Paper DS20

Transforming Legacy Data into CDISC Standards: Practical Approaches and Regulatory Insights

Veena Palle, Cytel, Boston, USA
Khalif Halani, Cytel, Vancouver, Canada
Patricia Feeney, Cytel, Austin, USA

ABSTRACT

Converting legacy clinical trial data into CDISC formats is essential for regulatory compliance and data reuse. Legacy datasets, often captured in older, study-specific formats such as excel, csv, txt, or SAS, lack the standardization required by regulatory agencies.

In a typical setting, SDTM and ADaM datasets serve as the base for the tables, listings, and figures (TLFs) included in the clinical study report (CSR). In some cases, the TLFs for the CSR are not written from CDISC-compliant data due to a variety of reasons. However, there may be a need for CDISC-compliant data after the TLFs are finalized.

In this presentation, we will explore practical strategies for converting legacy clinical trial data, where the TLFs for the CSR were not based on SDTM and/or ADaM, into CDISC-compliant formats. We will highlight common challenges and share key regulatory perspectives from agencies such as the FDA and PMDA.

INTRODUCTION

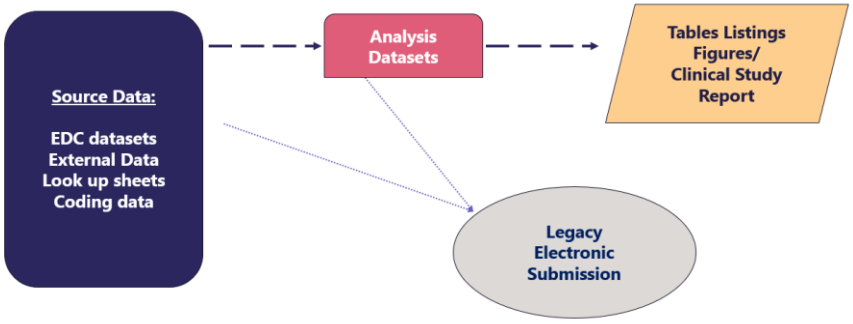
For contemporary studies, Clinical Data Interchange Standards Consortium (CDISC) compliance is addressed prospectively in the creation of the clinical study report (CSR); however, a significant volume of clinical research data either pre-dates these requirements or were developed without the CDISC structure. Many such “legacy studies” have finalized CSRs or other study reports that serve as the authoritative scientific record.

When CDISC transformation is initiated after the finalization of the CSR, driven by new regulatory submissions, integrated analyses, financial constraints, or acquisition, traditional forward mapping approaches are insufficient. Analyzing the data prospectively without regard to the CSR, is generally not acceptable, as it may lead to discrepancies with previously reported results. This paper addresses the unique challenges of post-CSR CDISC transformation and proposes strategies and guidance along with regulatory expectations and real-world experiences and constraints.

LEGACY VERSUS CDISC APPROACHES

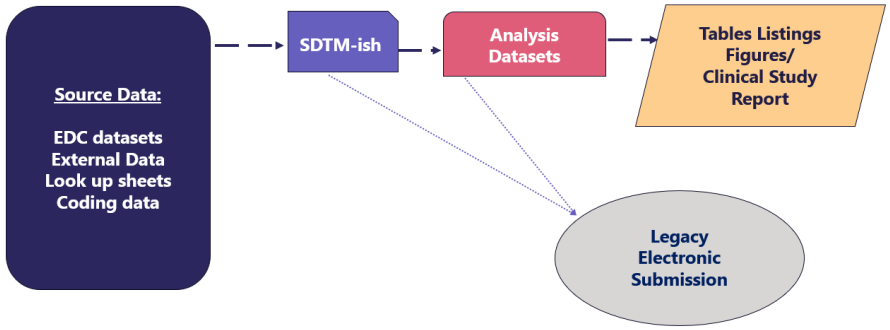
Prior to creation and implementation of the CDISC standards, each clinical trial’s tables, listings, and figures (TLFs) used for the CSR were collated and produced in a variety of ways, usually determined by the study’s Sponsor or analysis team. Typically, the source data were collated into analysis datasets which were then used to produce the CSR (Figure 1). The source and analysis datasets submitted to the regulatory agencies were not required to use common structures, naming conventions, domains, variables, controlled terminology, standardization that allows for common understanding or ease of review.

Figure 1. Producing TLFs and CSR prior to CDISC Standardization (Legacy)



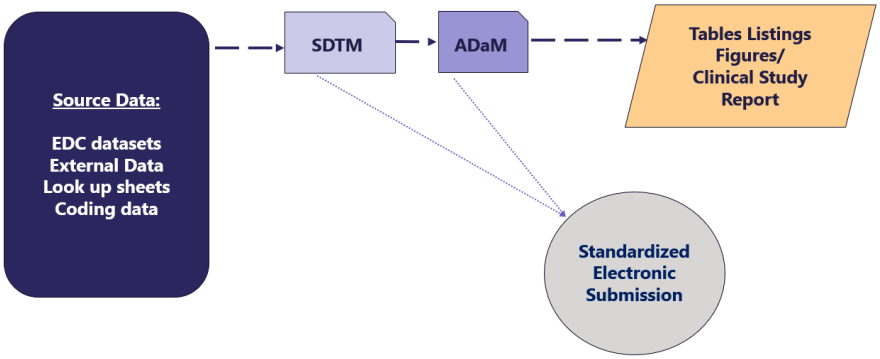
The CDISC standards were developed to create common approaches (rules of the road) for the collation and analyses of trial data to enable information system interoperability and improve medical research. Starting in the early 2000's the Consortium began the process of developing the core standards for the datasets and have regularly revisited and revised them based on user feedback and the various committee's recommendations. In 2004, the United States Food and Drug Administration (FDA) recommended that studies submit their data using SDTM. Many sponsors attempted to create SDTM-like datasets, but there were wide variations from the now common standard. These earlier versions of SDTM-like data look similar to the current day counterparts, but they are not necessarily conformant and require updates to be strictly submission-ready (Figure 2).

Figure 2. Variation of Legacy Data



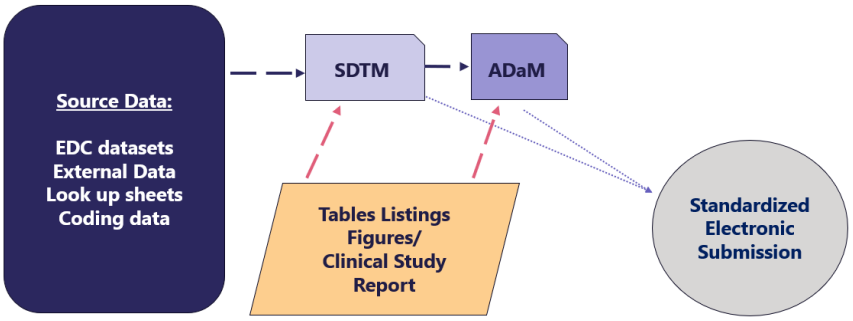
The FDA and Japan's Pharmaceuticals and Medical Devices Agency (PMDA) required studies that began after December 2016 (FDA) and submitted after April 2020 (PMDA) to submit their study's data according to the CDSIC standards, i.e. SDTM and ADaM (Figure 3). The CDISC standards have been updated since launching based on public feedback from users across programming, medical, statistics, data management; and the CDISC efforts continue to adjust to meet the needs of the medical research community.

Figure 3. Producing TLFs and CSR using CDISC Standards



The Conversion of Legacy data refers to a process by which original source data, which were used to produce the TLFs, are converted to the CDISC standard for submission independent of the original TLFs and CSR (Figure 4). The expected process (Source data → CDISC data → TLFs) is reversed; rather than the TLF results being produced by the datasets, the datasets are being produced, from the source data, but also to ensure they match the results in the TLFs and CSR. We refer to this retrofitting of the data as Conversion of Legacy data.

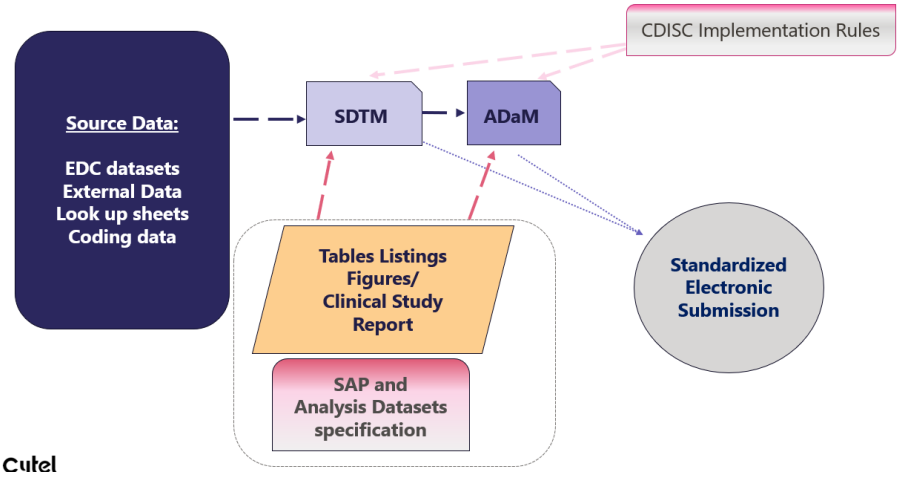
Figure 4. Conversion of Legacy Data



Producing the TLFs used for a CSR presents a variety of challenges, including the changes to the planned analyses (SAP and mock shells) during development. The CDISC conventions and standards are applied prospectively to the datasets (Figure 3) prior to the creation of the TLFs.

When converting legacy data to CDISC the task becomes more complex because the source data, the CDISC conventions, the specifications used to produce the legacy analysis data, and the TLFs and CSR become the inputs for the creation of the SDTMs and ADaMs. The retrofitted SDTM and ADaMs must be able to be "one-proc away" from producing all the results in the TLFs used in the CSR. The CDISC conventions and standards that would normally be applied prospectively are applied in tandem (Figure 5).

Figure 5. Conversion of Legacy Data to CDISC



FRAMEWORK AND REGULATORY PERSPECTIVES

One common reason to retrofit and create CDISC datasets is that the study predated the required CDISC conformance dates. However, this isn't the only reason. Others include:

- Study results and data were not originally intended for submission (e.g., investigator-initiated studies)
- Study was terminated by the Sponsor before the study team had developed the CDISC-compliant datasets
- Financial or practical reasons led to a decision to create a report without the CDISC convention
- Originally submission accepted via FDA waiver of CDISC requirement
- Other reasons
- Any combination of the above

It is common that early phase studies, identified during planning for a Clinical Summary of Safety or Efficacy for a specific compound, were performed using legacy approaches (Figure 1 and Figure 2). In order to submit the data for the Integrated Summary of Safety or Effectiveness the underlying data must be converted to CDISC standards.

Regulatory authorities (FDA, PMDA) will accept the post-CSR conversion of the data to CDISC for submission even though the original CSR is not written directly from the CDISC sources. The FDA provides guidance on challenges and methods for converting legacy data in its Study Data Technical Conformance Guide (SDTCG), Specification

Document (Dec 2025). While regulators expect the datasets to be technically CDISC-compliant and fully traceable to reported results, they do not expect legacy studies to conform perfectly to current implementation guides. They also do not recommend a specific approach. Rather, they expect:

- consistency between CDISC datasets and CSR outputs (traceability),
- transparent documentation of derivations (adequate to support review),
- reproducibility of the results, and
- clear justification for deviations from CDISC guidance

The Sponsor should prepare a Legacy Data Conversion Plan (Plan) describing the legacy data conversion (i.e., which standardized datasets will be created [SDTM and/or ADaM] and the sources). At the end of the conversion the Sponsor must produce the Legacy Data Conversion Report (Report) with the results of the conversion, issues encountered and resolved, and any outstanding issues (SDTCG - Section 8.1.4.2).

To begin the conversion process, the first step is to collect all of the information used to create the TLFs and CSR. These include:

- CSR or similar study report
- TLFs (if available)
- All source datasets
    - Corresponding Data Transfer Agreements or manuals for external data (if available)
- Case Report Forms (CRFs) from the electronic data capture
- Statistical analysis plan (SAP)
- TLF mock shells and any programming instructions (if available)
- Legacy analysis datasets (if available)
    - Specifications (if available)
- SAS ® programs used to create the analysis datasets or TLFs (if available)

Often not all these pieces of information are available – in the worst-case scenario the CSR and source data are the only data available. Occasionally, Sponsors may believe SDTM datasets are available when they are, in fact, SDTM-like. In this case, upon assessment, the Plan would be revisited and the work would commence.

Regardless of the availability of information, the next step is to deconstruct the TLFs and CSR to understand and identify the fundamental components needed to create SDTM and ADaM datasets. Ideally many of the supporting documents are available, but the final CSR content is the definitive authority as the final product must match the CSR. This information will need to be included in the Report. The essential components to map out from the CSR results include:

- Endpoint definitions
- Analysis population criteria
- Censoring, imputation, derivations, or other important data handling rules
- Statistical methods used to generate the results

The third step is to create the SDTM and ADaM datasets and the accompanying electronic submission package (define-XML, reviewer's guides, and annotated CRF), an annotated CRF that maps to the legacy data (for traceability), verify the CDISC datasets can produce the TLFs and CSR, and ensure that the resulting datasets meet submission requirements. This becomes an iterative process until the datasets are robust and complete. The SDTM reviewer's guide should include copies of the Plan and the Report.

CHALLENGES AND RECOMMENDATIONS

The following are a few challenges Cytel encountered when performing a post-CSR conversion of study data to CDISC.

- **Challenge:** The team performing the CDISC conversion did not work directly on the CSR which leaves an expected knowledge gap.
- **Recommendation:** Collect as much information as possible and available when starting the project. In the absence of the availability of programs or finalized specifications, an obscure note-to-file or even an email may contain a nugget of information that could lead to downstream data handling of derivations or, in some cases, hard-coding.

- **Challenge:** Some older databases require coding (such as MedDRA and WHO Drug) re-applied to the raw data.
- **Recommendation:** The availability of coding dictionaries should be contemporaneous with the version used for the CSR.

- **Challenge:** Applying terms using Controlled Terminology (CT), a central component of CDISC, which may not have been reported in the CSR.
- **Recommendation:** The CT Term should be agreed upon for submission. Ensure both the CT term and the CSR-reported are included in the datasets. If values don't match, adjudicate a decision and report in the Legacy Conversion Data Report and, if applicable, the Reviewer's Guide.

- **Challenge:** If programing of legacy data analyses involved hard coding (i.e., manually changing data points) and if the supporting documentation is either absent or insufficient for to trace, it can be time-consuming to match final results.
- **Recommendation:** If the SAS ® programs specifications are available, search through the code and documentation to identify hard coding. If these are not available and the summary statistics are extremely close to matching the CSR, but slightly off, consider that hard coding may have occurred. Review the data in the listings from the CSR against the SDTM/ADaM data to allow for the identification of hard coding needed from source-to-SDTM level in order to match the CSR. Report any handling in the Legacy Conversion Data Report and Reviewer's Guide.

- **Challenge:** Chronological data entry errors, such as start dates occurring after end dates, can exist in the legacy data. Unlike in a prospective setting, these errors cannot be queried and resolved. The issues will be captured in the SDTM and propagated in the ADaM Pinnacle 21 reports.
- **Recommendation:** Ensure the data match the CSR and highlight the discrepancies in the Legacy Conversion Data Report and Reviewer's Guide, both in the conformance findings and in the individual domain sections.

- **Challenge:** Legacy data values collected do not conform to the non-extensible CDISC code list (e.g., adverse event action taken with study drug, relationship to study drug), leading to CT validation errors and associated Pinnacle 21 findings.
- **Recommendation:** Ensure the original value is mapped to the SDTM so that the ADaM version of the variable is reflective of the original reported value. Ensure this is outlined in the Legacy Conversion Data Report and the SDTM Reviewer's Guide both in the conformance findings and in the individual domain sections.

- **Challenge:** Look-up lists developed by the Sponsor and implemented at the ADaM level are commonly used to classify and categorize data and those data may not be available in the source documents.
- **Recommendation:** Reverse engineering the look-up list from the results may be needed. Clear documentation and processes are required. Refer to these in the specification component of the define.xml and Reviewer's Guide.

TIPS AND TAKEAWAYS

We recommend the following:

- Standardize the SDTMs as much as possible to other submission-compliant datasets in the compound. This will reduce any effort when pooling for integration.

- Ensure the Reviewer's Guides and annotated CRF are clear regarding which data points were omitted from the CSR, or the converted datasets and the reasons.

- Ensure that the SDTM reviewer's guide includes the Legacy Data Conversion Plan and Report and that two sets of annotated CRFs (one for SDTM and one for legacy) are included in your submission.

- Establish a review plan for the datasets against the CSR-results prior to beginning the conversion process; review and revise as the conversion progresses.

- Ensure the ADaM datasets are "one proc away" from the results. For example, present an ADaM dataset that is ready to feed into a complex statistical procedure in SAS (GLMMIX, PHREG, MIXED) making use of the standard ADaM variables (ANLxxFL, CRITxxFL, PARCATx) and "where" statements without additional data manipulations. Since the reviewers will not have access to the output programs, take the extra step to ensure their review is eased by reducing the amount of data manipulation that needs to be done to verify the results.

- Consider using the PARCAT variables to define for the reviewer the purpose of the parameter (e.g., PARCATy = "Co-Primary Estimand", PARCATy = "Sensitivity #3 for Secondary Estimand"). Since the

reviewers will not have access to the output programs and you have done the work to recreate the analysis in the CSR, increase the transparency for them and reduce their review time.

- Use conformance software (such as Pinnacle) at intervals during the conversion to avoid re-work at the end.

- One advantage is that the CSR is complete; there is no chance someone will make updates to the statistical analysis plan or request additional output be added to the CSR. All surprises will come from within the data rather than outside the process.

CONCLUSION

Post-CSR conversion of legacy data to CDISC-compliant dataset is a necessary and, potentially time-consuming, step. Clear traceability and reproducibility of results while bridging multiple sources and guidelines is essential for submission purposes. We recommend gathering as much information from the original study as possible in your endeavors (leave no stone unturned). Start by having a good plan to confirm the reproducibility of the results but also the flexibility to pressure test the plan as you dig in and find out more. Ensure the proper submission documents are available (including the Legacy Data Conversion Plan and Report and the CRF annotated for legacy values) and that the Reviewer's Guides and ADaM design are easy to follow for the reviewers.

REFERENCES

CDISC Published Standard Library: https://www.cdisc.org/standards
US FDA, Study Data Technical Conformance Guide, Technical Specifications Document, December 2025.
US FDA, Electronic Common Technical Document v4.0, Technical Conformance Guide, March 2025.

ACKNOWLEDGMENTS

Many thanks to Nalini Doddapaneni and Agnes Nkenglefac for their feedback on recent conversion studies and contributions. A special thanks to Angelo Tinazzi for his helpful insights, expertise, and references.

RECOMMENDED READING

We recommend this paper by our colleague which provides some great suggestions:
https://www.lexjansen.com/phuse/2018/ds/DS01.pdf

CONTACT INFORMATION

We would love to hear from you. Contact the authors at:
Khalif Halani and Veena Palle
Cytel, Inc.
Emails: Khalif.halani@cytel.com, Veena.palle@cytel.com
Website: https://cytel.com/