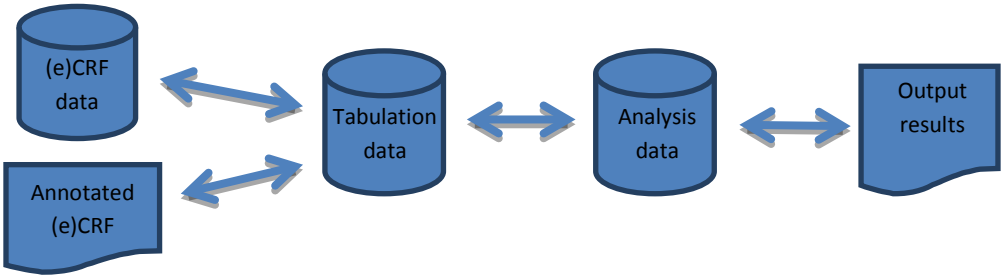
Study Level Traceability Considerations
Best Practices for Non-Linear Data Flow

October 2014

Background:

In 2013, The Traceability and Data Flow sub-team of the CSS working group *Optimizing the Use of Data Standards* produced a white paper titled 'Best Practices for Basic Linear Data Flow'. The purpose of that white paper was to provide fundamental best practices for data traceability to support Regulatory Agency review of electronic clinical data when the flow of clinical data processing is linear – see diagram below.



In this linear data flow, there are no post processing or parallel steps that may complicate or break the chain of data traceability. If CDISC standards are not utilized as part of this linear data flow, post processing steps are required to convert legacy data to meet the CDISC data standards formats. As a result of these post processing steps, data traceability is sometimes put at risk.

Scope:

This white paper will focus on traceability in a non-linear data flow, resulting from the conversion of legacy data to CDISC standards. In February, 2014, the FDA released the *Draft Study Data Technical Conformance Guide – Technical Specifications Document*[1]. In the *Draft Study Data Technical Conformance Guide*, the FDA outlined the following recommendations for sponsors to consider when converting legacy data:

1. Prepare and Submit a Legacy Data Conversion Plan and Report
2. Provide an annotated CRF that maps the legacy data elements
3. Record significant data issues, clarifications, explanations of traceability, and adjudications in the *Data Guide*. For example, data were not collected or were collected using different/incompatible terminologies, or were collected but will not fit into, for example, SDTM format.
4. Sponsors should consider the submission of legacy data (i.e., legacy CRF/aCRF, legacy tabulation data, and legacy analysis data) in addition to the converted.

[1] Study Data Technical Conformance Guide:
http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

The *Draft Study Data Technical Conformance Guide* also provides a list of common traceability issues resulting from three different types of legacy conversion approaches:

1. Legacy data conversion to SDTM only and submitted with legacy analysis datasets
2. Independent conversion to SDTM and ADaM, from legacy tabulation datasets and analysis datasets, respectively
3. Sequential conversion of legacy data to SDTM followed by creation of ADaM from the SDTM datasets (TLF (Tables, Listings and Figures) and CSR (Clinical Study Report) still produced from legacy data)

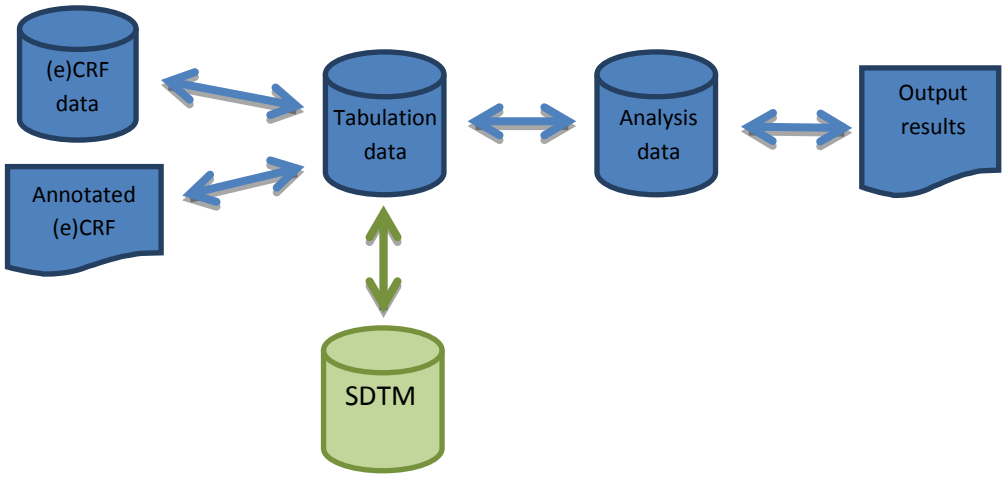
The remainder of this white paper will take a closer look at the issues identified in the *Draft Study Data Technical Conformance Guide* that result from these three conversion approaches. Primarily, recommendations will be made to help minimize the risk of these issues occurring. It should be noted that upon finalization of the *Study Data Technical Conformance Guide*, this white paper will need to be reviewed to ensure it is consistent with the final version of the guide.

The recommendations outlined in this white paper do not replace the need for sponsor communication with the FDA on data standardization plans. Such communication should occur as early as possible in the clinical trial process as individual reviewers may have different needs.

1. Legacy data conversion to SDTM only

- Assumes complete legacy traceability documentation is in place
- Assumes conversion to "true" SDTM (not SDTM like or +/-)

The following diagram shows where the legacy data conversion takes place relative to the linear legacy data flow. The ensuing table lists the issues identified in the *Draft Study Data Technical Conformance Guide* and provides possible solutions for minimizing the issue or alleviating it completely.



[1] Study Data Technical Conformance Guide:
http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

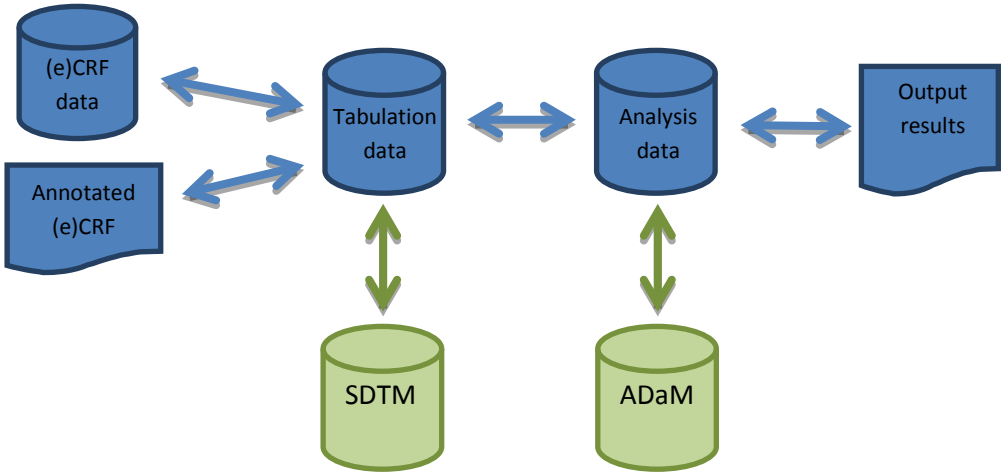
FDA Identified Issue	Risk Mitigation
No traceable path from legacy analysis data to SDTM.	• Submit legacy tabulation (raw) data, with relevant legacy traceability information, in addition to the SDTM • Submit traceability information between legacy tabulation (raw) data and SDTM, and/or • Submit two annotated CRFs, one for legacy tabulation (raw) data and one for the SDTM data
No ability to confirm analysis variable imputation or derived variables.	• Submit legacy tabulation (raw) data, with relevant legacy traceability information, in addition to the SDTM
Unable to replicate tables, listings and figures (TLFs) and legacy analysis datasets using SDTM datasets.	• Submit legacy tabulation (raw) data, with relevant legacy traceability information, in addition to the SDTM
No ability to confirm derivation of intermediate datasets or custom domains.	• Submit legacy tabulation (raw) data, along with intermediate datasets, with relevant legacy traceability information, in addition to the SDTM • Assuming the reference to custom domains means SDTM-like datasets then refer to core assumptions above.
No ability to determine location of collected CRF variables in the converted SDTM data.	• Submit annotated CRFs for the SDTM data
Difficulty in understanding the source or derivation methods for imputed or derived variables in integrated/pooled data, supplemental qualifiers, and related records.	• Submit legacy tabulation (raw) data, with relevant legacy traceability information, in addition to the SDTM • Submit traceability information between legacy tabulation (raw) data and SDTM, and/or • Submit two annotated CRFs, one for legacy tabulation (raw) data and one for the SDTM data for each source study

In addition to these recommendations, appropriate QC steps should be taken to ensure that the SDTM datasets and the legacy tabulation datasets produce the same results. For example, reproduce key summary results using the SDTM datasets and compare to TLF produced from the legacy data. The more rigorous the quality checking, the less risk that the FDA identified issues will occur.

If differences between the SDTM datasets and the legacy datasets cannot be resolved, these should be clearly documented in the reviewer's guide. This may occur when legacy/sponsor defined terminology cannot be re-mapped to CDISC controlled terminology.

[1] Study Data Technical Conformance Guide:
http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

2. Independent conversion to SDTM and ADaM, from legacy tabulation datasets and analysis datasets, respectively



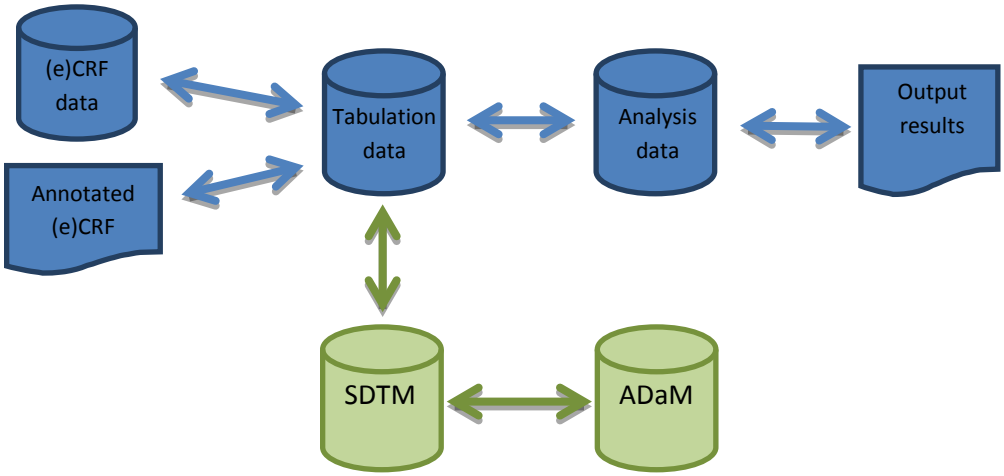
The above diagram depicts the independent SDTM and ADaM conversion scenario. The *Draft Study Data Technical Conformance Guide* identified the following issues that can occur when this approach is implemented:

- No traceable path from legacy to SDTM to ADaM and to Study Report.
- No explanation or source for analysis imputed or derived variables.
- No traceable path to ISS and ISE / pooled data.
- TLFs do not match datasets (analysis datasets or SDTM datasets (when used)).
- No traceable path to intermediate datasets or custom domains.
- No explanation or source for imputed or derived variables in datasets.

Per the ADaM Implementation Guide, it is an assumption that ADaM datasets are derived from SDTM. Taking this into consideration along with the fact that this conversion approach introduces two opportunities for traceability issues, it is strongly recommended that this approach not be taken.

[1] Study Data Technical Conformance Guide:
http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

3. Sequential conversion of legacy data to SDTM followed by creation of ADaM from the SDTM datasets (TLF and CSR still produced from legacy data)



The primary concern identified when this approach is implemented is that a reviewer may not be able to replicate the results in the TLF and CSR using the ADaM or the SDTM datasets (FDA TCG Risk)

During the review process, a Statistical Reviewer at the FDA will typically work backwards from the CSR, then to the TLF, and then to the analysis datasets to confirm and/or understand results. If the post-processed SDTM and ADaM datasets do not produce the same results as the legacy datasets and TLF a significant amount of time can be wasted reconciling the differences.

Although this conversion approach may be the most reliable, the following recommendations should be taken into consideration.

1. Use the post-processed ADaM datasets to reproduce key TLF, compare to the legacy produced TLF, and reconcile any differences prior to submission.
2. If it is not feasible to recreate TLF, use the ADaM datasets to reproduce the numbers in the output results as part of your internal QC process
    a. Full QC of all numbers is ideal
    b. If this is not possible, recommend QC of key results
    c. Document scope of QC in your metadata and reviewers guide or other supporting documentation delivered as part of your submission.
    d. Submit SDTM annotated CRF
3. Be prepared to submit legacy data components – tabulation (raw) data and analysis datasets – that were used to produce outputs, TLF and CSR.
    a. Submit traceability information between legacy tabulation (raw) data and SDTM
    b. Submit additional annotated CRFs for legacy tabulation (raw) data.

[1] Study Data Technical Conformance Guide:
http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf

   c. Submit SAS programs used to create legacy TLF
4. In the reviewer's guide, clearly document the fact that the ADaM datasets were not used to create the TLF and CSR and provide summary of validation steps taken to ensure accuracy and consistency of results.

[1] Study Data Technical Conformance Guide:
http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf