PhUSE 2014

Paper RG06

Traceability: Plan Ahead for Future Needs

Sandra Minjoe, Accenture Life Sciences, San Bruno, USA
Tanja Petrowitsch, Bayer Pharma AG, Wuppertal, Germany

ABSTRACT

Can you determine which records from which dataset(s) and what procedure(s) were used to create a table? Do you know how analysis values and records were derived from collected data? Do you know where a field from a CRF can be found in the data? Put another way, can you trace your data, from collection, through each dataset, and onto your output tables?

Traceability is something that, if not built into the process of data and output development, can be difficult to add later. However, when traceability is included, it helps streamline the development of programs, enables efficient quality control, assists in future analysis needs, and allows for easier review.

This paper describes some simple ways to incorporate traceability into the dataset and output development process, and elaborates on some of the benefits seen when traceability is incorporated. It includes references to documents that we, as two of the co-leads of the Computational Sciences Symposium working group on Traceability and Data Flow, helped develop and post to the wiki.

INTRODUCTION

At the time this document was written, Wikipedia's definition of traceability reads: "Traceability is the ability to verify the history, location, or application of an item by means of documented recorded identification." In terms of clinical trials, we talk about traceability at each step along the way between data collection and summarization, as shown in this diagram from the Computational Sciences Symposium (CSS) Traceability and Data Flow project draft white paper "Best Practices for Basic Linear Data Flow":

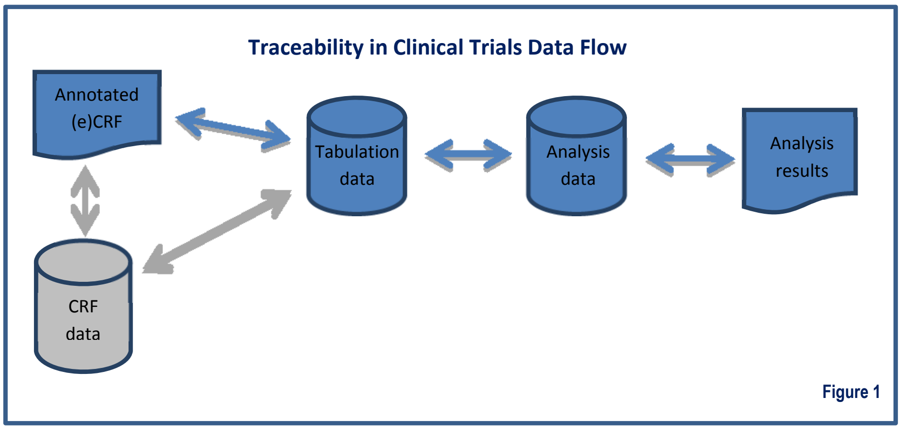


Figure 1

In this figure, traceability is represented as the arrows between each piece of the actual content.

We typically develop traceability from left to right, in the same order that we develop the corresponding materials we are tracing between. However, reviewers of these materials are often more interested in seeing traceability from right to left, in order to answer the question "Where did this come from?" In this paper traceability is described in both directions along the path, from collection to results and from results to collection, as shown above by the double-sided arrows.

USES OF TRACEABILITY

We often think about including traceability specifically for submission, but it can be useful for a variety of purposes.

AS PART OF THE QUALITY PROCESS

Specifications at all the levels shown in Figure 1 allow for a well-structured and complete set of documentation. Each data transition is described in study specifications: At a minimum, this suite of study specifications describes the relationship between each of the components in Figure 1, allowing a company to demonstrate that they did what they said they would do, and allows for tasks such as double-programming to take place.

By creating and saving this documentation, it can be pulled out ready-to-use years later, such as when creating a multi-study integrated analysis. On the other hand, if data is saved without the corresponding traceability, it can be difficult or even impossible to determine relationships. Thus, proper documentation in the form of traceability is a large component of quality. Some traceability can even be incorporated into the data itself.

Programs and other tools that were developed to actually convert data or create analysis results can be used to provide traceability, but there can be some issues with this option. First of all, programs often rely on underlying macros and other tools, which are not commonly stored at the study-level, making this an incomplete set of documentation. Additionally, they can be rather cumbersome to read through and understand. For these reasons, specifications are often a better solution.

WHEN SHARING DATA

Often it isn't just internal people who need to understand a study. For example, a sponsor company may outsource a study to a vendor, and the vendor then needs to provide not only the final products shown in Figure 1, but also the traceability or connections between each component.

When an integrated analysis uses studies created by multiple sponsors and/or vendors, it's helpful to have the study-level traceability in a similar format. Instead of maintaining the study-level traceability in various documents, it would be better to have the content standardized. A straightforward way to accomplish this is to transfer the specifications into a study-level define.pdf or define.xml file. This complete and standard documentation will help partners, vendors, and clients quickly review and understand the study data.

WHEN SUBMITTING TO A REGULATORY AGENCY

Submission to a regulatory agency is really just a common example of sharing data, described in the section above, but with more regulations.

Currently the United States Food and Drug Administration (USFDA) is the only regulatory agency in the world that requires the submission of clinical trial data. This means that they receive data and traceability documents from companies all over the world. In order to streamline their review process, they have been sending out strong messages about the need for standards in these areas. USFDA has also been working with both the Clinical Data Interchange Standards Consortium (CDISC) and the Computational Sciences Symposium (CSS) group to develop and maintain standards that they can use internally. On the USFDA website, there is a list of standards that can be accepted, including standards for traceability. A define file, plus additional documentation such as a data reviewers guide, are used to satisfy their traceability needs.

Other agencies, like PMDA and EMA, are now also considering accepting data with their submissions. If data submission becomes a requirement outside of the USA, it can only be expected that those agencies would have similar traceability needs and requirements as the USFDA.

TYPES OF TRACEABILITY

As described in the usage section above, specifications can provide a lot of traceability. The different types of specification documents, as well as other forms of traceability, are described within this section.

TRACEABILITY BETWEEN COLLECTION AND TABULATION

The first part of Figure 1 shows the relationships between the annotated (e)CRF, CRF (or raw) data, and the tabulation data:

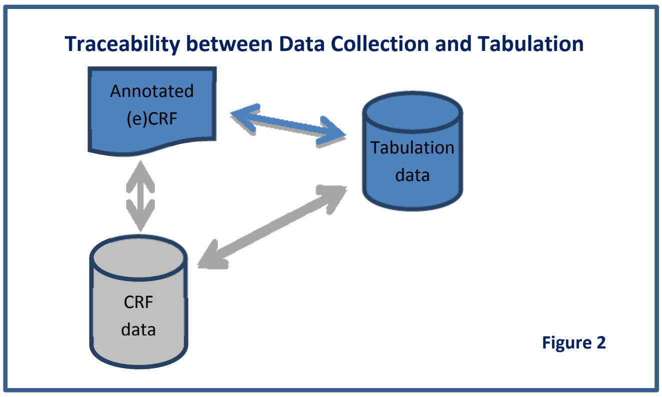


Figure 2

There are three components shown in Figure 2, each with traceability arrows between them. Although we may store CRF data separately from tabulation data, we typically submit only the tabulation data. Components and the traceability arrows that are primarily for internal use are shown in grey, and those that would also be used when sharing data are in blue. Each traceability step is described separately here:

- On the far left, the blank (e)CRF is annotated to describe the content location within the CRF (raw) data. This is typically for internal use only, developed when setting up the study CRF database.

- Along the top of the diagram, a similar blank (e)CRF annotation describes the content location within the tabulation database. This annotation is useful when sharing data, and may not be developed until data is ready to share.

- At the bottom of the diagram, traceability between the CRF (raw) data and the tabulation data is documented. Tabulation data specifications are typically for internal use only, developed at the time the tabulation data is created.

Annotation of the (e)CRF, as part of the traceability for both CRF (raw) data and tabulation data, includes at least the dataset name and variable name in the corresponding data structure. When possible, such as when a category or visit variable is pre-printed on the form, it also includes the actual text of the variable.

If someone reviewing the data in the tabulation dataset has a question about how it was collected, the (e)CRF annotated for the tabulation data provides the location of that data on, for example, the paper form where it was collected. Said another way, without the tabulation annotated (e)CRF, it would be virtually impossible to determine where to find this data.

Some data in the tabulation datasets can be derived. Because it is unusual to derive tabulation data, we can include a flag to denote that a row is derived. In this way we can include some traceability within the data itself, not only via some external metadata.

Following standard conventions for dataset names, variable names, variable labels, and content will help make data more easily understood. This is true within each set of data (within CRF data and within tabulation data), and also across the sets of data. Whenever possible, using the same dataset structures, variable names, and variable labels in the CRF data as in the tabulation data will add clarity and simplify documentation. In other words, minimizing the amount of transformations between the CRF data and the tabulation data helps with traceability.

As part of finalizing a study, especially when preparing it for submission to the USFDA, companies typically create a define file, with the (e)CRF annotations (shown in blue in Figure 2) as part of the input. For example, the mapping directions for a variable described on the annotated (e)CRF might be copied to the Source/Derivation/Comment column of that variable in the define file.

Another document created to describe the (e)CRF, collected data, and tabulation data is the Study Data Reviewers Guide (SDRG). Unlike the define file, which basically shows content in a set of tables, the SDRG is a text document. This document is a good way to provide additional traceability information that doesn't have a place in the define file but is helpful to communicate. Example content includes the version of standards used, the results of compliance checks against those standards, and whether screen failures are included in the data sets. Templates, instructions, and SDRG examples can be downloaded for free from the PhUSE wiki.

When using CDISC structures, Clinical Data Acquisition Standards Harmonization (CDASH) is the standard for CRF (raw) data, Study Data Tabulation Model (SDTM) is the standard for tabulation data, and define.xml is the standard for the define file. These standards, like all of the CDISC standards, are available for free from the CDISC website under the Standards tab.

TRACEABILITY BETWEEN TABULATION AND ANALYSIS DATA
The middle section of **Figure 1** shows the relationship between tabulation and analysis data:

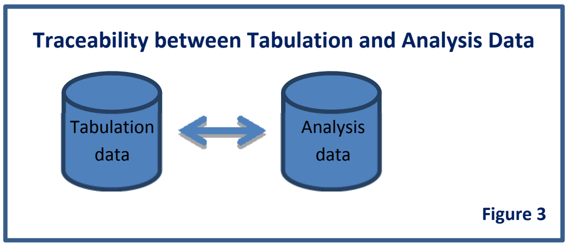


Figure 3

The arrow in Figure 3 shows the traceability between analysis data and tabulation data. Tabulation data is structured so that each collected data point appears only once across the suite of datasets that makes up a study. Analysis data, on the other hand, is compiled to contain all the necessary information for a specific set of analyses to make it analysis-ready; this means some information, such as population flags, will be found repeated across multiple analysis datasets. Tabulation data and analysis data each serve different purposes and thus will contain different information, even when based on the same concept, such as tabulation vs. analysis laboratory data.

Analysis data is commonly built off of tabulation data. Analysis data specifications provide the instructions to a programmer on how to create the analysis dataset, with details at the dataset, variable, and even value level.

These specifications can also be used to provide traceability between the tabulation datasets and the analysis datasets. If someone reviewing the analysis data has a question about how it was derived and which tabulation variables or observations were used, the content in the analysis dataset specifications provide that link. Without this information, many analysis variables are difficult if not impossible to understand or replicate.

As part of finalizing a study, especially when preparing the study for submission to the USFDA, companies typically create a define file, with the analysis dataset specifications as part of the input. For example, the derivation algorithm for a variable in the analysis dataset specifications might be copied to the Source/Derivation/Comment column of that variable in the define file.

Another document created to describe the relationship between the tabulation data and the analysis data is the Analysis Data Reviewers Guide (ADRG). The ADRG is similar to the SDRG mentioned in the section above. Unlike the define file, which basically shows content in a set of tables, the ADRG is a text document. This is a good way to provide additional traceability information that doesn't have a place in the define file but is helpful to communicate. Example content includes the version of standards used, the results of compliance checks against those standards, and whether screen failures were included in the data sets. Templates, instructions, and ADRG examples can be downloaded for free from the PhUSE wiki.

In addition to metadata, we can imbed some traceability into the data itself. When rows of analysis data map back to tabulation data, bringing some of the tabulation data forward can aid in traceability. For example, if a sequence number is maintained in the tabulation data, including this same sequence number in the analysis data provides traceability between these datasets at the row level. Alternately, analysis rows that have been created and don't map back to tabulation data can include a variable to denote that the row is derived.

Additional variables in an analysis dataset can be included just for traceability rather than for any analysis needs. Bringing forward some of the variables from the tabulation data can show similarities and differences from the analysis data. For example, if a partial character date exists in the tabulation and is imputed to a numeric full date for analysis purposes, including the original character date alongside the analysis numeric date is a simple way to quickly show where imputation was done.

When using CDISC structures, the Study Data Tabulation Model (SDTM) is the standard for tabulation data and the Analysis Data Model (ADaM) is the standard for analysis data. CDISC also provides a standard define.xml. These standards, like all the CDISC standards, are available for free from the CDISC website under the Standards tab.

TRACEABILITY BETWEEN ANALYSIS DATA AND ANALYSIS RESULTS

The last section of **Figure 1** shows the relationship between data tabulation and analysis results.

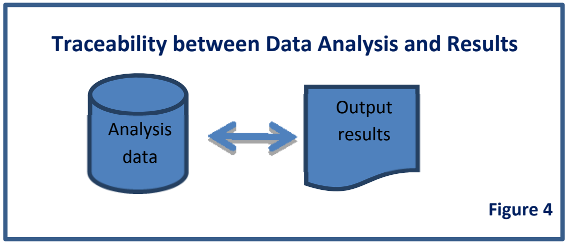


Figure 4

As described in the previous section, analysis data is created specifically for use in the generation of analysis results. In Figure 4 we are focusing on the need for traceability between these two pieces.

Analysis results specifications provide instructions to a programmer on how to create the output, including which dataset, variables, and even values to use. Table mockups, references to specific statistical models, and the location of input analysis data can all be part of these specifications.

Analysis results specifications can also be used to provide traceability between the analysis datasets and the analysis results. If someone reviewing the analysis results has a question about how the results were derived, including which statistical function and which analysis dataset, variables, or observations were used, the content in the analysis results specifications provides that link. Without this information, many analysis results are difficult if not impossible to understand or replicate.

In practice, analysis results specifications are often not well structured and not usually saved with the other study specifications and data. A reviewer instead tries to determine some of this information from the Statistical Analysis Plan (SAP) and/or footnotes included on the table. Not surprisingly, many reviewers often have questions about how a particular value on a table was generated. This gap in results traceability is currently being addressed by CDISC in a define.xml 2.0 extension, as described later this section.

Some traceability can be included within the dataset itself. For example, numeric visits can be created to represent the month, such as 1, 2, 4, 8, and 12, rather than just an ordinal number of 1, 2, 3, 4, and 5 that would only be useful for sorting. Flags can be included in the data to signify which rows are used for particular tables. Following standard conventions for dataset names, variable names, variable labels, and content can also help make data more easily understood and traceable.

When using CDISC structures, the Analysis Data Model (ADaM) is the standard for analysis data, and this data is used for analysis result generation. ADaM metadata includes not only dataset references via dataset-, variable-, and value-level metadata, but also results-level metadata. Results-level metadata is not required, but is useful for providing a link between each analysis dataset and the results created from it. Results-level metadata was not included by CDISC as part of the define.xml 2.0, but as of this writing an extension is being written to add this. As noted previously, the ADaM and define.xml standards, like all the CDISC standards, are available for free from the CDISC website under the Standards tab.

The ADRG, as described in the TRACEABILITY BETWEEN TABULATION AND ANALYSIS DATA section above, can also provide additional traceability between the analysis data and the analysis results.

COMPUTATIONAL SCIENCES SYMPOSIUM WORK

The Computational Sciences Symposium (CSS) has several working groups, one of which is called “Optimizing the Use of Data Standards” that has developed material directly related to traceability.

REVIEWER GUIDES

The “Optimizing the Use of Data Standards” working group has produced two documents that are useful for adding traceability to a submission: the Study Data Reviewers Guide (SDRG) and the Analysis Data Reviewers Guide (ADRG). These documents were each described separately above as part of the discussion on types of traceability, and now here together for reference:

- The SDRG provides a standard structure for describing information about the tabulation data that isn't adequately covered in the define file accompanying SDTM data. This includes document mapping decisions and sponsor-defined domains. The intent of the SDRG is that it would be submitted in addition to the SDTM data and define.xml, to provide further traceability.

- The ADRG provides a standard structure for describing information about the analysis data and reporting that isn't adequately covered in the define file accompanying ADaM data. This includes human-readable documentation of analysis methods, data sets, and programs. The intent of the ADRG is that it would be submitted in addition to the ADaM data and define.xml, to provide further traceability.

Content of these documents includes the version of all standards and dictionaries, whether screen failure subjects are included in any of the datasets, and results of any conformance checks that were run. Each document is intended to provide a quick overview of the clinical study, and not repeat details found elsewhere. Templates, instructions, and examples of each of these reviewer guides can be downloaded for free from the PhUSE wiki.

TRACEABILITY AND DATA FLOW

The “Optimizing the Use of Data Standards” working group has a project called “Traceability and Data Flow”. As of this writing, this group has several white papers finalized or in progress:

- “Traceability: Current State Analysis” (final)

- “Preliminary Recommendations for Traceability Documentation using Define-XML 2.0” (final)

- “Traceability: Best Practices for Basic Linear Data Flow” (final, with update in progress)

- “Study level traceability for FDA needs” (in progress)

- “Integration traceability for FDA needs” (in progress)

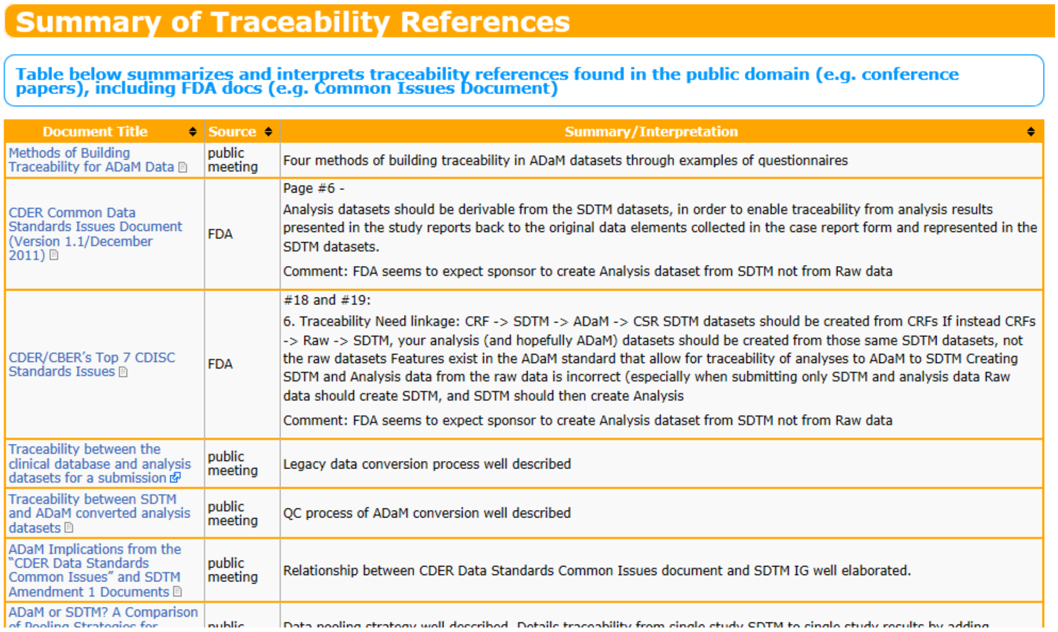
Additionally, a team is developing a “Legacy Data Conversion Plan” template, instructions, and at least one example to describe studies that were, for example, started without the use of CDISC standards but converted to CDISC later.

All of the final documents are available to download for free from the wiki. Each of the current in-progress documents will be made available for download when finalized.

We encourage the reader of this paper to download and use the suggestions contained within these white papers. Of particular use is the Basic Linear Data Flow white paper, which contains many examples of traceability that can be used at various places in the process, via both the data itself and with external metadata.

To find other references to traceability, including how and when to use it, consider reviewing the “Summary of Traceability References”. This is simply a list of documents we've found across industry that make reference to traceability. The list includes publications from industry meetings and documents that have come out of regulatory agencies (such as USFDA) and standards groups (such as CDISC). Not only does it include a link to the actual resource, but also a summary of what the document says about traceability. It is helpful for gaining an understanding of what our industry has been and is now saying, including opinions and recommendations, on this topic of traceability.

Here is a screen shot showing some of the first few items included on the list at the time of this writing:



Summary of Traceability References

Table below summarizes and interprets traceability references found in the public domain (e.g. conference papers), including FDA docs (e.g. Common Issues Document)

Document Title	Source	Summary/Interpretation
Methods of Building Traceability for ADaM Data	public meeting	Four methods of building traceability in ADaM datasets through examples of questionnaires
CDER Common Data Standards Issues Document (Version 1.1/December 2011)	FDA	Page #6 - Analysis datasets should be derivable from the SDTM datasets, in order to enable traceability from analysis results presented in the study reports back to the original data elements collected in the case report form and represented in the SDTM datasets. Comment: FDA seems to expect sponsor to create Analysis dataset from SDTM not from Raw data
CDER/CBER's Top 7 CDISC Standards Issues	FDA	#18 and #19: 6. Traceability Need linkage: CRF -> SDTM -> ADaM -> CSR SDTM datasets should be created from CRFs If instead CRFs -> Raw -> SDTM, your analysis (and hopefully ADaM) datasets should be created from those same SDTM datasets, not the raw datasets Features exist in the ADaM standard that allow for traceability of analyses to ADaM to SDTM Creating SDTM and Analysis data from the raw data is incorrect (especially when submitting only SDTM and analysis data Raw data should create SDTM, and SDTM should then create Analysis Comment: FDA seems to expect sponsor to create Analysis dataset from SDTM not from Raw data
Traceability between the clinical database and analysis datasets for a submission	public meeting	Legacy data conversion process well described
Traceability between SDTM and ADaM converted analysis datasets	public meeting	QC process of ADaM conversion well described
ADaM Implications from the "CDER Data Standards Common Issues" and SDTM Amendment 1 Documents	public meeting	Relationship between CDER Data Standards Common Issues document and SDTM IG well elaborated.
ADaM or SDTM? A Comparison of Pooling Strategies for	public	Data pooling strategy well described. Details traceability from single study SDTM to single study results by adding

Figure 5

This reference list is updated periodically, as new documents come out. However, because it is maintained by volunteers, completeness isn't guaranteed. In fact, if you are aware of any missing documents, we encourage you to send that information to us so that we can have it added to this list.

DEVELOPING MATERIAL

Everyone involved in creating the CSS documents are volunteers who do this in addition to their day jobs. If you have an interest in working on any in-progress items, please contact the authors or others on the leadership team listed on the wiki. Note that while some of the volunteers who contribute to the development of these materials also attend the spring CSS meeting in Maryland, USA, it is by no means a requirement to do so. You can help by contributing/developing material offline and attending regular teleconferences.

TRACEABILITY IN SUBMISSIONS

As mentioned earlier in this paper, one need for traceability is with submissions to a regulatory agency. In a submission, we want to include material that will not only provide the required data and documentation, but also allow understanding of the data flow and traceability. The Computational Sciences Symposium (CSS) Traceability and Data Flow project draft white paper "Best Practices for Basic Linear Data Flow" contains this figure that includes both a define file and reviewers guide in addition to each set of data:

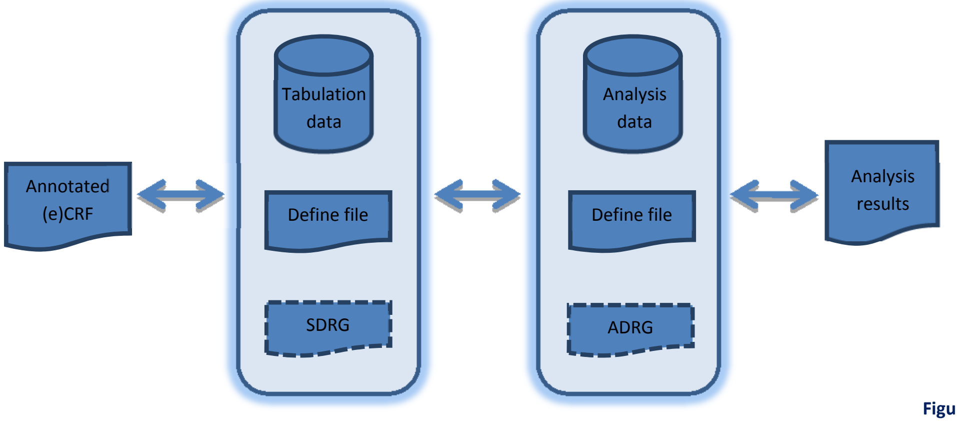


Figure 6

Although the SDRG and ADRG documents shown in Figure 6 are not required for submission to USFDA, they help represent the full set of traceability components. These standard documents were each designed to help address questions that might surface when data is reviewed. We recommend including them whenever the corresponding data is submitted.

Also note that Figure 6 applies to any submission that follows a linear data flow, whether using CDISC or not. When using CDISC, replace the "Tabulation data" with SDTM and the "Analysis data" with ADaM.

TRACEABILITY AS PART OF THE PROCESS

When traceability is not part of the process, it can be a lot of work to add it later. For example, when study data and analysis results are found without the corresponding traceability documentation, we can only guess at what was done. Maybe we create programs to try to reproduce analysis results, and then do a comparison to see if these results match what was originally reported. This is not only time-consuming, but also error-prone. Alternately, when traceability documents are developed as part of the process, they are, instead, a complete and accurate representation of what was done.

We've describe in this paper some of the documents that can be used to show traceability, including annotated (e)CRFs, specifications, define files, and data reviewer guides. All of this information is really metadata, or data about data. It can be created as part of our specifications, but much of it is also used in a slightly different form for traceability.

Metadata is often stored in Excel spreadsheets, such as when creating analysis dataset specifications, and then copied to other traceability documents, such as the define.xml for analysis data. A growing industry trend is to collect this metadata in a true metadata repository, rather than in spreadsheets. A repository allows for different views of the same underlying data to be used for different purposes. A metadata repository also allows you better control and consistency across studies, and makes the creation of filing documents, such as the define.xml, much more trivial.

We can't stress enough the value of creating traceability as part of the process, in whatever form you can. Documenting what was done as you do it involves very little time, adds value right away in terms of quality, provides reviewers the material they need to understand the data and analysis, and gives future users all the information they need when revisiting the study such as for integration analysis. Put another way, skipping traceability might save a few hours short term, but can cost days, weeks, or even months of work later.

CONCLUSION

Traceability has become an important component of our workload. It allows us to follow the data from collection to output tables and back. It is not only helpful for a submission to a regulatory agency, but also when sharing data and for internal use. Many different types of documents can be used to show traceability, including annotated (e)CRFs, specifications, data reviewer guides, and define files. Additionally, sometimes variables within the data itself can be used to aid in traceability. A growing industry trend is to make use of a metadata repository, allowing metadata to be captured once and used for all of our different traceability needs, including both specification and submission materials. Providing traceability documents along with the data itself help address questions that arise during review. It is much faster and less error-prone to create traceability documentation as part of the development process rather than to try to add it later.

REFERENCES

CDISC website: http://cdisc.org/

CSS Optimizing the Use of Data Standards wiki:
http://www.phusewiki.org/wiki/index.php?title=Optimizing_the_Use_of_Data_Standards

CSS Traceability and Data Flow project: http://www.phusewiki.org/wiki/index.php?title=Traceability_and_Data_Flow

CSS traceability references: http://www.phusewiki.org/wiki/index.php?title=Summary_of_Traceability_References

CSS wiki: http://www.phusewiki.org/wiki/index.php?title=FDA_Working_Groups

EMA: European Medicines Agency - Transparency - Release of data from clinical trials

PMDA: "Overview of CDISC implementation at PMDA" presented at CDISC European Interchange 2014

USFDA Study Data Standards Resources: http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm

Wikipedia definition of traceability: http://en.wikipedia.org/wiki/Traceability

ACKNOWLEDGMENTS

We would like to thank all the volunteers in the CSS working groups, especially those involved in the Optimizing the Use of Data Standards. Their contributions to the white papers and other reference materials included on the CSS wiki were invaluable to creating this paper.

CONTACT INFORMATION

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

Sandra Minjoe
Accenture Life Sciences
Sandra.Minjoe@Accenture.com

Tanja Petrowitsch
Bayer Pharma AG
Tanja.Petrowitsch@Bayer.com