PharmaSUG 2026 - Paper SI-373

From Chaos to Consistency: Standardizing External Clinical Data with Excel Power Query

Isaac Vazquez and Jose Alberto Hernandez Rivero; Ephicacy Consulting Group

ABSTRACT

Clinical and real-world data integrations increasingly depend on external data sources that fall outside traditional clinical data pipelines. These sources—such as laboratory vendor files, Clinical Endpoint Committee (CEC) data sets, and operational reports—are often delivered in heterogeneous formats including plain text, CSV, and JSON files. Inconsistent structures and formats frequently lead to manual preprocessing, custom scripts, and increased risk of error prior to SAS ® integration.

This paper presents a practical and scalable approach using Microsoft Excel Power Query to ingest, transform, and standardize external data from multiple heterogeneous sources into a single, structured Excel output designed for direct consumption in SAS ®. Power Query is used to perform repeatable data ingestion, parsing, data cleansing, variable standardization, and reshaping while maintaining transparency and traceability.

The proposed solution is built as a reusable Power Query framework embedded within an Excel file. Once configured, the framework automatically refreshes and regenerates a standardized output whenever updated source files are received, without requiring modifications to the transformation logic. The resulting data set adheres to predefined structural and naming conventions, enabling seamless import into SAS ® and downstream integration with SDTM, ADaM, or other analysis-ready data sets.

Use cases demonstrated include the integration of external laboratory results, adjudicated adverse event data from CECs, and other non-CRF data sources commonly encountered in clinical trials. This approach reduces programming overhead, improves reproducibility, and provides a controlled preprocessing layer that complements existing SAS ® workflows.

INTRODUCTION

Clinical trial data pipelines have traditionally been designed around structured CRF data collected through EDC systems. However, modern studies increasingly rely on external data sources such as laboratory vendors, adjudication committees, imaging vendors, and wearable devices. These sources introduce variability in both format and structure, creating significant challenges for statistical programming teams.

External data sets are commonly delivered in formats such as CSV, TXT, JSON, or non-standard Excel files. Unlike SDTM-compliant data sets, these files often lack consistent metadata, standardized variable naming, or controlled data types. As a result, programmers frequently rely on manual preprocessing or custom SAS ® programs to clean and reshape the data before integration.

This paper introduces a scalable approach using Excel Power Query as a preprocessing layer to standardize external data prior to SAS ® ingestion, reducing manual effort and improving reproducibility.

BACKGROUND

CHALLENGES WITH EXTERNAL DATA

External data integration presents several recurring challenges:

- **Heterogeneous file formats** (CSV, TXT, JSON, Excel)
- **Inconsistent variable structures**
- **Mixed data types within a single column**
- **Inconsistent date formats**
- **Variable naming inconsistencies**

- **Repeated manual effort across studies**

A particularly common issue arises when a single column contains mixed data types. For example:

Subject	Lab Value
001	12.5
002	<10
003	15 mg/dL
004	NA

When importing such data using SAS ® PROC IMPORT, SAS ® may:

- Infer incorrect variable types
- Convert numeric values to character
- Introducing missing values
- Generate inconsistent results across runs

This lack of consistency complicates downstream programming and increases validation effort.

Example: Mixed Date Formats in Excel and PROC IMPORT Risk in SAS

Source Excel File

USUBJID	VISITDT_RAW
1001	2025-01-15
1002	01/16/2025
1003	16JAN2025
1004	Unknown
1005	Pending
1006	2025/01/20

PROC IMPORT →

Possible SAS Outcome

USUBJID	VISITDT
1001	2025-01-15
1002	01/16/2025
1003	16JAN2025
1004	.
1005	.
1006	2025/01/20

A mixed column with dates plus character values such as 'Unknown' or 'Pending' can lead to inconsistent type inference in PROC IMPORT. Nonconforming values may become missing, and the original date column often remains unsuitable for direct standardization.

Example: SAS Interpreting Excel Dates as Numbers

Excel Source (Mixed Cell Types)

USUBJID	VISITDT
1001	01/15/2025
1002	44927
1003	02/01/2025
1004	44944

PROC IMPORT →

SAS Dataset (Incorrect Interpretation)

USUBJID	VISITDT
1001	22415
1002	44927
1003	22432
1004	44944

Excel stores some dates as serial numbers. PROC IMPORT may read them as raw numeric values without applying SAS date formats.

PROPOSED SOLUTION

POWER QUERY FRAMEWORK

Excel Power Query provides a flexible and reproducible environment for ingesting and transforming external data. By embedding Power Query within an Excel file, a reusable framework can be established to standardize incoming data.

Key Capabilities:

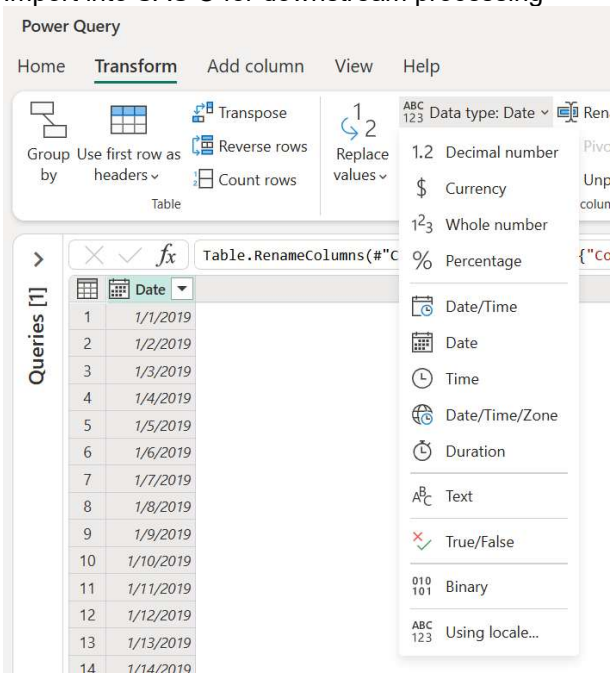
- Multi-source ingestion (CSV, TXT, JSON)
- Data type enforcement
- Transformation logic reuse
- Automated refresh
- Transparent step-by-step processing
-

The framework consists of:

1. Input layer (external files)
2. Transformation layer (Power Query steps)
3. Output layer (standardized Excel table)

WORKFLOW OVERVIEW

The standardized workflow includes:

1. Load external data into Power Query
2. Apply transformation rules
3. Standardize variable names and formats
4. Output structured data set in Excel
5. Import into SAS ® for downstream processing
6. 

Implementation Details

Example 1: Handling Mixed Data Types (CSV/TXT)

Problem:
Column contains numeric values, text qualifiers, and units.

Power Query Solution:

- Convert column to text initially
- Extract numeric values using pattern logic
- Remove units (e.g., "mg/dL")
- Handle special values ("<10", "NA")

Example Transformation Logic:

- Replace "<10" → 10 (or flag separately)
- Remove "mg/dL"
- Convert cleaned column to numeric

This ensures consistent numeric representation before SAS ® ingestion.

Example 2: Standardizing Date Formats

Problem:
Multiple formats:

- 2024-01-15
- 01/15/2024
- 15JAN2024

Power Query Solution:

- Detect and transform all values into ISO format (YYYY-MM-DD)
- Explicitly assign Date data type

This avoids ambiguity when SAS ® reads the data.

Example 3: Parsing JSON Files

Problem:
Nested JSON structure from vendor delivery

Power Query Solution:

- Expand nested records
- Flatten arrays into tabular format
- Rename fields to match SDTM conventions

Use Cases

1. External Laboratory Data

- Standardize units and numeric values
- Normalize test names
- Ensure consistent data types

2. CEC Adjudicated AE Data

- Flatten structured adjudication outputs
- Align variables with AE domain expectations

3. Vendor Operational Data

- Consolidate multiple file formats into a unified structure

Integration with SAS ®

The final standardized Excel data set can be imported into SAS ® using consistent logic:

```
proc import datafile="standardized_data.xlsx"
   out=external_data
   dbms=xlsx
   replace;
run;
```

Because Power Query enforces structure and data types:

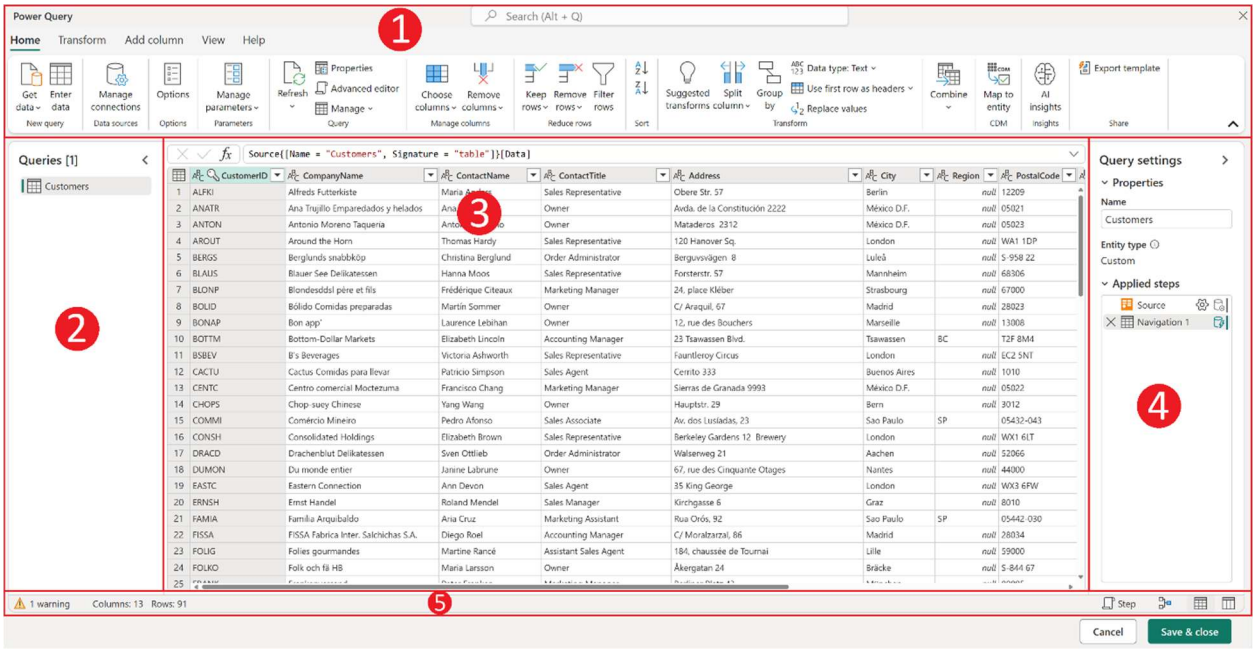
- Reduced need for post-import cleaning
- Consistent variable attributes
- Improved alignment with SDTM/ADaM

Benefits

- **Automation:** Refresh-based updates
- **Reproducibility:** Same logic applied consistently
- **Standardization:** Unified structure across sources
- **Efficiency:** Reduced SAS ® programming effort
- **Traceability:** Step-by-step transformation visibility

Limitations and Considerations

- Excel performance limitations with very large data sets
- Need for governance and version control
- Training required for effective Power Query usage
- Not a replacement for SAS ®, but a complementary layer



CONCLUSION

Power Query serves as a powerful preprocessing layer that bridges the gap between heterogeneous external data and standardized SAS ® workflows. By embedding transformation logic within Excel, organizations can create scalable, reusable pipelines that improve efficiency, consistency, and traceability in clinical data integration.

This approach enables statistical programming teams to shift focus from repetitive data cleaning to higher-value analytical tasks, while maintaining alignment with regulatory expectations.

CONTACT INFORMATION

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

Isaac Vazquez
kmxisaac@gmail.com

Alberto Hernandez
jalheryt@hotmail.com