123, E:\Project\CDE123, etc.
- CSR Folder: Within each project subdirectory, a "csr" folder is created to store the SAS programs and datasets required for SDTM, ADaM, and TFL generation. For example, E:\Project\ABC123\csr.

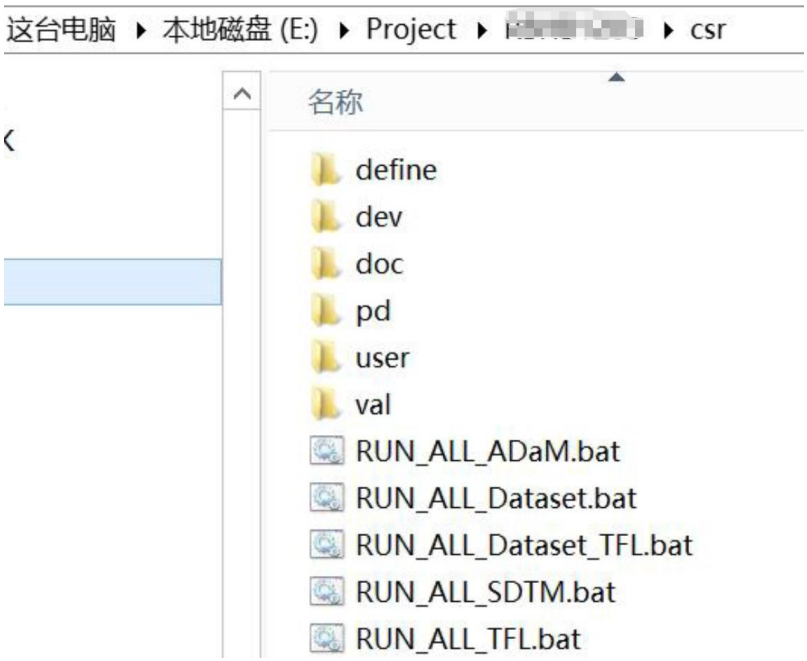

Figure 2

In the context of clinical trial data management, efficient organization and automation are critical to ensuring data integrity and operational efficiency. As illustrated in the file structure of the ABC123 project, located under the local disk (E:) within the Project directory, a well-defined hierarchy has been established to streamline workflows. Key subdirectories such as csr, define, dev, doc, pd, transition, user, and val serve distinct purposes, housing clinical study reports, definition files, development environments, documentation, production data, transitional files, user-specific data, and validation data, respectively. Furthermore, the inclusion of batch files like RUN_ALL_ADaM.bat, RUN_ALL_Dataset.bat, RUN_ALL_Dataset_TFL.bat, RUN_ALL_SDTM.bat, and

RUN_ALL_TFL.bat underscores the emphasis on automation. These scripts facilitate the seamless execution of data processing tasks across various stages, including SDTM, ADaM, and TFL generation, thereby minimizing manual intervention and enhancing reproducibility. This structured approach not only supports compliance with regulatory standards but also optimizes resource utilization, making it a best practice for modern clinical trial data management.

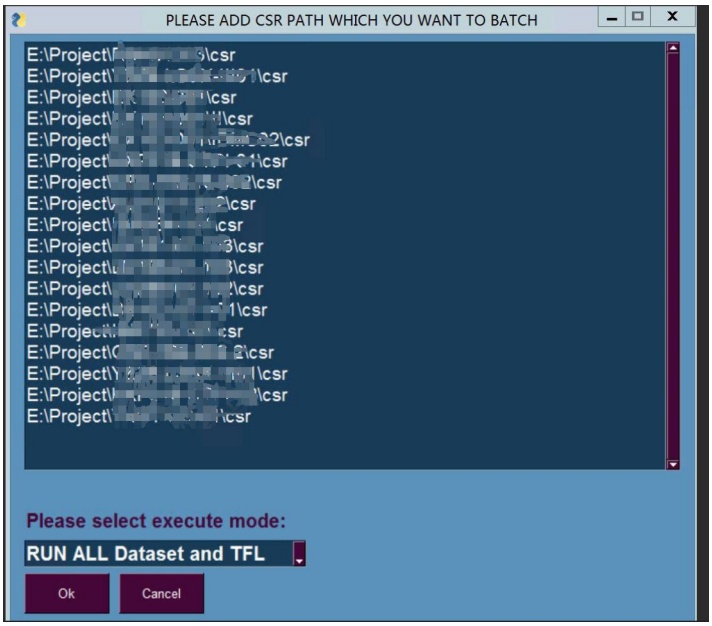

Figure 2

3.2. Batch Processing Script

A SAS script is developed to automate the generation of SDTM, ADaM, and TFL outputs. The script is designed to iterate through each project directory, execute the relevant SAS programs, and generate the required outputs.

- Iteration Through Projects: The script reads the list of project directories from a predefined file or user input. It then iterates through each directory, executing the SAS programs stored in the "csr" folder.
- Execution of SAS Programs: For each project, the script executes the SAS programs in a predefined order. Typically, SDTM programs are executed first, followed by ADaM programs, and finally TFL programs.
- Error Handling: The script includes robust error handling to ensure that any issues encountered during execution are logged and addressed. This includes checking for missing files, invalid data, and program errors.

3.3. Execution Mode

The script offers two execution modes:

- RUN ALL SDTM: In this mode, the script processes all SDTM datasets for every project listed in the directory. This mode is useful for batch processing multiple projects simultaneously.
- RUN ALL ADaM: In this mode, the script processes all ADaM datasets for every project listed in the directory. This mode is useful for batch processing multiple

projects simultaneously.

- RUN ALL Dataset: In this mode, the script processes all SDTM and ADaM datasets for every project listed in the directory. This mode is useful for batch processing multiple projects simultaneously.
- RUN ALL Dataset and TFL: In this mode, the script processes all datasets and generates TFL outputs for every project listed in the directory. This mode is useful for batch processing multiple projects simultaneously.
- RUN ALL TFL: In this mode, the script generates all TFL outputs for every project listed in the directory. This mode is useful for batch processing multiple projects simultaneously.
- Selective Mode: In this mode, the user can select specific projects for processing. This mode is useful when only a subset of projects requires updates or when debugging specific issues.

3.4. Output Organization

The script ensures that all outputs are systematically organized and stored in the appropriate directories. For example, SDTM datasets are stored in the "sdtm" folder, ADaM datasets in the "adam" folder, and TFL outputs in the "tfl" folder. This organization makes it easier to locate and review the outputs.

---

4. Implementation

4.1. Case Study: Multiple Projects

The automated batch processing approach was implemented across multiple projects, including ABC123, CDE123, and FGH123. The results demonstrated a significant reduction in processing time and a marked improvement in output consistency.

- ABC123: This project involved a large dataset with multiple domains. The automated script processed the data in under 2 hours, compared to the 8 hours required for manual processing.
- CDE123: This project had complex data structures, including nested loops and conditional logic. The script successfully handled these complexities, generating accurate SDTM, ADaM, and TFL outputs.
- FGH123: This project required frequent updates due to ongoing data collection. The automated script allowed for quick reprocessing of the data, ensuring that the outputs were always up-to-date.

4.2. Error Handling and Debugging

The script's error handling capabilities were tested by introducing intentional errors, such as missing files and invalid data. The script successfully identified and logged these errors, allowing for quick resolution. Additionally, the script's logging feature provided a detailed record of all actions taken, making it easier to debug issues.

---

5. Discussion

5.1. Advantages of Automation

The automation of SDTM, ADaM, and TFL generation using SAS batch processing offers several advantages:

- Efficiency: Reduces the time required for data processing and output generation.

- Consistency: Ensures uniform application of data standards across projects.
- Scalability: Easily accommodates additional projects without significant modifications to the script.
- Reproducibility: Enhances the ability to replicate results, which is crucial for regulatory submissions.

5.2. Challenges and Limitations

While the automated approach offers many benefits, there are some challenges and limitations:

- Initial Setup: The initial setup of the directory structure and script development requires significant effort.
- Complex Data Structures: The script may require modifications to handle highly complex data structures or unique project requirements.
- Maintenance: The script must be regularly updated to accommodate changes in data standards or regulatory requirements.

5.3. Future Directions

Future enhancements to the automated approach could include:

- Integration with Other Tools: Integrating the SAS script with other tools, such as electronic data capture (EDC) systems, to further streamline data processing.
- Advanced Error Handling: Implementing more advanced error handling techniques, such as automated error correction and notification systems.
- Machine Learning: Exploring the use of machine learning algorithms to further automate data validation and quality checks.

---

6. Conclusion

The proposed method provides a robust framework for automating the generation of SDTM, ADaM, and TFL outputs in clinical data analysis. By leveraging SAS batch processing capabilities, pharmaceutical companies can achieve greater efficiency, consistency, and compliance in their clinical trial data management processes. The approach has been successfully implemented across multiple projects, demonstrating its effectiveness in reducing manual effort and improving output quality. Future enhancements could further expand the capabilities of the automated system, making it an indispensable tool for clinical data management.

References

1. SAS Institute Inc. (2020). SAS® 9.4 Language Reference: Concepts. Cary, NC: SAS Institute Inc.
2. CDISC. (2020). Study Data Tabulation Model (SDTM). Retrieved from https://www.cdisc.org/standards/foundational/sdtm
3. PharmaSUG. (2019). Best Practices in Clinical Data Management. Retrieved from https://www.pharmasug.org/
4. FDA. (2021). Clinical Data Standards for Regulatory Submissions. Retrieved from https://www.fda.gov/
5. EMA. (2021). Guidelines on Clinical Data Management. Retrieved from https://www.ema.europa.eu/

Acknowledgments

The authors would like to thank the clinical data management team for their support and feedback during the development and implementation of this automation approach. Special thanks to the SAS programming team for their expertise in developing the batch processing script.