PharmaSUG 2017 - Paper SS11

Documenting Traceability for the FDA: Clarifying the Legacy Data Conversion Plan & Introducing the Study Data Traceability Guide

David C. Izard, Chiltern International Ltd.
Kristin C. Kelly, Merck & Co. Inc.
Jane A. Lozano, Eli Lilly & Company

ABSTRACT

Traceability from data collection through to the presentation of analysis results has always been a concern of the US Food & Drug Administration (FDA). The introduction of electronic data as part of submission added additional steps to confirm provenance of information. Now the requirement to provide clinical and non-clinical data based on a set of FDA endorsed data standards adds exponentially to the challenge, especially if legacy format data structures were utilized when the study was originally executed and reported but data meeting FDA requirements must be present in your submission. The PhUSE organization, made up of volunteers across industry, has worked closely with the FDA to develop tools to support the organization, presentation and interpretation of clinical and non-clinical data to regulatory bodies. Examples include the Study & Analysis Data Reviewer's Guides and the Study Data Standardization Plan. These documents describe routine situations where FDA endorsed data standards are deployed at the time a study is initiated; additional support is needed when the provenance of the data is not as straightforward. The FDA's Study Data Technical Conformance Guide calls out for the need to provide a Legacy Data Conversion Plan & Report when legacy data is the source of deliverables based on FDA endorsed data standards, but it is not very clear as to when you must provide one. This paper will leverage recent PhUSE efforts to develop a template and completion guidelines for this document to clarify when it must be provided and introduce the concept of the Study Data Traceability Guide.

INTRODUCTION

This paper will walk the reader through the FDA's expectations and desires with respect to establishing the provenance of data and related assets from collection through to analysis, submission & regulatory review. It will start with steps the FDA has driven as part of their regulatory review, then proceed through the recently articulated and widely used assets such as the Study Data Standardization Plan and Reviewer's Guides, take an in-depth look at the role of the little used Legacy Data Conversion Plan & Report, and wrap up with a discussion of the concept of one "document" that would serve as an anchor for all aspects related to traceability at the study and, potentially, submission level.

HISTORY OF TRACEABILITY IN THE AGENCY'S EYES

ROUTINE FDA AUDIT DURING SUBMISSION REVIEW CYCLE

During the review of a New Drug Application (NDA) or Biologics Licensing Application (BLA) the FDA conducts a routine audit to ensure traceability of the analysis results back to the delineation and collection of the source data. Typically this involves FDA representatives coming onsite to a Sponsor organization in order to review study assets. For example, an FDA representative might start with an adverse event summary table present in the clinical study report (CSR) of a pivotal study and then try to work their way back through the assets, reviewing the actual table, the data and the associated program used to create the table, any programs and versions of data that preceded the data used to produce the table, and possibly samples of actual CRF pages of subjects that contributed records included in the summary table.

This is a clear case of where the FDA has not requested documentation of traceability up front but rather pursues establishing traceability of their own accord.

STUDY DATA STANDARDIZATION PLAN

The Study Data Standardization Plan (SDSP) is described in the FDA binding guidance *Providing Regulatory Submissions In Electronic Format — Standardized Study Data* [1] and the *Study Data*

Technical Conformance Guide (Study Data TCG) [2]. The SDSP is the beginning of the traceability story for a compound. The document is a high level summary of the exchange and terminology standards for nonclinical and clinical studies submitted as part of the IND or NDA filed under an indication. This is an opportunity for FDA to look at the standards for planned, ongoing, and completed studies early in the development of the compound development cycle.

The document is not a submission document – yet. It is possible the drug could fail and not reach submission. It is also possible a trial may fail and would not potentially be included in a submission. However, it is important to document the trial as FDA wants to see what the plan for the compound is at the time the SDSP is created. There is discussion that should take place within the sponsor organization to determine what trials should be included based on consultation with the compound team and regulatory group. FDA does not want to be surprised as it may be too late in the development of the compound to make needed changes. It is important to be in alignment with FDA and the standards that are supported in the FDA's Data Standards Catalog [3]. As sponsors move into providing standardized data (as opposed to legacy analysis data) and there is a version of SDTM that is no longer supported per the Data Standards Catalog, the information provided in the SDSP is an early look at the standards that are being/will be applied to studies within a compound/indication. The standards provided in the SDSP align with the information that is in the catalog. Information about pooling of data is also included in the SDSP, when it is known. The SDSP is a living document and should be updated appropriately prior to the submission. It can be submitted as a stand-alone document or should be included in the General Investigational Plan when it is part of an IND application.

The SDSP is also useful when an NDA has been submitted and the compound has a new line indication. The SDSP is the beginning of the traceability story for the new indication. Information about the new indication is included in the SDSP with a reference to the original NDA.

ROLE OF DATA DEFINITION FILES (DEFINE.XML)

In contrast to the SDSP, which provides a high-level overview of the studies in an application, a data definition file (define.xml, a.k.a. define file) is submitted at the study level. The define file is another important component to establishing traceability. Each study submitted should include a define.xml (or define.pdf for legacy studies) in the package that is a tabular collection of metadata that describes the data in the corresponding datasets (SAS v5 XPORT files, .xpt). The define file can be compared to a table of contents one might find in a book that lists the datasets, variables, and origin of the data. The requirement to include a define.xml is outlined in the Technical Conformance Guide and the supported versions of the Define standard are listed in the Data Standards Catalog.

In the submission package of tabulation data for a clinical study, an annotated CRF (aCRF) should also be included that links the collection field to the corresponding variable in the dataset. For each variable listed in the define.xml, the origin of where the data came from in the CRF should be linked back to the corresponding CRF page. Since not all data for a study is collected on the CRF, the origin of that data whether it be external data transfer (eDT), assigned, or derived will also be noted in the define.xml in the 'Origin' column. It should be emphasized that for a nonclinical study or an analysis package for a clinical study an aCRF will not be included, but the origin of each data point is provided just as in the define.xml for clinical tabulation data. For nonclinical, there is no CRF used to collect data because data is managed by specific collection instruments. In regard to analysis data for a clinical study, the data is derived from the tabulation datasets and a CRF is not necessary to be provided.

The define file will also contain value-level metadata which describes the actual values that are used to populate each variable in a submitted dataset for that particular study as well as any computational algorithms used to derive values. Any corresponding codelists or dictionaries used in the study, e.g. MedDRA, WHODrug, CDISC controlled terminology, will also be listed in the define.xml.

STUDY & ANALYSIS DATA REVIEWER'S GUIDES

In addition to providing the datasets, define.xml and aCRF, if applicable, the TCG recommends to include a Reviewer Guide (RG) to convey any further information about the study data that cannot be described in the aCRF or the define file that would aid the FDA reviewer in understanding the data submitted.

Study Data Reviewer's Guide (SDRG)

For tabulation data (SEND or SDTM), this document is referred to generically as the Study Data Reviewer's Guide (SDRG). For nonclinical studies, it is recommended to be named 'nsdrg.pdf' ('n' designates 'nonclinical') and provided as a PDF file for each study in module 4 (m4) of the eCTD. For clinical, it should be named 'csdrg.pdf' ('c' designates clinical) and provided as a PDF for each study in module 5 (m5).

The cSDRG typically contains the SDTM version used as well as dictionary/terminology (e.g. MedDRA, CDISC) versions for that particular study. It should also contain information about the trial design and any domain-specific information that cannot be described in the define.xml. There is a section for results of automated validation items that check for conformance to SDTM that cannot be resolved due to data oddities. Each issue that remains should be listed in this section with a comprehensive explanation for why the specific validation check cannot be rectified. The cSDRG should also contain content to explain instances where data collected on a CRF may not be present in the SDTM datasets as well as any issues encountered during conduct of the study or creation of the submission deliverables. Transparency about the study data enhances traceability throughout the submission.

The nSDRG typically contains the SEND version used as well as the version of CDISC Controlled Terminology. It may also contain information about the study design as well as domain-specific information that may be important to the reviewer. As in the cSDRG, there is a section that lists the results of automated validation checks for SEND and explanations for those that remain. The nSDRG also describes instances where there are discrepancies between the SEND datasets and the Study Report and explanations for those discrepancies.

Analysis Data Reviewer's Guide (ADRG)

For clinical analysis data (ADaM), this document is referred to as the Analysis Data Reviewer's Guide (ADRG). The ADRG provides context for analysis datasets and terminology similar to the SDRG. The ADRG also provides a summary of ADaM conformance findings. The ADRG purposefully duplicates limited information found in other submission documentation (e.g., the protocol, statistical analysis plan (SAP), clinical study report, define.xml) in order to provide FDA reviewers with a single point of orientation to the analysis datasets.

LEGACY DATA CONVERSION PLAN AND REPORT

The Legacy Data Conversion Plan & Report (LDCP) is described in section 8.3.2 of the Study Data Technical Conformance Guide which is associated with the FDA binding guidance Providing Regulatory Submissions In Electronic Format — Standardized Study Data. Traceability is important, especially to a regulatory reviewer, when collected study data that is non-standardized data is converted to standardized data. This will aid the reviewer to follow the data from collection to analysis when the analysis was originally done with legacy analysis data. Non-standardized data and legacy data are used interchangeably to refer to data that does not conform to standards currently supported in the Data Standards Catalog. Standardized data, on the other hand, refers to data that conforms to standards currently supported in the Data Standards Catalog. While sponsors work through the period of providing standardized study data based on the rule when the requirement begins, legacy data conversions will occur for studies in which legacy analysis data was used.

The Legacy Data Conversion Plan was a byproduct of an FDA project initiated in the early part of this decade. The FDA engaged with an organization to produce select SDTM domain datasets & ADaM analysis datasets from data that had been provided to the agency in support of regulatory submissions. They focused on a number of studies from a number of vendors within a single therapeutic area/indication to support a proof of concept that SDTM & ADaM could be successfully used to analyze data across studies and Sponsors. This activity primarily involved creating SDTM & ADaM from legacy tabulation data guided by legacy analysis assets that were present. As the agency sought to confirm results and assess usability they ran into issues where existing tools to record data migration decisions did not adequately explain all they needed to know in order to perform their work. From this a process and

template was developed to produce what is now the predecessor to the Legacy Data Conversion Plan to document detailed information about conversion decisions.

The Legacy Data Conversion Plan & Report should be added to the appropriate reviewer's guide for clinical studies (cSDRG, ADRG). The Technical Conformance Guide describes three situations in which the LDCP should be included in the reviewer's guide.

LEGACY TABULATION DATA CONVERTED TO SDTM

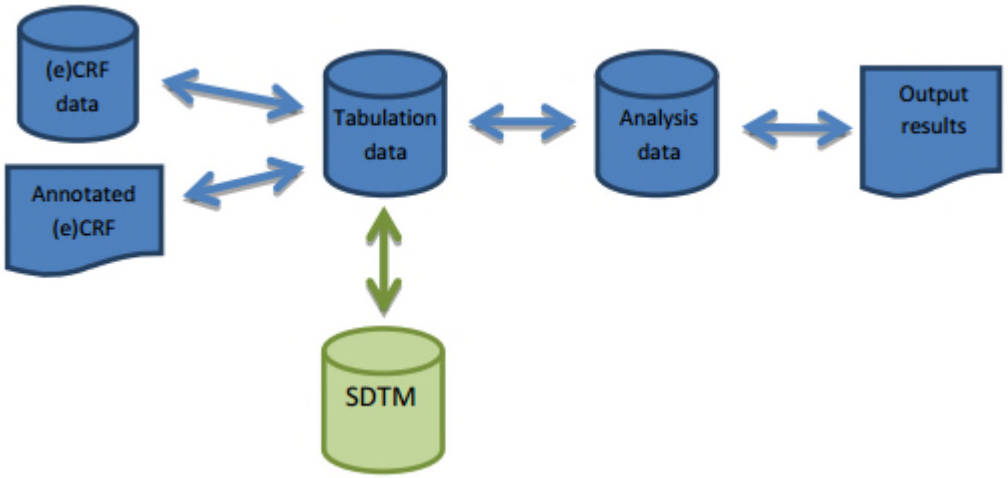

Figure 1 - Legacy Data Conversion to SDTM

Per the TCG, issues observed when data flows in this manner include the following:

- Limited ability to determine location of collected CRF variables in the converted SDTM data unless the legacy aCRF is re-annotated.
- Limited traceable path from SDTM to the legacy analysis data.
- Limited ability to replicate/confirm legacy analysis datasets (i.e., analysis variable imputation or derived variables) using SDTM datasets.
- Limited ability to confirm derivation of intermediate analysis datasets or custom domains.
- Difficulty in understanding the source or derivation methods for imputed or derived variables in integrated/pooled data, supplemental qualifiers, and related records.

LEGACY TABULATION DATA CONVERTED TO SDTM AND THEN ADAM IS CREATED FROM THE SDTM

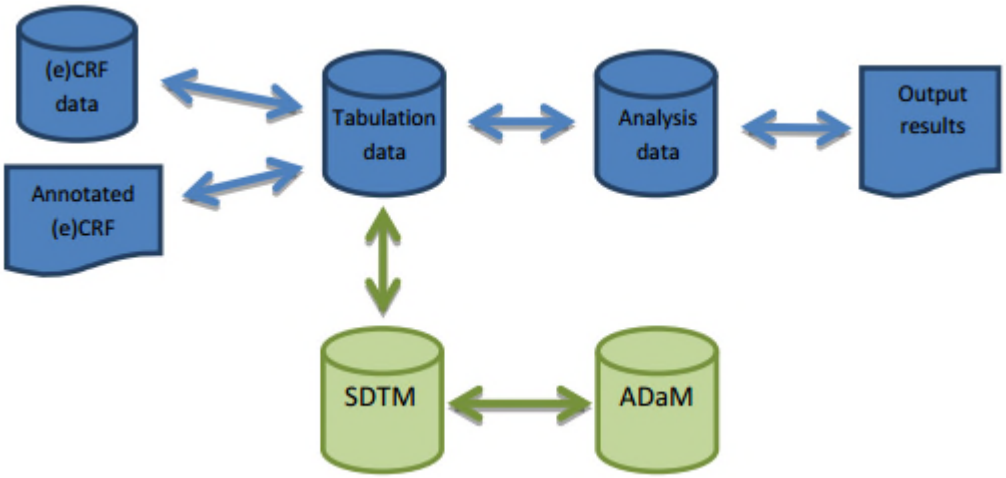

Figure 2 - Legacy Tabulation Data Converted to SDTM, ADaM Created from SDTM

Per the TCG, issues observed when data flows in this manner include the following:

- Limited ability to determine location of collected CRF variables in the converted SDTM data unless the legacy aCRF is re-annotated.
- Limited traceable path from SDTM to the legacy analysis data.
- Limited ability to replicate/confirm legacy analysis datasets (i.e., analysis variable imputation or derived variables) using SDTM datasets.
- Limited ability to confirm derivation of intermediate analysis datasets or custom domains.
- Limited traceable path from ADaM to the Tables, Figures and the CSR.
- Difficulty in understanding the source or derivation methods for imputed or derived variables in integrated/pooled data, supplemental qualifiers, and related records.

LEGACY TABULATION DATA CONVERTED TO SDTM AND LEGACY ANALYSIS DATA CONVERTED TO ADAM IN PARALLEL

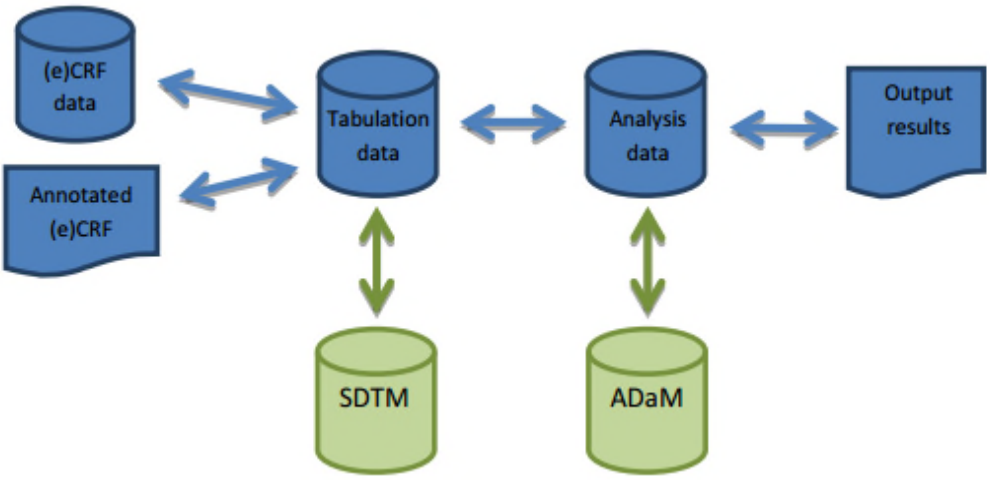

Figure 3 - Legacy Tabulation & Analysis Data Converted to SDTM & ADaM in Parallel

Per the TCG, issues observed when data flows in this manner include the following:

- Limited ability to determine location of collected CRF variables in the converted SDTM data unless the legacy aCRF is re-annotated.
- Limited traceable path from SDTM to the legacy analysis data.
- Limited ability to replicate/confirm legacy analysis datasets (i.e., analysis variable imputation or derived variables) using SDTM datasets.
- Limited ability to confirm derivation of intermediate analysis datasets or custom domains.
- Limited traceable path from SDTM to the ADaM datasets.
- Limited ability to replicate ADaM datasets (i.e., analysis variable imputation or derived variables) using SDTM datasets.
- Limited traceable path from ADaM to the Tables, Figures and the Clinical Study Report (CSR).
- Difficulty in understanding the source or derivation methods for imputed or derived variables in integrated/pooled data, supplemental qualifiers, and related records.

WHEN THE LEGACY DATA CONVERSION PLAN SHOULD BE CONSIDERED

Another situation when an LDCP may be needed is when a version of standardized data is no longer supported per the Data Standards Catalog (e.g. SDTM IG 3.1.1) and is up-versioned to a version that is supported per the catalog (e.g. SDTMIG v3.1.3, v3.2).

There are other situations in which a LCDP is appropriate. Every conversion of legacy data to standardized data is different. The decision falls to the sponsor to determine when the LDCP should be included in a reviewer's guide. If the reviewer cannot tell the story of the data via the eCRF, the define document, and the reviewer's guide, then the extra information per the LDCP should be included.

All eDC systems are not 100% CDASH conformant. One could make the argument that any study in which standardized data (i.e. SDTM, ADaM) is created would make the case to have the Legacy Data Conversion Plan & Report included in the reviewer's guide. That is not the intention of this document as the sponsor should make the determination on the extent of the information that should be in the document.

WHAT ELSE SHOULD ACCOMPANY THE LEGACY DATA CONVERSION PLAN

The Technical Conformance Guide also includes guidance on other pieces of supporting information that should be provided if a sponsor chooses to submit a Legacy Data Conversion Report for a given study. These would include the following:

- Two versions of the aCRF: One version that traces the collection field to the legacy variables and one that links the collection field to the correct SDTM variable
- Legacy tabulation and analysis datasets may need to be provided in addition to the standardized data to support traceability to the standardized data, legacy CSR, and/or TLFs

These assets associated with the legacy data would be placed in the appropriate sections of the eCTD pertaining to legacy data, /tabulations/legacy for tabulation data and /analysis/legacy for analysis data.

THE FDA WEIGHS IN – WHEN THEY TRULY EXPECT A LDCP

During a meeting of the PhUSE working group dedicated to developing a template, completion guidelines and examples for the Legacy Data Conversion Plan & Report in early January 2017, the discussion centered around the fact that a literal interpretation of the text in Section 8.3.2 of the Study Data TCG would lead many to believe that this report must be included for virtually every study, as almost all data goes through some sort of transformation from collection through to tabulation, analysis and reporting. Members of the FDA present at this meeting stressed that, while they are not in a position to dictate data collection methods, their biggest challenge is understanding how data conforming to FDA endorsed data

standards came into being, particularly if the original study data collection did not utilize CDISC controlled terminology and other FDA endorsed dictionaries such as MedDRA & WHODrug and/or did not capture data utilizing FDA endorsed data standards, for example, following the CDISC Clinical Data Acquisition Standards Harmonization (CDASH) data collection principles.

The FDA went on to further describe the situation where a Legacy Data Conversion Plan & Report would be needed:

- The data definition file (define.xml) cannot adequately explain where data present in SDTM came from
- The aCRF to SDTM is not clear enough on how variables were transformed from how they were collected to how they were represented in SDTM
- The reviewer's guide cannot adequately document the migration strategy and results without using the Legacy Data Conversion Plan & Report section of the guide to convey this information

STUDY DATA TRACEABILITY GUIDE

The FDA, in their efforts to establish traceability of data from collection through to analysis, have requested that aspects of traceability be documented at varying levels across a number of documents included in a regulatory submission. The Study Data Standardization Plan, described earlier in this document, not only records the standards utilized at the study level but also includes sections on how sets of pooled data are produced from study level data as well as basic traceability documentation associated with clarifying why multiple versions of data are present in a submission. The Study Data Reviewer's Guide, associated with a set of SDTM data, contains a question, "Was [this data] utilized to produce the analysis datasets?" Interestingly, while this question may make sense on the surface, it does not take into account the role the particular set of SDTM domain datasets play in the generation of the CSR; they might be on the critical path to CSR results, alternatively they might be produced solely to support pooling and standardization efforts as the CSR results were produced at an earlier point in time using legacy format tabulation and analysis data. The Legacy Data Conversion Plan & Report, extensively reviewed in this paper, serves as a key document to document a particular aspect of traceability, the transformation of previously collected non-standardized data to data based on FDA endorsed data standards.

While all of these documents nobly attempt to document traceability at an atomic level, they don't succeed in fully establishing traceability, either independently or collectively. At the same PhUSE meeting described earlier the concept of introducing documentation that gathered all traceability concepts into a single location was considered and met with enthusiasm by FDA representatives and that it should be further explored.

At minimum, a Study Data Traceability Guide would contain, for each individual study:

- Documentation of each set of data that has been created, including its purpose at the time it was created
- The source of information contained in each set of data
- High level description of transformations that occurred as one set of data was created from another set of data

There are many details that need to be worked out with respect to documenting study data traceability:

- Is this a living document, similar to the Study Data Standardization Plan, that is first developed at the time original sets of data are conceived and developed in support of some activity? For example, do you first create this document at the time that source data, tabulation data and analysis data are designed and implemented in support of the study CSR, and then subsequently updated if new sets of data are produced for other purposes, such as pooling / standardization on an integrated level or developed for specialized analysis purposes such as population PK analysis?

- Do you aggregate this information at the study level or submission level? How do you avoid repetition across study level documents but at the same time ensure that all aspects of traceability are adequately and accurately documented?

As this concept is further developed we will continue to explore if this information needs to stand alone in its own document, if it is developed as a document at the study level but ultimately incorporated into something else at the time of submission, such as a section or appendix to the Study Data Standardization Plan. This will be influenced most profoundly by the uptake of developing and maintaining this type of documentation by Sponsors & Service Providers and prevailing views of this type of information as core study level documentation versus its relevance only at the time of submission of an application for regulatory review.

CONCLUSION

As the FDA and other regulatory agencies move from reports and other documents to data and other machine readable sources of information to facilitate regulatory reviews, there will be an increasing push for both adherence to standards and transparency with respect to the provenance of submission assets. We see documents such as the Legacy Data Conversion Plan filling a need as we move from proprietary data collection standards and practices to methods rooted in the utilization of FDA endorsed data standards in the design and implementation of clinical trials at the point of data collection. The concept of study data traceability should follow a similar path, although it will be interesting to see the impact that the use of electronic health/medical records (EHRs/EMRs) might have on this process, giving all of these concepts new life as we move towards standardizing data practices across all aspects of medical information.

REFERENCES

[1] U.S. Food & Drug Administration, *Providing Regulatory Submissions in Electronic Format - Standardized Study Data: Guidance for Industry*, December 2014, Accessed April 21st, 2017 - https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM292334.pdf.

[2] U.S. Food & Drug Administration, *Study Data Technical Conformance Guide* (current version at time of access – v3.3 / March 2017), Accessed April 21st, 2017 - https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf.

[3] U.S. Food & Drug Administration, *FDA Data Standards Catalog* (current version at time of access - v4.5.1 / August 31st, 2016), Accessed April 21st, 2017 - https://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM340684.xlsx.

ACKNOWLEDGMENTS

The authors would like to thank all of the individuals who have attended the PhUSE/CSS annual meetings and participated in the various teams within the Optimizing the Use of Data Standards working group. Their efforts have made this paper and many of the tools we use today to document study data traceability a reality.

RECOMMENDED READING

The FDA documents listed in the references above provide the regulatory framework for how the FDA expects traceability to be maintained, documented and provided to the for review. The efforts of many PhUSE working groups dedicated to operationalizing many of the FDA's concepts can provide further helpful insight:

- PhUSE Working Group – Optimizing the Use of Data Standards - http://www.phuse.eu/optimizing-data-standards
- PhUSE efforts on the Study Data Standardization Plan - http://www.phusewiki.org/wiki/index.php?title=Study_Data_Standardization_Plan_(SDSP)

- PhUSE efforts on the Legacy Data Conversion Plan & Report - http://www.phusewiki.org/wiki/index.php?title=Legacy_Data_Conversion_Plan_%26_Report

CONTACT INFORMATION

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

Dave Izard
Chiltern International, Ltd.
david.izard@chiltern.com

Kristin Kelly
Merck & Co., Inc.
kristin.kelly@merck.com

Jane Lozano
Eli Lilly & Company
j.a.lozano@lilly.com