PharmaSUG China 2026 - Paper AA-145

Breaking Silos via Recycled Metadata: TOC-Driven Automation for Standard Title and Footnote Generation

Tracy Cui, Dizal Pharmaceutical Co., Ltd.

ABSTRACT

In clinical statistical programming and document development, fragmented file management and repetitive manual configuration of title and footnote (TLaFT) create significant information silos. These silos result in redundant metadata entry, inconsistent formatting and inefficient document iteration. This paper presents a practical R-Shiny automation solution that breaks down document silos by reusing TLaFT metadata.

The application offers three core functions:

1. Metadata Recycling — Automatically retrieves TLaFT metadata from previous deliverables, even across compounds or studies.
2. Intelligent Matching — Automatically aligns standard title-footnote templates with key information derived from pre-defined TOC metadata.
3. Regulatory-Ready Output — Automatically generates accurate, standardized, and regulatory-compliant titles and footnotes for final TFL production.

INTRODUCTION

Clinical TFL (Tables, Figures, and Listings) development depends on standardized TLaFT metadata to guarantee consistent formatting, complete data traceability, and regulatory compliance for clinical submission deliverables. In routine clinical statistical programming, traditional document management workflows suffer from severe metadata fragmentation and manual repetitive operations.

Historical TLaFT resources are scattered across independent study folders and archived deliverables, forming cross-project information silos. Programmers frequently rewrite and configure highly similar TLaFT content for new studies, leading to redundant metadata entry, inconsistent document formatting, and inefficient document iteration, while also increasing manual error risks and QC review workloads.

With a practical, interactive automation tool built with R Shiny, we can break TLaFT metadata silos and streamline routine TFL development. Centered on reusable clinical metadata, the application integrates three core functional modules to optimize the end-to-end TLaFT development workflow.

By replacing repetitive manual TLaFT compilation, manual template matching and scattered file sorting, the tool effectively eliminates redundant metadata operations, unifies TFL formatting specifications, and enhances data traceability and regulatory compliance. This lightweight, scalable, and low-threshold automation framework greatly improves the efficiency and quality of routine clinical statistical programming and TFL document iteration.

METHODS

As illustrated in **Figure 1**, the proposed tool's metadata reuse and assembly pipeline involves the centralization of scattered archived TLaFT assets into a reusable metadata library, with the TOC of a specified task providing official key identifiers. The tool matches and merges these two data sources to generate standardized target TLaFT records, combining reused title/footnote content from historical files with TOC-driven formal identifiers to produce fully structured, submission-ready TLaFT outputs with unified mandatory fields.

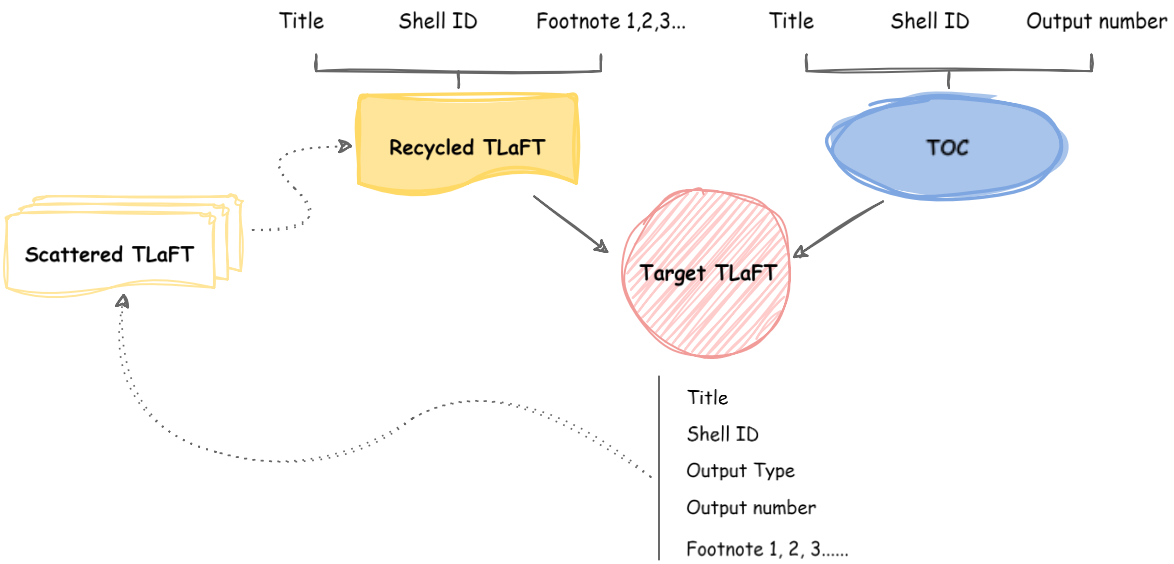


Figure 1. Automation from TOC to TLaFT

DEVELOPMENT ENVIRONMENT AND CORE PACKAGES

The tool is developed based on R 4.5.0 and the Shiny web framework, requiring no commercial software deployment and supporting lightweight, portable execution across enterprise Windows clinical workstations. All packages adopt fixed offline locked versions extracted from local offline installation archives to guarantee runtime stability, result reproducibility and clinical QC audit traceability:

- **shiny** (v1.13.0): Interactive web UI construction, drag panel resizing, custom JavaScript interaction and event-driven human-computer interaction;
- **stringdist** (v0.9.17): Optimized cosine distance calculation for normalized title semantic similarity scoring;
- **openxlsx** (v4.2.8.1): Regulatory-aligned Excel read/write, sheet styling, freeze panes, adaptive column width and bilingual CN/EN worksheet generation;
- **fs** (v2.1.0): Recursive directory traversal, temporary file filtering, secure path parsing for scattered cross-study TLaFT resource recycling;
- **rhandsontable** (v0.3.8): Editable interactive table for match review, locked protected core columns, right-click context menu batch operation;
- Auxiliary supporting packages: **dplyr** (v1.2.1) and **stringr** (v1.6.0) for vectorized data cleansing, text normalization and structured metadata integration;
- Additional dependent utilities: **DT** (0.34.0), **readxl** (1.5.0), **shinyjs** (2.1.1) for table rendering, raw Excel reading and frontend page interactive optimization.

CORE CUSTOM FUNCTION

Alongside foundational R packages, customized wrapper functions encapsulate domain-specific clinical logic for data cleaning, intelligent matching, metadata fusion, and compliant formatting. **Table 1** summarizes key self-developed functions, their functional objectives and real-world clinical programming application scenarios.

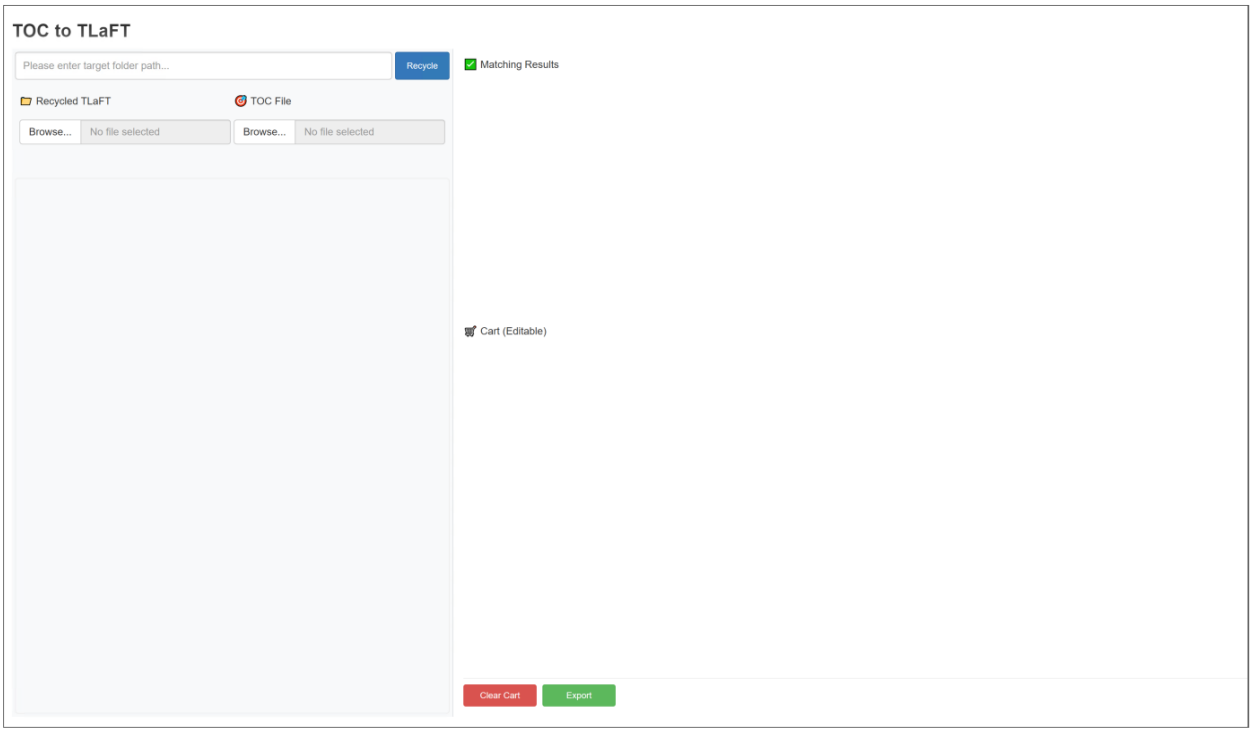
Function Name	Core Purpose	Example
`flatten_col()`	Flatten nested/cell-merged Excel content into clean single-line character strings	Resolve merged footnote cells stored as list objects into plain trimmed text for matching

Function Name	Core Purpose	Example
`cosine_match_optimized()`	Weighted matching: title cosine similarity + Shell_ID exact match, threshold filtering & top-N sorting	Match one TOC title against all historical TLaFT records, return top 8 candidates with score ≥0.7
`split_output_number()`	Split raw output identifier from TOC into standardized `OUTPUT_TYPE` and `OUTPUT_NUMBER` in TLaFT	Parse `T12.3` → `OUTPUT_TYPE = "T"`, `OUTPUT_NUMBER = "12.3"` for regulatory formatting
`process_cart_data()`	Hierarchical field mapping, fuse recycled TLaFT base data with authoritative TOC key values	Populate fixed `CORE_COLUMNS`, overwrite Shell_ID / Output_Number from current TOC while retaining historical footnotes
`apply_sheet_style()`	Regulatory-compliant Excel styling, freeze panes, zebra striping, auto column width	Format bilingual CN/EN TLaFT sheets with locked header rows and audit-ready visual formatting

Table 1. Core functions and purpose

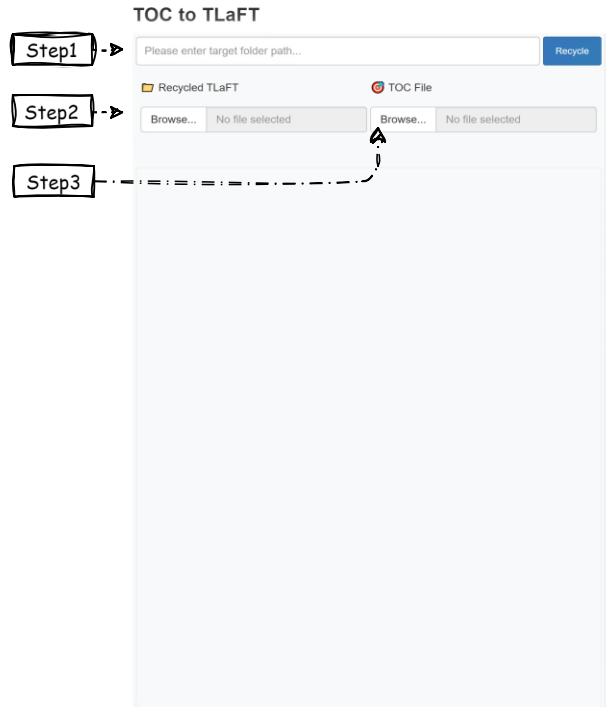
RESULTS

Following the proposed methodological framework, a dedicated Shiny web application is developed to realize automated TLaFT lifecycle management.



Display 1. Main Interface of TOC to TLaFT

KEY STEPS IN THE APP



Step 1: Cross-Project TLaFT Recycle

Users input the root directory path of archived clinical study deliverables and trigger the Recycle button. The tool recursively scans target folders; aggregates scattered historical Title & Footnote files and consolidates standardized reusable TLaFT metadata with full source traceability.

Step 2: Recycled TLaFT Upload and Sheet Select

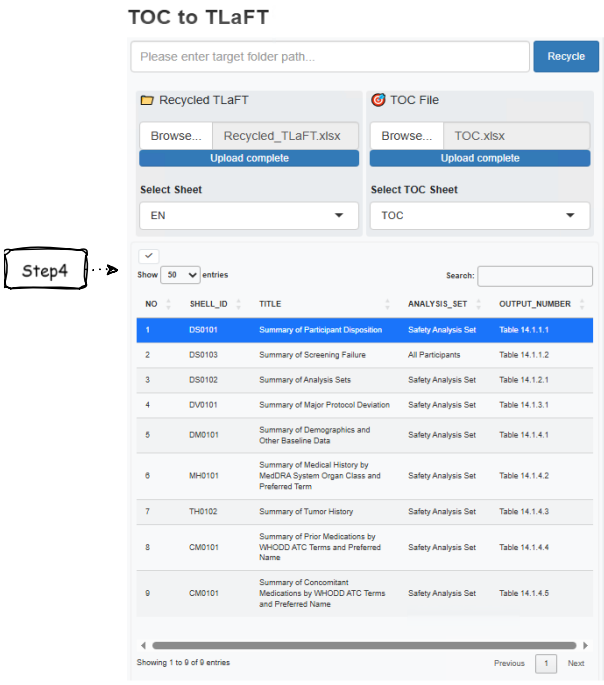
1 of the mandatory input files is archived reference TLaFT. The system parses sheet contents, standardizes column names, and cleans raw text to prepare datasets for intelligent matching.

Step 3: TOC Upload and Sheet Select

The other mandatory file is the current trial's TOC, which contains official identifiers such as Shell ID, target title, analysis set and output number.

Display 2. First 3 Steps on Left Panel

After the above 3 steps, embedded preprocessing functions clean, messy cell content, unify column naming conventions, and normalize text format to prepare standardized inputs for subsequent intelligent matching.



Step 4: Target Record Selection & Dual Matching Mode Switching

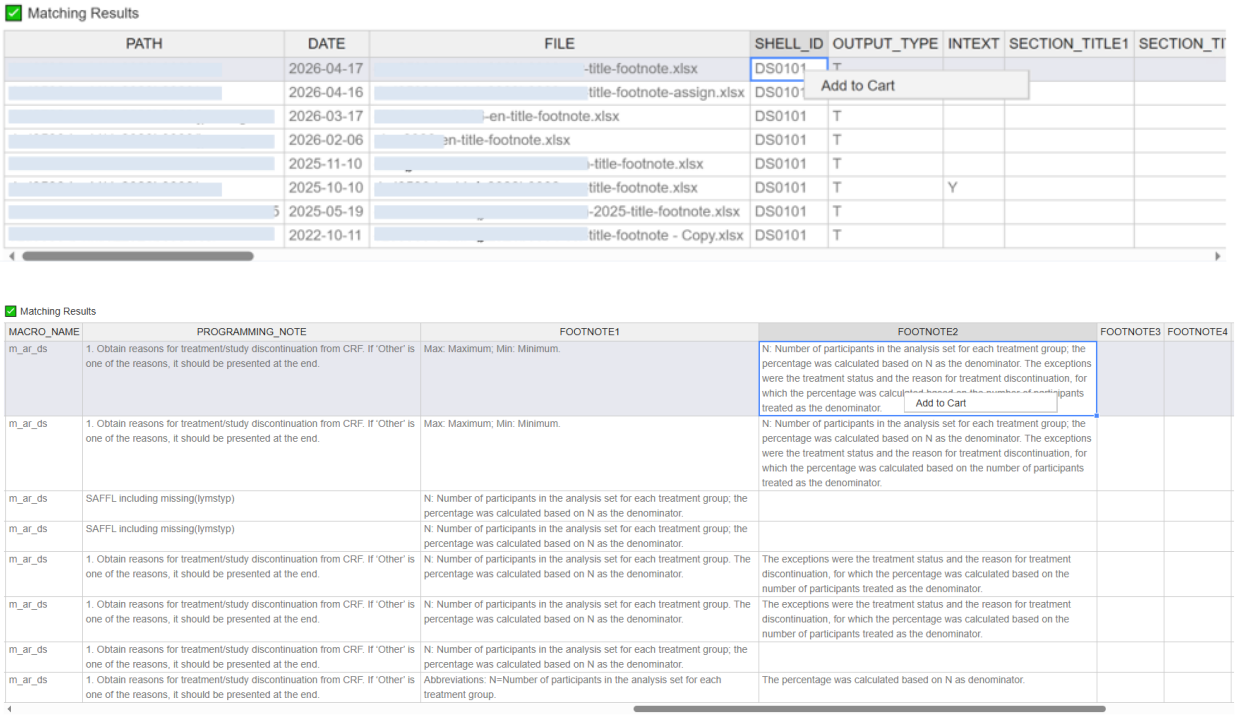
There are 2 matching modes that users can select based on project situations.

- Full Auto Batch Matching Mode: The system automatically pairs all TOC records with top-scoring historical TLaFT above the predefined 0.7 similarity threshold, delivering one-pass batch candidate generation to the cart in the right panel.
- Manual Single-Entry Selective Matching Mode: Programmers manually tick individual target rows inside the TOC table to designate only specific entries for similarity alignment. The top matching historical records were delivered to containers in the right panel.

Display 3. TOC Table on Left Panel

Step 5 : Add Qualified Records to Editable Cart

The screenshot in **Display 4** comprehensively presents all high-similarity matches filtered by the predefined similarity threshold, containing full audit-oriented fields: absolute file path (PATH), file last modification timestamp (DATE), source file name (FILE), core trial identifier SHELL_ID, output classification OUTPUT_TYPE, as well as key footnotes (FOOTNOTE1 - FOOTNOTE10). All the fields are set as read-only to avoid accidental alterations of the source metadata.



Display 4. Matching Result on Right Panel

Step 6: Export Final TLaFT

After click Add to Cart with the specified record, it will be synchronously appended to the Cart (Editable) panel on the lower right of the interface. The cart serves as the final manual quality control hub for all pre-confirmed TLaFT metadata before export.

Within the cart table, field-level permission control is implemented to balance edit flexibility and regulatory data integrity: mandatory trial identifiers such as SHELL_ID, OUTPUT_TYPE and standardized section titles sourced from the official TOC are set to read-only mode to preserve compliance with clinical trial documentation rules. In contrast, descriptive fields including supplementary footnotes, programmer annotations and explanatory text are fully editable, allowing users to revise wording, supplement contextual notes or adjust phrasing to fit project-specific submission requirements.

Multiple batch management functions are integrated for post-selection administration. The Clear Cart button clears all pending entries in one click for restarting candidate selection, while the Export button activates formatted Excel generation once manual validation finishes.

TOC to TLaFT

Cart (Editable)

TOC_No	SHELL_ID	OUTPUT_TYPE	INTEXT	SECTION_TITLE1	SECTION_TITLE2	OUTPUT_NUMBER	TITLE1
1	DS0101	T				14.1.1.1	Summary of Participant Disposition
2	DS0103	T				14.1.1.2	Summary of Screening Failure
3	DS0102	T				14.1.2.1	Summary of Analysis Sets
4	DV0101	T				14.1.3.1	Summary of Major Protocol Deviation
5	DM0101	T				14.1.4.1	Summary of Demographics and Other Baseline Data
6	MH0101	T				14.1.4.2	Summary of Medical History by MedDRA System Organ Cl
7	TH0102	T				14.1.4.3	Summary of Tumor History
8	CM0101	T				14.1.4.4	Summary of Prior Medications by WHODD ATC Terms and
9	CM0101	T				14.1.4.5	Summary of Concomitant Medications by WHODD ATC Ter

Clear Cart Export

Export Successed!

Display 5. Cart (Editable) on Right Panel

EFFICIENCY IMPROVEMENT

The sequential automated workflow integrated within the Shiny application delivers substantial efficiency gains compared with conventional manual TLaFT compilation workflows for clinical trial documentation, covering file retrieval, content matching, formatting standardization, and final deliverable generation.

1. **Batch Historical Metadata Recycling Efficiency**

   Manually locating, opening, and aggregating scattered legacy title-footnote files distributed across multi-study folder hierarchies typically consumes 25–40 minutes per project, with repetitive manual copy-and-paste prone to missed files and human omission. The tool's recursive folder scanning completes full metadata collection within 5–10 seconds. It automatically filters valid Excel files, extracts core structured fields including modification timestamp, file source path, and standardized title text, and assembles a unified reusable metadata library in one batch, eliminating manual file navigation and consolidation labor.

2. **Intelligent Matching Replaces Subjective Manual Alignment**

   Traditional practice requires programmers to manually screen historical records and compare textual similarity between current TOC entries and archived TLaFT content, a process that takes several minutes per TOC line and introduces inconsistent subjective matching judgment. The weighted cosine similarity algorithm executes bulk matching for the entire TOC dataset instantly, filtering qualified candidates above the preset confidence threshold (0.7) and sorting results by matching score. Automated semantic alignment cuts per-line matching time from minutes to milliseconds, unifies matching criteria across different programmers, and reduces inconsistency caused by individual experience differences.

3. **Human-In-The-Loop Review Optimizes Revision Workflow**

   Valid matching records can be added to the editable cart via one-click right-click operation, separating locked regulatory mandatory fields (Shell ID, output type, official TOC title) and editable descriptive footnote content. This design prevents accidental alteration of critical trial identifiers while centralizing all revision operations in a single panel. Instead of switching between dozens of source Excel files for footnote editing, clinical programmers conduct all fine-tuning within the app interface, lowering context-switching overhead and shortening manual QC review duration by approximately 60%.

CONCLUSION

This paper developed a dedicated R Shiny automation tool for clinical TLaFT metadata reuse, addressing the inefficiency and inconsistency of manual title-footnote compilation across clinical trials. Following structured multi-step workflows including recursive historical metadata recycling, dual-file uploading,

weighted cosine similarity intelligent matching, human-in-the-loop cart review and formatted one-click export, the system centralizes scattered legacy resources and standardizes matching judgment criteria.

This lightweight, dependency-friendly web application improves productivity for clinical programming teams, shortens trial documentation turnaround cycles, and provides a stable reusable solution for standardized metadata management in multi-study clinical development. Future optimization may expand batch multi-study export and configurable matching threshold customization to adapt to diverse submission rules.

ACKNOWLEDGMENTS

I would like to thank Dizal programming team for the discussion and testing on the features and functions of the tool.

CONTACT INFORMATION

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

Tracy Cui
Dizal Pharmaceutical Co., Ltd.
tracy.cui@dizalpharma.com