CDISC SDTM/ADaM Pilot Project[1]
Project Report

Executive Summary

Background

CDISC is a non-profit, multidisciplinary consensus based standards development organization founded over a decade ago that has established open, worldwide biopharmaceutical data standards to advance the continued improvement of public health by enabling efficiencies in medical research. In addition to enabling FDA submissions, many other advantages come with the use of clinical data standards. Research studies have shown that standards enhance the performance of clinical studies in several other key areas such improved internal data warehousing, data integration and data transport, as well as enabling collaborative research through timely and efficient data-sharing.

The collective power and borderless innovation provided by the CDISC constituency is well represented by the performance of this CDISC SDTM/ADaM Pilot Project to generate an ICH E3/eCTD clinical study report (CSR) using the CDISC data models.

Pilot Project Goals

The CDISC SDTM / ADaM Pilot Project was conducted as a collaborative pilot project with FDA and Industry. The objective of the pilot project was to test how well the submission of CDISC-adherent datasets and associated metadata met the needs and the expectations of both medical and statistical FDA reviewers[2]. In doing this, the project also assessed the data structure, resources and processes needed to transform source data into the SDTM and ADaM formats and to create the associated metadata.

Overview

This report documents the efforts made by the Pilot Project core team to successfully accomplish the above stated objectives. The legacy data used in the Pilot Project was provided by Eli Lilly and Company from a phase II clinical trial. Each step of the pilot process and work completed are easily followed in this report beginning with the de-identification of the pilot legacy data, application of CDISC Standards (including SDTM, ADaM, and CRTDDS), and resulting in the creation of a CDISC-compliant electronic clinical study report submission.

---

[1] This Pilot Project is also referred to as "Pilot 1." It was conducted during 2006 and 2007.

[2] **Disclaimer: All comments, statements, and opinions attributed in this document to the regulatory (FDA) review team reflect views of those individuals conveyed as informal feedback to the pilot project team, and must not be taken to represent guidance, policy, or evaluation from the Food and Drug Administration.**

This pilot project effort represented an unprecedented amount of work and collaboration between CDISC[3], the Industry and FDA and led to a number of valuable learnings. These learnings are documented in this report in Section 6, and were presented at the 2006 and 2007 CDISC Interchange conferences.

Conclusion

All of the aforementioned goals were met by the CDISC SDTM/ADaM pilot project. The project established that the package submitted using CDISC standards met the needs and the expectations of both medical and statistical reviewers participating on the regulatory review team. The regulatory review team noted the importance of having both data in SDTM format to support the use of FDA review systems and interactive review, and data in ADaM format to support analytic review. The project also demonstrated the importance of having documentation of the data (e.g., the metadata provided in the data definition file) that provides clear, unambiguous communication of the science and statistics of the trial.

The regulatory review team expressed a favorable impression of the pilot submission package. They were optimistic about the impact that data standards will have on the work associated with their review of new drug applications.

[3] **Disclaimer: As defined in CDISC Core Principles, CDISC standards support the scientific nature of research and allow for flexibility in scientific content; however, CDISC does not make the scientific decisions nor drive scientific content; rather, our primary purpose is to improve process efficiency and provide a means to ensure that submissions are easily interpreted, understood and navigated by medical and regulatory reviewers.**

Table of Contents

1. Introduction

Submission of data to the Food and Drug Administration (FDA) has been necessary for years in order for the FDA to conduct a thorough review and electronic submission of data will likely become a regulation in the future. The Clinical Data Interchange Standards Consortium (CDISC) is a non-profit, multidisciplinary consensus based standards development organization founded over a decade ago and has established open, worldwide biopharmaceutical data standards to advance the continued improvement of public health by enabling efficiencies in medical research. During this 10-year period, CDISC has focused considerable effort on developing standards to help FDA in its review and approval process of safety and efficacy data. To this end, the CDISC data models have been successfully used to help FDA better understand industry data, by providing a platform of standard data content. This standard data minimizes programming and rework of the data during FDA review, and greatly facilitates the integration and reuse of data from multiple submissions for broader scientific and medical evaluation.

The development of CDISC standards has been informed by descriptions of FDA reviewers' needs expressed by FDA Liaisons. Over time, the Submission Data Tabulation Model (SDTM) and the Analysis Data Model (ADaM) have matured to the point that references to them in industry forums are now common. The standards have garnered the attention of the mainstream of the pharmaceutical industry, which is working on ways to implement these standards in hopes of streamlining submission and facilitating review of the data. CDISC recognizes that the unity and interoperability of data standards is a necessity for both the submission and the review and approval process.

This report describes the CDISC SDTM/ADaM Pilot Project, hereafter referred to as the "pilot project." The objective of the pilot project was to test the effectiveness of data submitted to FDA using CDISC standards in meeting the needs and the expectations of both medical and statistical FDA reviewers. In doing this, the project would also assess the data structure/architecture, resources and processes needed to transform data from legacy datasets into the SDTM and ADaM formats and to create the associated metadata.

1.1. *Outline of this pilot project report*

This project report is intended to describe the pilot submission package and the processes followed, including the decisions made to produce the package, and lessons learned from the experiences of the pilot and from feedback from the regulatory review team.

A basic outline of this project report is:

- Section 1 provides an overview of the pilot project and details about the report itself.
- Section 2 describes the process followed by the pilot project team in creating the pilot submission package, including the datasets, the analysis results, and the various documents included in the pilot submission package.
- Section 3 focuses on the metadata - how it was collected and its use in the project.
- Section 4 provides details about the Define file created by the pilot project team.

- Section 5 describes the interactions and communications between the pilot project team and the regulatory review team.
- Section 6 summarizes the key points and outstanding issues noted in the report.
- Section 7 contains the appendixes of the report.
  - The first appendix summarizes key points about the management of the pilot project, including a list of participants.
  - The second appendix provides an overview of the repository used by the pilot project team.
  - The remaining appendixes supplement the information in the body of the report with more detailed information.
  - A list of abbreviations used in the project report is also included.

1.1.1. Additional pilot project materials available

The revised pilot submission package is available to CDISC members (on the "members-only" section of the CDISC webpage) for use as an example of the application of the CDISC standards. The programming code used to generate the pilot submission package is not included, as some of it was proprietary to corporate sponsors. However, as detailed a description as possible of how the work was done is included in this report. As stated previously, the processes followed by the pilot project team were often dictated by the timelines and constraints of the project, and were not necessarily "best practice" or even "good practice."

The reviewer's guide and cover letter included with the pilot submission package provide additional helpful information. (Both documents were included in the same PDF file, with appropriate bookmarks.)

In addition, various presentations about this pilot project have been made during 2006-2007; those presentations can be found on the CDISC webpage in the Publications and Presentations section (http://www.cdisc.org/publications/index.html). For example, presentations made during the 2006 CDISC Interchange are available at that location.

1.2. *Terms and phrases used in the report*

"Define file" refers to the data definition file, which is the roadmap for the submission package. It is the file containing the metadata for the tabulation and analysis datasets, as well as the analysis results metadata. The file can be in portable document format (PDF) as traditionally used, or an extensible markup language (XML) format, as recommended in current guidance[4], and is referred to as the Define.pdf or Define.xml file, respectively. The Define.xml file is also known as the Case Report Tabulation Data Definition Specification (CRT-DDS).

[4] April 2006 FDA guidance regarding regulatory submissions in electronic format ("Guidance for Industry: Providing Regulatory Submissions in Electronic Format – Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications, April 2006, Electronic Submissions, Revision 1," and the associated document "Study Data Specifications"). Refer to the following website: http://www.fda.gov/cder/regulatory/ersr/ectd.htm

The CDISC Define.xml team has written a document specifying the standard for providing Case Report Tabulations Data Definitions in an XML format for submission to regulatory authorities (e.g., FDA). The XML schema used to define the expected structure for these XML files is an extension to the CDISC Operational Data Model (ODM).

The term "SAS transport files" refers to SAS® XPORT (version 5) transport files (XPT), i.e., data in the SAS XPORT Transport format[5].

Tabulation datasets contain the data collected during a study, organized by clinical domain. These datasets conform to the CDISC Submission Data Standards (SDS), as described in the CDISC Study Data Tabulation Model. The SDTM was developed by the CDISC Submissions Data Standards (SDS) team, and precursors to the SDTM were called SDS standards. The terms "tabulation dataset" and "SDTM dataset" are used interchangeably in this document.

Analysis datasets contain the data used for statistical analysis and reporting by the sponsor. The Analysis Data Model describes the general structure, metadata, content, and accompanying documentation pertaining to analysis datasets. The terms "analysis dataset" and "ADaM dataset" are used interchangeably in this document.

The term "pilot project team" refers to the group of individuals from industry who worked on the pilot project. Refer to Appendix 7.1.1 for a list of pilot project team members.

The term "regulatory review team" refers to the group of FDA volunteers who participated on this pilot project, providing input and feedback based on their areas of expertise and interest. The views expressed by these volunteers are their own opinions and experience and are not, necessarily, those of FDA. Refer to Appendix 7.1.1 for a list of regulatory review team members and contributors.

Refer to Appendix 7.9 for a list of the abbreviations used in this report.

1.3. *Description of the project*

In April 2005 two of the CDISC Board members, Edward Helton (SAS Institute) and Stephen Ruberg (Eli Lilly and Company), discussed the concept of a pilot project to test the use of the CDISC standards. They developed a draft charter for the project, which CDISC leadership approved. According to that document, the objectives of the project included:

- Assess the data structure/architecture, resources and interoperability needed to transform data from legacy datasets into the CDISC SDTM and/or ADaM formats.
- Perform case studies that demonstrate the effective transformation of legacy data into CDISC SDTM domains and ADaM datasets and their associated metadata. A case study (or series of studies) would allow CDISC to understand the use of SDTM in submission of derived data and the specific needs for separate ADaM datasets/programs. CDISC wants to repetitively test and learn the very best application of its interoperable standards to meet the industry regulatory data submission needs.

[5] SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

- Gather the input, evaluation, and review from a group of FDA reviewers, in a collaborative software environment, of real clinical trial data based on the CDISC standard.
- Assess the boundaries between SDTM and the parallel elements in ADaM. Understand the requirements and working relationships between observed data, derived data, specific analysis datasets, and program files.

Optional or "next step" objectives were to explore submission of data using an XML file format versus the SAS® System XPORT format and to explore the use of ODM and the CRT-DDS (also called Define.xml) in providing metadata for the submission package.

The presentation of the proposal for the pilot project occurred at the CDISC Interchange Meeting in September 2005. The pilot project team was identified and the first team meeting held in November 2005. Table 1 provides highlights of the timeline between the CDISC Interchange Meetings in 2005 and 2006 and the receipt of the regulatory review team's final comments on the pilot submission package.

Table 1 Timeline for CDISC SDTM/ADaM Pilot Project

November 18, 2005	First pilot project team teleconference
January 25, 2006	Planning meeting with CDISC Board representatives
February 17, 2006	Legacy study documents (redacted protocol, abbreviated study report, case report form) provided to pilot project team
February 28, 2006	Face-to-face kick off meeting for the project, included roundtable discussion with regulatory team members
April 10, 2006	Pre-submission encounter with FDA participants
April 19, 2006	De-identified legacy data provided to pilot project team
June 30, 2006	Submission package sent to the regulatory review team
August 28, 2006	Pilot project team received regulatory review team's comments
September 26, 2006	Announcement of results at CDISC Interchange
February 13, 2007	Revised submission package sent to regulatory review team
April 4, 2007	Pilot project team received regulatory review team's comments on revised submission package

The timelines for the project were driven by the early agreement (at the January 2006 planning meeting) that results would be reported at the CDISC Interchange 2006 conference. To achieve this deadline, the pilot submission package needed to be sent to the regulatory review team by the end of June 2006. All activities in producing the pilot submission package were geared towards meeting that target date.

It was agreed at the January planning meeting that the primary focus of the pilot project would be to produce a submission package as an example of the application of the CDISC standards, and that FDA statistical and medical reviewers would evaluate the submitted datasets (SDTM and ADaM), metadata and documentation. The phrase "pilot submission package" will refer to this submission package in this report. Additionally, the team identified a set of success criteria to help assess the overall efficacy of the pilot submission package from the perspective of the regulatory review team. These criteria were: 1) is the submission evaluable with current tools; 2) can the reviewers reproduce the analyses and derivations; and 3) can the reviewers easily navigate through the pilot submission package.

The goal of the pilot project was not to prove or disprove efficacy and safety of a drug; therefore not all components of the legacy study (referred to as Study CDISCPILOT01) discussed in the legacy protocol were included in the pilot submission package. The pilot submission package included one abbreviated study report that documented the pilot project team's analyses of the legacy data. The purpose of providing a study report was to test the summarizing of results and the linking to the metadata, as well as providing results or findings for the regulatory review team to review and/or reproduce. Accompanying the study report were the tabulation datasets, analysis datasets, Define.xml files containing all associated metadata, an annotated case report form (aCRF), and a reviewer's guide.

With the objectives of the pilot project in mind, the completeness of the pilot submission package was considered adequate for the purpose of this pilot project by the regulatory review team; however the pilot submission package falls far short of the standard requirements for a complete application to market a new drug or biologic. The pilot submission package is for illustration only; there is no intention to imply in any way that it constitutes a complete submission package.

The regulatory review team had a favorable overall impression of the pilot submission package. Through several meetings (teleconference and face-to-face), the individuals participating on the review team provided constructive feedback and specific details of what they considered best practices with regard to the content, structure, and format of clinical study reports (CSRs), the clinical data, and the metadata that describe the clinical data. Although the regulatory review team was generally pleased with the original pilot submission package, they noted a few issues. The primary issues related to functionality available for the Define.xml file and the format and structure of the analysis datasets. The pilot project team and the regulatory review team agreed that a revised pilot submission package would be created, to address these issues as much as possible. The pilot project team sent the revised pilot submission package to the regulatory review team in February 2007 and received comments back in April 2007. Based on a small survey among the regulatory review team, the issues with functionality and navigation of the Define file appeared to have been addressed. The feedback from the regulatory review team regarding the revised analysis datasets was positive, stating that the revised versions are a good illustration of what information is critical to understanding the lineage of the data from case report form (CRF) to analysis.

1.4. *Caveats*

The pilot project team was primarily focused on the "What" (i.e., content) of a CDISC-adherent submission, not the "How" (i.e., process). Although the "How" (i.e., process) was addressed in the efforts of the pilot project team, optimizing the process was not a focus of the project due to a variety of factors (refer to Section 2), including the fact that the amount of time available to produce the pilot submission package was shorter than envisioned. Tight timelines affected the project because the reasons for choosing certain *ad hoc* processes were often that they were the fastest "good" processes to implement rather than the "preferred" process. Difficulties with process are not necessarily inherent in the standards; indeed, these issues might not exist with better tools and more time to think about processes. Therefore, one should not interpret the processes described in this report as the only, or the best, way to proceed with the creation of a submission using the CDISC standards.

CDISC is moving towards having a harmonized set of standards. The experiences gained in this pilot project, and in future projects, promise to be very helpful in furthering integration of standards. Accordingly, some *ad hoc* decisions were required to facilitate integration for the pilot package. While these decisions resulted in a "legitimate" submission using the standards available at the time, the resulting product does not necessarily represent a future version of the standards. For example, the pilot submission used extensions to the Define file (as described in Appendix 7.5.5) that may not necessarily be incorporated into future versions of the Define.xml standard. One of the purposes of this project report is to explain the various decisions made by the pilot project team and the implications of those decisions.

Clearly, the pilot project differs from a "real-world" creation of a package for submission to FDA. Wherever possible, the report highlights these differences so that readers will not assume that CDISC or the pilot project team advocates real-world use of these processes. For example, the use of MedDRA terms in the pilot submission was constrained under the terms of an agreement with the MSSO, which controls licensing of MedDRA, as described in Section 2.2.3 of this report.

Readers should note that this pilot project did not examine how the CDISC standards interact with every aspect of clinical data processing and review. For example, the pilot project did not test whether certain sets of required, expected, and permissible variables in SDTM were more useful to the review process than other sets. In addition, since the pilot project used only one clinical trial from one therapeutic area, it did not address the question of how well the CDISC standards would apply to clinical trials in general. One of the benefits of standard data is the possibility of combining data across different submissions. This pilot project did not have the data or the resources necessary to test this benefit. By using only one team to produce the submission, this pilot did not test the reproducibility of the CDISC standards across multiple teams.

As noted throughout this report, all comments, statements, and opinions attributed in this document to the regulatory (FDA) review team reflect views of those individuals conveyed as informal feedback to the pilot project team, and must not be taken to represent guidance, policy, or evaluation from the Food and Drug Administration.

1.5. *Orientation to the legacy study*

This section of the report provides a brief orientation to the legacy data used in this pilot project. Full descriptions of the legacy study are in the protocol, found in Appendix 1 of the CSR.

The study was a prospective, randomized, multi-center, double blind, placebo-controlled, parallel-group study conducted on an outpatient basis. Patients with probable mild to moderate Alzheimer's disease were to be studied in a 3-arm, placebo-controlled trial of 26 weeks duration. The objectives of the study were to evaluate the efficacy and safety of two doses of active drug as compared to placebo.

The scales used to assess efficacy in this pilot project were:

- Alzheimer's Disease Assessment Scale - Cognitive Subscale, total of 11 items [ADAS-Cog (11)]
- Clinician's Interview-based Impression of Change (CIBIC+)

- Revised Neuropsychiatric Inventory (NPI-X)

Safety was assessed using:

- Adverse events
- Vital signs (weight, standing and supine blood pressure, heart rate)
- Laboratory evaluations

2. Process

2.1. *General description*

Figure 1 illustrates the content and general structure of the pilot submission package submitted to the regulatory review team. Note that the blue rounded rectangles represent folders, with the text in the box providing information rather than the precise folder names described in the eCTD specification; not all folders are illustrated.

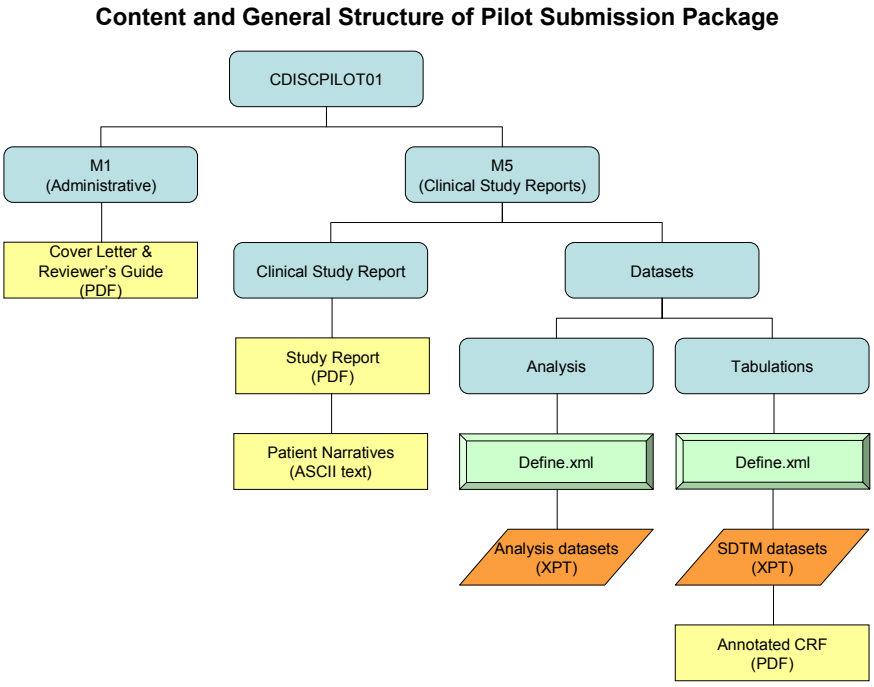


Figure 1 Pilot Submission Package Structure

An agreement reached early in the pilot project was that the emphasis of this first pilot project would be the final product – the actual pilot submission package, rather than the process of creating it. Several factors influenced the decision to focus on "what" instead of "how":

- Before an attempt can be made to provide guidance for process, it was important to first verify that the CDISC standards themselves met the needs of reviewers.
- How and when the CDISC standards are applied will be very sponsor-specific.
- Having an example to work from and to use for discussion is important for future process discussions.

- It was understood that producing the pilot submission package might necessitate the use of "coat hangers, duct tape, and bandages" to get everything to harmonize properly. These patches would definitely not be part of a recommended process, but would facilitate meeting the timelines.
- Future pilot projects will build on the work done for this pilot project.

Consequently, the process described here is only a basis for future development – both to consolidate things that worked well and to avoid or improve on things that worked poorly. To provide that basis, this report includes detailed descriptions of the processes used in this pilot project, including the rationale for various decisions as appropriate.

Figure 2 illustrates a general outline of the process followed by the pilot project team. The term "derived data" refers to data that involve calculations or manipulations of the CRF data. At the onset of the project, it was agreed that the tabulation datasets (i.e., SDTM datasets) would be created from the legacy data, with only a very minimum amount of derived data included. These datasets, referred to as "SDTM-without-derived," were the input for the creation of the analysis datasets. With one exception, analysis results were based on analysis datasets; the concomitant medications summary was based on SDTM datasets, as described in Section 2.7. The pilot team wanted to test the utility of including derived data in SDTM, so a set of potentially useful variables in the analysis datasets were selected for inclusion in SDTM. The origins of these variables were to be identified as variables in the analysis datasets and appropriate links provided. A separate step in the process added these derived data to create the "SDTM-with-derived" tabulation datasets submitted to the regulatory review team. Quality control conducted by the pilot project team verified that the derived data incorporated in the SDTM datasets were consistent with the original data in the analysis datasets. Refer to Section 2.6 for more details regarding including derived data in the tabulation datasets.

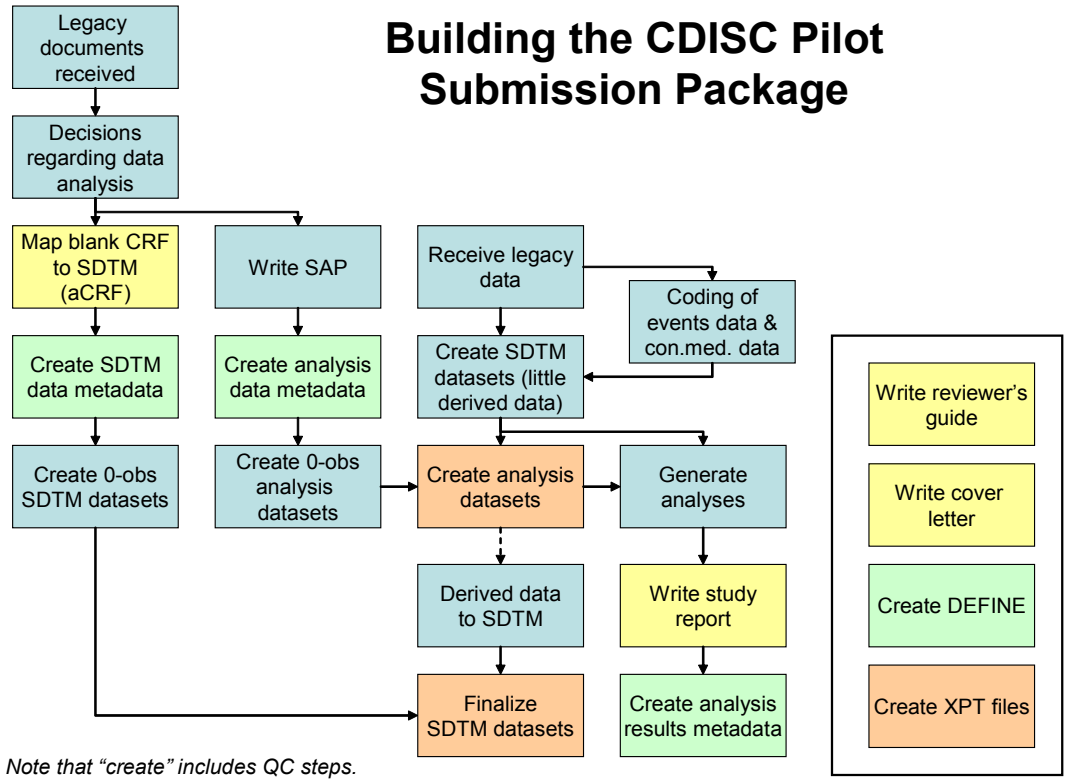


Figure 2: Process followed in building the pilot submission package

2.2. *Data and tools used*

2.2.1. Legacy data

Eli Lilly and Company (the Legacy Sponsor) provided the legacy data used in CDISCPILOT01 for the purposes of this pilot project. De-identification of the data and redaction of documents occurred prior to release to the pilot project team. De-identification included changing dates of data elements while maintaining all chronological relationships and sequences within the data elements for each subject (e.g., no change in the relationship of timing of adverse events with respect to dosing).

The protocol provided is from the original study, although redacted. The statistical analysis plan created specifically for this study as part of the pilot project included descriptions of deviations from the protocol-specified analyses. The statistical analysis plan (included as Appendix 9 of the CSR) also describes some additional analyses included to test other aspects of the standards.

This pilot project did not reproduce all of the Legacy Sponsor's analyses and reports, nor did it include all of the data from the legacy study. Instead, the pilot project addressed only the more common elements of a submission. These included the primary and some secondary safety data, the primary efficacy endpoints, a few secondary efficacy endpoints, and a representative set of analyses of these endpoints as specified in the statistical analysis plan (SAP).

2.2.2. Standards / tools used

The creation of this pilot submission package involved the following standards:

- SDTM Implementation Guide Version 3.1.1
- SDTM Version 1.1
- Analysis Data Model Version 2.0 (referred to as "ADaM v2" in this document) as issued for public comment in March, 2006 (note that the ADaM Implementation Guide was not available at the time of this pilot project)
- CRT-DDS version 1.0
- ODM version 1.3 (public comment period closed on May 2, 2006)

Consistent with the direction of CDISC and at the request of the regulatory review team, the data definition tables were provided in XML format (CRT-DDS) for greatest flexibility. The datasets provided were SAS Version 5 transport files.

The XML schema provided for Define.xml in the pilot submission package is an extension of the ODM 1.3 schema, with new elements added to support the ADaM analysis results metadata. This extension is only illustrative of how analysis results metadata could be implemented. The schemas will likely change when formally vetted by the CDISC standards teams.

A style sheet presents the XML in a human-readable format via a web browser. Members of the pilot project team developed the style sheet used for the pilot submission package. (Refer to Section 4.2 for more details.) It illustrates what the pilot project team thought to be a reasonably functional and desirable presentation of the CRT-DDS in a web browser. The present rendering resembles the traditional Define.pdf, but this is not a requirement.

Several software packages and tools were used in the production of the pilot submission package, and the pilot project team particularly appreciates the vendors who provided products and support for their use. However, to avoid any implication of endorsement of one vendor or system over another by CDISC or the pilot project team, no specific vendor mention will be made in this report.

2.2.3. MedDRA coding of event data

At the request of the regulatory review team, the legacy event data were coded using the Medical Dictionary for Regulatory Activities (MedDRA). Because the data were intended to be available to a wide audience, it was necessary to obtain agreement from the MSSO (Maintenance and Support Services Organization) regarding the use of MedDRA. The MSSO serves as the repository, maintainer, and distributor of MedDRA.

The MSSO's general policy is to limit the public distribution of MedDRA to a very small subset of MedDRA (i.e., 100 or fewer terms). This is done to protect the investment of MedDRA subscribers.

The MSSO stated that this pilot project test is in the best interest of their user community and that the use of MedDRA should not be a limiting factor. Consequently, for the case of the CDISC SDTM/ADaM Pilot Project, the MSSO agreed to allow CDISC to utilize MedDRA with the following limitations:

- All MedDRA terms except the lower level term, preferred term, and system organ class were to be masked in the pilot submission package.
- CDISC were to identify and inform the MSSO of the fixed period of time that this pilot program will be in effect. This is simply an identification of a fixed period of time for the pilot project and the use of MedDRA in the pilot project, not a limitation.
- The total number of terms lower level terms and preferred terms used in this pilot project would not exceed 10,000 terms.

MedDRA version 8.0 was the coding dictionary used for the adverse event data.

The regulatory review team requested that all five levels of MedDRA coding be included in the tabulation datasets. The three levels not currently included in the SDTM adverse event (AE) model [Higher Level Group Term (HLGT), Higher Level Term (HLT), Lower Level Term (LLT)] were included in the supplemental qualifiers domain for AE (i.e. SUPPAE). To protect the copyright and licensing agreement of MedDRA non-informative terms masked the actual values of HLGT and HLT (e.g. HLGT_0152, HLT_0617). The pilot project team also chose to mask the AE verbatim text, replacing the actual text with a randomly generated coded text (e.g. "VERBATIM_0013") with each unique term corresponding to unique coded text.

It is important to note that due to the considerations outlined above, the coding of adverse events for this project was NOT consistent with MedDRA coding rules and conventions. It is important to clarify that in submissions sponsors should adhere to the rules of the dictionary used in the submission.

2.2.4. Process for concomitant medication coding with WHODD

The coding of concomitant medications used a sample of the World Health Organization Drug Dictionary (WHODD) downloaded on 25 April 2006 (http://www.umc-products.com/DynPage.aspx?id=2844). The sample WHODD was used to perform concomitant medication coding. The coding process involved creating a single dataset from the medicinal product, ingredient, therapeutic group, substance, and anatomical therapeutic chemical (ATC) code ASCII files. The merging by drug name of this dataset with the SDTM concomitant medication (CM) domain produced coded terms. Since this is a sample dictionary, coded terms were not available for all medication records.

2.3. *Annotating the CRF*

At the time of the pilot project, the SDS metadata team was drafting an appendix to the SDTM implementation guide called Metadata Submission Guidelines. In annotating the CRF, the pilot project team followed the advice in this draft document; refer to Appendix 7.2 for more detailed information on the creation of the annotations.

Each page where data were collected and reported was annotated. References to annotations on other pages (e.g., "see visit 1") were not used to provide information on the origin of variables.

Links from the Define.xml to the blank CRF could have been established via hard links to one or more page numbers or via a PDF Advanced Search that would provide a reviewer with all "hits" for the searched values. The pilot submission package implemented the search

capability as well as providing the more traditional links to the appropriate page numbers in the blank CRF. The pilot project team elected to do both so that the familiar method of referring to the blank CRF was also available to the review team. The reviewer's guide sent with the pilot submission package explained that the Acrobat Advanced Search, using the "Search Comments" option, would facilitate finding annotations more efficiently. By combining "Search Comments" and "Whole Words", a reviewer could find all variables for a particular domain using the 2-letter domain prefix that was placed in the "Subject" field.

The comments could also be printed by using that option in Adobe Acrobat (i.e., select to print the document and then select the comments option). A note explaining this additional attribute of comments should probably have been included in the reviewer's guide, to make reviewers aware of the functionality.

The CRF was annotated with "Not Entered in Database" on those pages/panels/date entry fields where data were not reported in the datasets due to data de-identification. (This was not done in the original pilot submission package and the oversight was noted by the regulatory review team and corrected in the revised pilot submission package.)

2.4. *Creation of SDTM datasets from the legacy data*

The mapping of legacy data to SDTM began with the creation of a blueprint for converting the data, a tabular document referred to as the "mapping-specifications document." If a legacy variable was needed for the SDTM data then the target dataset(s) and variable(s), as well as any other pertinent information needed for the conversion, were recorded in the appropriate cells. If the variable was not contributing to the SDTM data, then "NOT MAPPED" was indicated. The mapping-specifications document also contained all of the controlled-terminology that needed in the SDTM data. Figure 3 shows such a screenshot of this mapping-specification document.

Mapping Specifications Document

Source Dataset	Variable	Data Type	Label	Target Domain	Target Variable	Mapping Comments
DEMOG	VISIT	Char	Scheduled Visit	SC,DC	VISITNUM	
DEMOG	UNVISIT	Char	Unscheduled Visit	NOT MAPPED		
DEMOG	SSSEX	Char	Sex	DM	SEX	
DEMOG	ORGIN	Char	Origin Code	DM	RACE	When 'CA' then 'CAUCASIAN'; When 'AF' then "AFRICAN DESCENT (NEGRO,BLACK)' when 'EA' then 'EAST/SOUTHEST ASIAN (BURMESE, CHINESE, JAPANSE, KOREAN, MONGOLIAN, VIETNAMESE)' when 'AB' then 'WEST ASIAN (PAKISTANI, INDIAN SUB-CONTINENT)' WHEN 'HP' THEN HISPANI
DEMOG	RESLTVAL	Char	Number of Years of Education	SC	SCORRES, SCSTRESN, SCSTRESC	SCTESTCD=YEARSEDU" and SCTEST="YEARS OF EDUCATION"
DEMOG	MMSESUM	Num	Baseline Severity	NOT MAPPED		
DEMOG	VSDTE	Num	Visit Date	DM,SC,DC	--DTC	
DEMOG	DIAGDATE	Num	Date of Onset of AD	DC	DCORRES, DCSTRESN, DCSTRESC	DCCAT=ALZHEIMER'S DISEASE HISTORY, DCTEST=DATE OF ONSET OF ALZHEIMER'S DISEASE,DCTESTCD=ADONSET
DEMOG	TRTMENT	Char	Treatment Assignment - Character	DM	ARM	If TRTMENT=" then ARMCD=SCRNFAIL and ARM=SCREEN FAILURE
DEMOG	TREAT	Num	Treatment Assignment - Numeric	DM	ARMCD	
DEMOG	AGE	Num	Age	DM	AGE, AGEU= YEARS	
DEMOG	CTPATNO	Num		DM, SC	USUBJID, SUBJID(DM)	USUBJID=01-INV-CTPATNO

Figure 3 Illustration of Mapping Specifications Document

Upon completion of the mapping specifications for each legacy domain, QC was performed on those specifications, corrections were made, and the cycle repeated until those specifications were considered final.

Only when the specifications for a legacy dataset were considered final did programming for that dataset commence. The data were uploaded into an ETL software tool where, armed with the mapping specification document and the annotated CRF, a developer was able to convert the data to SDTM.

The pilot project team agreed at the outset that the SDTM datasets initially produced from the legacy data would include only a minimum amount of "derived" data, meaning data that did not originate on the CRF. The pilot project team referred to these datasets as "SDTM-without-derived." The only derived data included in the tabulation datasets at the end of the ETL processing were unique subject identifiers (USUBJID), visit numbers (VISITNUM), visit names (VISIT), study day variables (--DY), and baseline flags (--BLFL).

Several programming tasks were done after the ETL processing, including removing data for screen failures from all domains other than DM, adding other derived data (questionnaire scores in the QS domain, population flags in SUPPDM, etc.), and at the very end, converting the datasets to SAS transport files.

Upon the finalization of each SDTM dataset, quality control (QC) checks verified that the data were mapped accurately and according to the specifications.

To simplify programming and data manipulation, the pilot project team elected to split the questionnaire domain (QS) into multiple datasets, based on questionnaire type, in such a way that concatenation of the datasets back into a single domain was possible. To facilitate this

reassembly of the dataset, QSSEQ was made unique across the entire set of split QS domains by the addition of a questionnaire-specific value to the sequence number. For example, by adding a questionnaire-specific value of 5000 to the sequence numbers of the records in the QSAD dataset, an original QSAD sequence number of 1 became 5001. The pilot submission package contained the re-assembled QS domain.

The pilot project team elected to order the variables in the SDTM datasets using the dataset's key (i.e., index) variables (as listed in the dataset metadata) and the order of variables used in the SDTM Implementation guide. The key variables were placed first, followed by the remaining variables. This variable ordering scheme was applied consistently for all domains in the pilot submission package.

2.5. *Analysis datasets*

The pilot project team decided that the most pragmatic approach to creating analysis datasets was to use the SDTM domains as input. This ensured that reviewers would be able to trace the creation of derived variables contained in the analysis datasets back to their source in the SDTM datasets and ensured that the analysis dataset creation programs would be of value if requested by the review team. The first step was to outline the analysis datasets that would be required to perform the primary and secondary efficacy and safety analyses. This allowed the pilot project team to identify the relevant CRF pages and SDTM domains, ensuring that all of the expected data would be mapped from the legacy datasets into SDTM.

Once the identification of the analyses to be included in the study report was complete, the specifications of the analysis datasets were developed.

An organized approach was used to create what are often referred to as "analysis dataset specifications." These specifications were essentially the metadata needed to document how a variable was derived, what sources (from the SDTM datasets) were used, and what decision rules and exceptions to these rules were used. These specifications were entered into a suite of Excel spreadsheets, described below in Section 7.4. The pilot project team used this prescriptive approach to creating the analysis datasets by defining the metadata first and then using this metadata to guide programming of the final analysis dataset. This approach ensured that the analysis datasets and the accompanying metadata in Define.xml were in harmony. It also facilitated the pilot project team's creation and quality control of the analysis datasets by providing analysis information per variable in a readable columnar format rather than relying on gleaning this information from the analysis dataset creation programs.

The development of the analysis datasets proceeded in a commonly used process as follows. The statistician responsible for the analysis dataset completed the metadata spreadsheet with a detailed description of all variables to be contained in the analysis dataset. The statistical programmer used these specifications to construct the analysis dataset. When the draft analysis dataset was available, the statistician validated that the derived variables, etc. were programmed correctly. If necessary, this process was iterated and iterations continued until a final analysis dataset was produced.

A required analysis dataset was the subject level analysis dataset (ADSL). As per ADaM v2, this analysis dataset had one record per subject and contained all of the important variables needed to describe a subject, such as values of baseline characteristics, treatment variables,

population indicators, clinical milestones, and completion status. This analysis dataset was used as input to other analysis datasets and thus was pivotal to the work stream.

The principles specified in the published ADaM v2 were utilized in this pilot project. However, in parallel to the work on the pilot project, the CDISC ADaM team was developing the ADaM Implementation Guide, which presents standards for the structure and content of analysis datasets, including standard variable names. Therefore, it should be kept in mind that the analysis datasets submitted with the pilot project represent the concepts in ADaM v2 but do not necessarily reflect those included in the ADaM Implementation Guide.

According to ADaM v2, analysis datasets only need to be provided for key (i.e., important) analyses, as defined and agreed upon by the sponsor and reviewers. The pilot project team provided analysis datasets for each analysis included in the pilot submission package with the exception of the concurrent medication summary. The pilot project team decided to provide analysis datasets for almost all their analyses, key or not, because the number of analyses in the pilot submission was relatively small and because they felt it was important to provide a broad range of illustrative examples. The concomitant medications summary was the only analysis for which an analysis dataset was not provided.

2.5.1. Issues addressed as a result of review team comments

The regulatory review team provided very helpful comments regarding the structure and documentation of the analysis datasets. For instance, the comments were considered by the ADaM team in the development of the ADaM Implementation Guide.

Comments from the regulatory review team regarding the original pilot submission package identified the following goals for the analysis datasets and the associated metadata:

- transparency regarding how values from the SDTM data were handled for the efficacy analysis data, specifically the primary efficacy analysis datasets
- an analysis dataset for the primary efficacy variable that would facilitate the production of a meaningful graph of the data
- an analysis dataset for the primary efficacy variable that would facilitate the exploration of the sensitivity of certain algorithms such as last observation carried forward (LOCF) and windowing as well as alternative statistical methodologies the reviewer might want to try
- a dataset that is structured and described in the DEFINE file in such a way as to make clear the rules applied for using the windowing and LOCF algorithms as described in the statistical analysis plan

The primary efficacy analysis datasets, ADQSADAS and ADQSCIBC, were used for the analysis of the ADAS-Cog and CIBIC questionnaire data, respectively.

The ADQSADAS and ADQSCIBC datasets included in the revised pilot submission package were structured as one record per outcome variable (ADAS-Cog total score and CIBIC score, respectively) per analysis visit per subject. (For the purposes of this pilot project, each of these datasets had only one outcome variable.) In addition, all observations from the QS tabulation dataset for the ADAS-Cog and CIBIC questionnaires were included in the corresponding analysis datasets, with flags included to identify records created by the LOCF

algorithm and by the windowing algorithm and those that were "as observed" (i.e., included with no changes from the tabulation dataset). The datasets are described in Appendix 7.3.

Additional changes to analysis datasets in the revised pilot submission package as a result of regulatory review team feedback included:

- Population flag variables were modified to contain either Y or N. No blank values were allowed.
- Dates of the first and the last dose were included in all analysis datasets
- All three variables containing treatment information were included in all analysis datasets (as opposed to only one or two of the variables). Within the pilot submission, the three treatment variables were TRTP, TRTPN, and TRTPCD (referring to the text, numeric, and coded versions, respectively, of the planned treatment).
- A flag variable was added to all relevant analysis datasets (i.e., all except ADSL and ADTTE) to indicate whether the observation occurred while the subject was on-treatment.
- Variables within each analysis dataset were ordered in a logical pattern, rather than alphabetically.

Changes to the metadata associated with the analysis datasets included changing the description of "structure" to be more consistent with that used in SDTM. For example, the structure of the LB domain was described as "one record per lab test per time point per visit per subject" in the metadata. The metadata description of structure for the lab analysis datasets (ADLBC and ADLBH) was changed from "one record per subject per visit per lab parameter" to "one record per lab test per visit per subject."

2.6. *Derived data in SDTM*

SDTM consists of "collected" data with limited derived data added (e.g. baseline flags). The purpose of having additional derived data in SDTM is to meet reviewers' needs for viewing data in domain structures rather than facilitate complicated analyses.

The pilot project team was charged with testing the adding of derived data to the tabulation datasets. The goals were to explore how derived data are added to the datasets, determine how the links between the SDTM and analysis datasets are expressed in the Define file, and assess how useful derived data would be to reviewers.

The pilot project team and the SDS Standards Review Committee discussed various options for adding derived data to SDTM, including: no additional derived data on SDTM, adding derived data as columns, adding derived data as rows, adding derived data as SUPPQUAL, and some combination of the last three. The agreed decision was to create a "SUPPQUAL column" for flags and add derived data as variables in SUPPQUAL. In addition, some derived data were added as rows in the dataset.

Consequently, derived data in the pilot project tabulation datasets included baseline and population flags, as described in the SDTM Implementation Guide, as well as:

- an adverse event treatment emergent flag (in SUPPAE)
- a total score for the ADAS-Cog(11) (added as a record in QS)
- an endpoint flag for lab data (in SUPPLB)

- a derived variable defined as result divided by upper limit of normal (i.e., LBTMSHI) for lab data (in SUPPLB)

While the SDTM supplemental qualifier datasets were not originally created with "numeric" qualifiers in mind, the pilot team chose to test the use of the supplemental qualifiers structure for the LBTMSHI variable.

In addition, as described in Section 2.2.3, three coding levels for MedDRA were included in the supplemental qualifiers domain for AE (i.e., SUPPAE).

There has been much debate over what process should be used to produce SDTM and ADaM datasets, including if and how derived variables should be incorporated into SDTM datasets. The pilot project team decided to add derived data to the SDTM datasets after the analysis datasets were created, using the same algorithms. Once the programming was complete, a separate QC was performed to ensure that the derived values were consistently represented in both the SDTM and the ADaM domains.

In adding the derived data to SDTM, some limitations of ODM with respect to providing a linkage in the Define file between SDTM and ADaM were identified. The intention was to provide a link in the metadata from the SDTM derived variable to the corresponding ("original") variable in the analysis dataset. The pilot project team found that the ability to link derived data in the tabulation datasets back to the analysis datasets was not available in the version of ODM being used. For example, in QS (the domain containing the questionnaire data), QSSTRESN contains CRF data and contains the derived total score from the corresponding analysis dataset. The "patch" for identifying this in the metadata was to use "Computational Algorithm or Method" to provide text describing the various sources for the value in the QS dataset. In this example, the text described that if the QSCAT variable is XXX and the QSDRVFL is set to yes, then the value in QSSTRESN was from the record containing the total score (computed using observed values) for the appropriate subject and visit in the corresponding analysis dataset; otherwise the value for QSSTRESN came from the CRF. The actual text used in the description was:

> *if QSDRVFL='Y' and the QS data pertain to ADAS-Cog or NPIX, then QSSTRESN is from ADQSADAS.ACTOT or ADQSNPIX.NPTOT, respectively, using the windowed data (i.e., where VISIT=AVISITC and ITYPE=' '), else if QSDRVFL = ' ' then QSSTRESN is from the CRF Page*

Given this limitation, the pilot project team elected to add the computational method for a minimum number of the SDTM variables. In addition, the content of the "Computational Algorithm or Method" and "Comment" columns in the pilot project define file differ from other examples in the public domain (where computational method is incorporated in the comments column). As noted in Section 4.4, reconciliation of differences between the elements recommended for the ADaM and SDTM metadata was still under discussion at the time of this pilot project. Since the pilot project team wanted to have a consistent format for the two sets of dataset metadata, with both using the same column headings, some "tweaking" of the information contained in existing metadata columns was necessary. Consequently, the pilot project team populated a minimum number of these fields within the SDTM dataset metadata as an illustration. In constructing a "real" define file, many other variables would include explanations of how they were derived (e.g., RFSTDTC, RFEDNDTC, AGE).

In attempting to address issues regarding derived data in the tabulation datasets, the pilot project team found that the meaning of the term "derived data" is not universally agreed. There is confusion between the term "derived data" and the use of the term "derived" for the origin in the SDTM metadata.

2.7. *Analysis results*

As noted previously, only the more common elements of a submission are addressed in the pilot submission package. These included the primary and some secondary safety data, the primary efficacy endpoints, a few secondary efficacy endpoints, and a representative set of analyses of these endpoints as specified in the SAP.

Once the analysis datasets were created, the programming for generating the analysis results was done. The analysis datasets were designed to be analysis-ready, consequently the generation of the results primarily consisted of the analysis procedure itself plus manipulation of the results into presentation format. The only exception to this was the summary of concomitant medications. This summary was programmed using the SDTM dataset as input. This provided an example of an analysis that did not have an analysis dataset included in the pilot submission package.

According to ADaM v2, analysis results metadata should be provided for key or difficult analyses. For illustrative purposes, analysis results metadata was created for every analysis included in the abbreviated report for this mock submission. Thus, the analysis results metadata in the pilot submission package provides a table of contents for all the analyses. The analysis results metadata includes a link to the relevant analysis dataset and a link to the section of the SAP where the specified analysis is described. In a real-world submission, the analysis results metadata would provide a table of contents for the key analyses, which would have been agreed between the sponsor and the reviewers.

2.8. *Writing the study report*

The CSR was based on the outline described in the guidance document entitled "E3: Structure and Content of Clinical Study Reports" from International Conference on Harmonisation (ICH) of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH E3). Text describing the study and the planned analyses were based on the legacy (redacted) protocol and the SAP written by the pilot project team.

ICH E3 includes descriptions of the format for the synopsis and the items to be included in the appendices. Tables included in the appendix were ordered and numbered in the same order in which results were described in the CSR. Hyperlinks to tables and figures (both in-text and in those in CSR Section 14) and to other sections of the report were included in the study report.

Because of the nature of this pilot project, several sections and appendices of the CSR were not completed. The incomplete sections and appendices were those that would not materially affect a reviewer's ability to assess the pilot submission package with respect to the goals of the project. Because the regulatory review team specifically requested that no listings be provided, no data listings were included in the appendix. The study report appendices included the redacted protocol, the statistical analysis plan created specifically for the pilot project, and the sample blank CRF. To maintain section and appendix numbering that would

be consistent with ICH E3, incomplete sections and appendices were included in the study report with a notation that text for that section or appendix was not included.

Raw statistical output from the primary efficacy analyses and from the repeated measures analysis were included in a subsection to Appendix 9 of the CSR, "Documentation of Statistical Methods," as requested by the regulatory review team. It was noted that statistical reviewers often expect raw statistical output from at least the primary efficacy analysis to be provided, and the provision of such output should be discussed and agreed between the sponsor and the reviewer. Such documentation is helpful for examining and understanding discrepancies between a reviewer's results and the results reported in the CSR.

2.9. *Assembling and publishing the pilot submission package*

The pilot project team decided that the submission format would be an eCTD/eNDA hybrid, utilizing PDF Table of Contents (TOCs) while maintaining the electronic common technical document (eCTD) folder structure. This decision was made to keep the submission "simple" and to keep the focus of the submission on the CDISC components, without the complications of an eCTD XML backbone.

Assembly began with identifying all the components that would comprise the pilot submission package. Study-specific components included the abbreviated clinical study report, the tabulation and analysis datasets, a Define.xml file for both the analysis and tabulations datasets, and an annotated CRF for the tabulation datasets. Also included in the pilot submission package were the necessary PDF TOCs, the schemas and style sheets required by the Define.xml documents, a cover letter, and a reviewer's guide. The cover letter and the reviewer's guide were concatenated into one PDF and submitted as the cover letter.

As each component was verified (for quality control) and considered final, it was placed into the publishing process. If necessary, the files were converted to PDF and/or concatenated with other PDFs. For all PDF documents, bookmarks and hyperlinks were added to facilitate navigation. For quality control, a review of each published PDF was performed, with corrections made as necessary. Finally, a QC review was performed on the entire pilot submission package, primarily to review the navigation, although content issues were addressed as well.

All pilot submission package components that were submitted in PDF format were converted using Acrobat 5 except for the annotated CRF. That file was created using Acrobat 7 because Acrobat 7 offered advanced searching capabilities. For additional information regarding the techniques used to annotate the CRF, see Section 2.3 and Appendix 7.2.

2.10. *Quality control*

To ensure high quality deliverables, the pilot project team applied QC processes to all the files, including mapping specifications, SDTM-without-derived datasets, analysis datasets, SDTM (including derived data) datasets, tables, and figures. These QC processes usually involved confirming the result through independent programming by another pilot project team member. Quality control of documents involved review by pilot project team members in addition to the authors. The pilot project team tested the Define.xml file by verifying links and content.

3. Metadata

The specifications for the analysis datasets and the SDTM datasets were written in metadata prescriptively, prior to developing the programs to create the analysis datasets. In contrast to a descriptive approach, this prescriptive approach leveraged the value of metadata by making the data specifications accessible by a suite of (SAS) macros that automated some processes of building and validating SDTM and analysis datasets as well as the accompanying Define.xml content. The analysis specifications and variable level metadata were entered into Excel spreadsheets. (Other options for collecting the information included data and catalog editors.) Software programming was used to convert the Excel spreadsheets into the following metadata elements:

- a dataset specifying dataset level attributes
- a dataset specifying variable level attributes
- a dataset specifying codes/decodes and valid values of variables
- a catalog containing entries that contain text descriptions and comments that could be attached to datasets, variables and other parameters
- a dataset specifying value-level information about variables that contained multiple types of data (e.g., vital signs result that might be blood pressure or heart rate)

These five metadata elements were then used to create an HTML file that included all the details required by a programmer to write a program to create the datasets. If any ambiguities or gaps in the data specification were identified by the programmers, the metadata was updated appropriately, and the HTML file recreated from the revised metadata. The metadata content was evaluated several times during the data build phases and kept consistent with the desired derived datasets. A programming macro used the attributes defined in the metadata to create 0-observation datasets. These 0-observation datasets thus conformed to the data specification in dataset names, dataset labels, variable names, variable labels, variable lengths, variable types, etc. As the last step of creating the final version of an analysis dataset, the programmer would append the data file created by the analysis dataset creation program to the appropriate 0-observation dataset, thus applying all pre-specified variable labels, lengths, types and variable content to the dataset. This process ensured that the Define file was consistent with the datasets described within it. The regulatory review team identified lack of consistency between the Define file and the data as a problem in many submissions. The process used by the pilot team addressed this regulatory concern, in addition to adding efficiency.

Other macros were used to help automate the many steps in the creation of analysis datasets. These included:

- A macro that created a format catalog containing formats created from the code/decode values defined in the data specification.
- A macro that sorted the observations in the datasets by the key variables identified in the metadata and re-ordered the variables within the dataset according to the variable order defined in the metadata. (Regulatory review team members expressed a preference for datasets whose variable order matched the order of variables in the Define file.)
- A macro that produced a report of the actual allocated lengths of all character variables along with the minimum length required to contain the maximum text string length. This

report helped to ensure that character variables were only as long as needed to contain the data values.

- A macro that compared the structure and attributes of the draft analysis datasets with the data specifications and compared the actual values found in variables with lists of allowed values in the metadata.
- A macro that used the metadata to generate the Define.xml file. The resultant XML file was syntactically validated by using an XML parser that compared the XML file to the CDISC ODM schema.

The SAS macros used in this process were developed by Gregory Steffens (Eli Lilly and Company) and can be found at the same location as the published pilot submission package.

The Define.xml file was also reviewed by the pilot project team and CDISC ODM/XML experts. A separate XML file was created for each of the two databases – the analysis database and the SDTM database. These two XML files were subsequently combined with each other and with the analysis results XML file to create a single XML file containing all of the dataset metadata.

The XML file created in the above process is a valid Define.xml file. As described in the next section, the pilot project Define.xml also includes some non-dataset metadata that were added in a subsequent step.

Figure 4 illustrates the process described above.

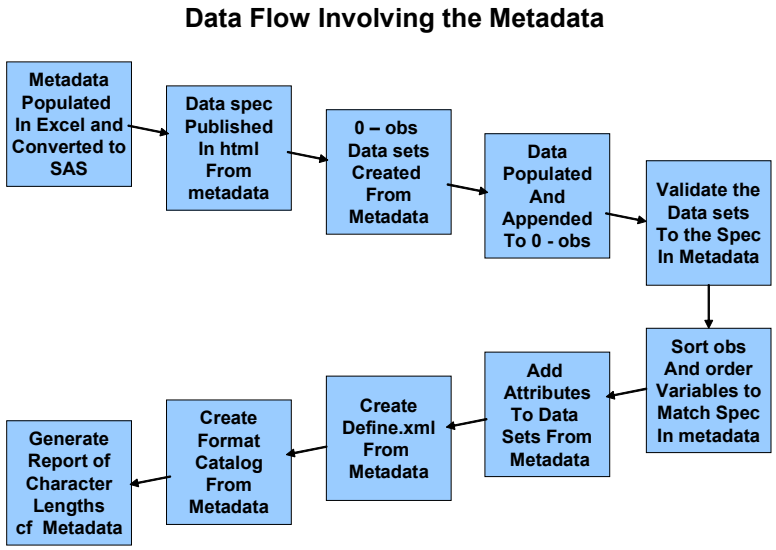


Figure 4 Illustration of steps followed in creation and use of metadata

Refer to Appendix 7.4 for screenshots of the Excel spreadsheets used to collect metadata components.

4. The pilot project Define.xml

4.1. *Overview*

The pilot submission package included a Define.xml file with supporting schema files and a style sheet for rendering that Define in a browser with the look of Define.pdf files. The schema files integrated metadata for the SDTM datasets, for the analysis datasets and for the analysis results.

4.2. *Appearance of the Define file*

The pilot project team decided it was best to provide a consistent interface to all three components of the Define file (SDTM dataset metadata, analysis dataset metadata, and analysis results metadata), and to provide one interface for all users.

The pilot project team provided a style sheet that accomplished the desired rendering in a web browser. (There is no "standard" style sheet advocated by FDA or CDISC.) The rendering of the Define.xml for the pilot project differed from prior Define.pdf-inspired renderings as follows:

- SDTM datasets, analysis datasets, and analysis results metadata were in separate sections of the Define, although the two dataset sections are similar in appearance.
- A brief table of contents, including, links to the reviewer's guide, analysis results metadata, analysis datasets, and SDTM datasets was provided at the top of the Define.xml.
- A powerful left-side navigation bar was included, although it could be used only with Internet Explorer.

These constructs assist a reviewer by organizing content into sections, and providing anchors for navigation to those sections. In Internet Explorer specifically, the left-side navigation bar allows the reader to move to any named section of the Define, from wherever they might currently be browsing. This is both more precise than the Table of Contents at the top of the Define.xml, and more direct. It also addresses a concern identified by the regulatory review team regarding the difficult navigation for the original pilot submission package.

Screenshots of the Define.xml are included in Appendix 7.5.1.

4.3. *Internal structure and creation of the Define file*

As noted earlier in this document, this pilot project was about "What" to create, not "How." In the case of the Define file, the "What" that the pilot project delivered concerns how the human reader interacts with the Define file, more than the fine details of the structure of the Define. However, some "how" details regarding the placement of the Define files in the pilot submission package, the structure of the Define, and the process for creating it are included in Appendix 7.5.

Appendix 7.5 is provided with the caveat that some portions of the pilot project implementation of Define.xml – while legitimate uses of the current standard – will undoubtedly change.

4.4. *Metadata implementation issues*

Because the pilot project team had to "cobble together" some things to produce the final pilot submission package, there are some idiosyncrasies within the metadata depicted in the Define file.

The analysis results metadata identifies the analysis dataset used for the analysis, often with a phrase identifying the appropriate records for the analysis.

As noted in Section 2.6, ODM does not currently provide mechanisms for linking from the derived data in the SDTM datasets back to the analysis datasets. The pilot project team filled this gap by including text in the "Computational Algorithm or Method" column that described the source and derivation of the derived data.

One of the ADaM core principles is that sponsors must provide clear and unambiguous communication of how a variable was derived. To facilitate this communication, ADaM supports providing metadata to describe the immediate predecessor for a variable as well as either a textual description of the derivation algorithm and/or a link to a software program or other documentation relative to the derivation. At the time of the pilot project, how this ADaM metadata would be coordinated with SDTM metadata and supported by ODM was still under development. Because the pilot project team decided to have a consistent format for the two sets of dataset metadata, with both using the same column headings, it was necessary to fit this information into existing metadata columns. The "Comment" column was therefore used in the analysis dataset metadata to identify the immediate predecessor data file. For the analysis dataset metadata, the "Origin" column identified the location of the first occurrence of the variable. The "Computational Algorithm or Method" column contained a hyperlink to a description of the derivation of the variable. For example, in the adverse event analysis dataset the treatment-emergent flag is created within the dataset (Origin=created here), using data from the dataset itself and from the subject-level analysis dataset (Comment=data from ADAE, ADSL), using the specified algorithm (hyperlink from Computational Algorithm or Method). Therefore, though the pilot project team devised a way to incorporate the desired information in the Define file, exactly how and where this information will be supplied in the future will be topics for future CDISC standards.

Because the Define file contains both the tabulation and analysis dataset metadata, with each given a separate and distinct portion of the Define file, the "purpose" column in the presentation of the dataset metadata did not contribute any useful information. In the future, consideration should perhaps be given to refining the contents of the column.

4.5. *Issues addressed as a result of review team comments*

When the issues noted by the regulatory review team with respect to the Define file (see Section 5.3.1) were explored, most problems were traced to the style sheet being used for the original pilot submission package. For the revised pilot submission package, major changes to the style sheet were implemented. Some of these modifications addressed errors, while others made it easier to navigate within the Define file. Additions that improved navigation included a bookmarks pane and links to the reviewer's guide.

When the original style sheet was used with Internet Explorer, the browser's Back button did not work consistently. This was traced to a known bug in Internet Explorer 6. This issue was

corrected by using a "framed" version of the style sheet (i.e., a version with a left-side navigation pane). The framed version works only with Internet Explorer, but offers much superior navigation capabilities. The non-framed version can be used with browsers other than Internet Explorer, but can be difficult to use with Internet Explorer 6, because of the bug mentioned earlier.

Some pilot project team members were unable to open the framed version of the Define file even when using Internet Explorer. This issue was traced to differences in internet browser settings, and the cause was ultimately traced to a difficulty caused by a reference, in the XML, to a specific version of Microsoft XML Services. When this specific reference was removed (as it has been in the public release of the pilot submission package), conflicts with users' internet browser settings were eliminated.

These issues illustrate the value of providing a sample of the data and define file to determine that the rendering provides the functionality expected.

A major issue identified by the regulatory review team was the difficulty in printing the Define file. The style sheet used in the pilot submission package was developed with the primary target of web browser rendering, which is not readily suited to printing. Reviewers who attempted to print the Define file found that the file did not fit on portrait pages, that page breaks were not clean, and that printing only a portion of the file was difficult. Opening the document in another application (e.g., Microsoft Word) provided a work-around, but was not an option that was user friendly or efficient. Instead, the pilot project team created a PDF file of the rendering that could be printed. This PDF file is not included in the public release of the pilot submission package because this solution required some non-standard procedures. As this shows, there is a need for XML standards evolution and accompanying tools that are accommodate the need for printing as well as screen rendering, without imposing further development work at style sheet-creation time.

4.6. *Issues to be addressed regarding metadata*

The pilot project team's implementation of the CDISC standards highlighted several metadata issues that require attention from the appropriate CDISC teams.

- Ability to link from the derived data in the SDTM datasets back to the analysis datasets (Section 2.6)
- Support of "source/computational methods" in analysis dataset metadata (Section 4.4)
- Support of value-level metadata in ODM/DEFINE (Section 7.5.10)
- Support of analysis results metadata in ODM/DEFINE (Section 7.5.5)

5. Interactions with the regulatory review team

Disclaimer: All comments, statements, and opinions attributed in this document to the regulatory (FDA) review team reflect views of those individuals conveyed as informal feedback to the pilot project team, and must not be taken to represent guidance, policy, or evaluation from the Food and Drug Administration.

One key factor in the success of the pilot project was the unprecedented level of interest and support by individuals at FDA. The regulatory review team participated in the

teleconferences and made time to meet with the pilot project team at several face-to-face meetings. At one of the face-to-face interactions with the regulatory review team, someone commented, "In order to get a standard we have to suffer." This became the unofficial mantra of the pilot project team.

5.1. *Identifying expectations and requirements*

A face-to-face meeting was held in February 2006 to kick-off the work on the pilot project. At this meeting, thirteen volunteers from FDA participated in a roundtable discussion of reviewer expectations and requirements. These volunteers included both statistical and medical reviewers, as well as data management and technical support experts. This discussion set the tone for many of the decisions made in the pilot project. The key messages from the discussion were:

- Consistency, accuracy, and completeness are essential in a submission. Sponsors should follow the specifications, including their own standards.
- The Define file is crucial and must be accurate. Too often changes made to other elements of the submission package (e.g. datasets) are not reflected in the Define file.
- Computer programs are necessary if the Define file is inadequate. Making the Define file accurate is likely to require incorporating some code in the metadata.
- Both SDTM and analysis datasets should be available to both medical and statistical reviewers. Members of the pilot project team members, as well as others outside the team had thought that medical reviewers rely on the tabulation datasets and statistical reviewers rely on the analysis datasets. The reality is that medical and statistical reviewers use both types of datasets.

Because so many of the comments made by the volunteers from FDA directly influenced the pilot submission package, a summary of the notes taken by pilot project team members during the discussion is provided in Appendix 7.6.

5.2. *Planning for the pilot submission package*

Because reviewers had recommended that a conversation be held regarding the structure and content of the pilot submission package, a "pre-submission encounter" was held. Although the pilot project team realized that, for a real-world submission, meetings between FDA and sponsors are difficult to arrange, this "encounter" provided an opportunity to have a discussion specifically related to statistical issues. In April 2006, the pilot project team met with the regulatory review team to discuss the plans for what would be included in the pilot submission package. The intended result of the meeting was for both the regulatory review team and the pilot project team to have a clear understanding of what needed to be done to make the pilot submission package sufficient for an assessment of reviewability. Timelines were also discussed.

In a real-world pre-submission encounter, the FDA reviewers would have received a briefing package beforehand. This would have eliminated the need to spend time going over the study description, as well as allowing the FDA representatives to come to the meeting with thoughts already in mind about additional analyses or different data structures. Ideally, a formal plan of the data structure, variable naming, etc., and a mock of what the data would look like, as well as a sample of the data, would have been provided. The timing of the

project did not allow the pilot project team to send a briefing package to the regulatory review team.

In addition, a real-world encounter would have involved only reviewers familiar with the particular therapy area. For the pilot project, eleven volunteers from FDA attended the meeting, representing multiple therapeutic areas and disciplines. This allowed the pilot project team to get a broader view of expectations for the pilot submission package.

The meeting began with an overview of the pilot project goals followed by an overview of the study. The analysis strategy was then presented, including what endpoints and analyses would and would not be included in the pilot submission package. The proposed data structures and descriptions of the contents of the SDTM and analysis datasets were presented. A Define.xml example and the annotated CRF were demonstrated.

In addition to numerous agreements regarding the specific pilot submission package (listed in Appendix 7.7), several key agreements were reached:

- Individual programs would not be included in the pilot submission package, though it was strongly encouraged by at least one reviewer. Instead, the pilot project team hoped to illustrate that the metadata, which would include sections of program code or pseudo-code, would be sufficient without providing complete programs.
- All levels of the MedDRA coding would be included in the SDTM datasets. This would provide a good opportunity to test the effectiveness of tools used at FDA with respect to the handling SDTM supplemental qualifiers.
- Individuals on the regulatory review team expressed a preference to avoid all listings, since they thought they would not be needed. Even listings of subjects with serious adverse events and deaths were thought to be unnecessary, since these subjects could be identified easily.

5.3. *Review team comments*

During various teleconference and face-to-face interactions between the pilot project team and the regulatory review team during August and September of 2006, comments on the original pilot submission package were collected. As stated previously, the regulatory review team had a favorable overall impression of the pilot submission package, but also provided constructive criticism in the form of many helpful comments. The revised pilot submission package (refer to Section 1.1.1 regarding availability of the package) addressed many of these comments. The key revisions made to the pilot submission package are noted in Appendix 7.8 of this report, indicating issues of interest to the regulatory review team.

The regulatory review team originally requested the inclusion of the MedDRA dictionary HLGT, HLT, and LLT in the tabulation datasets for the purposes of safety analyses. However, it was noted by some regulatory review team members that the inclusion in the tabulation data was actually unnecessary since these terms were included in the adverse event analysis dataset (ADAE).

The issues identified by the regulatory review team as being the most important involved the navigation within the Define file, the inability to print the Define file, and the structure of the primary efficacy analysis datasets.

5.3.1. Define file issues in the original pilot submission package

The following list is a summary of the issues noted by reviewers regarding navigation within the Define file:

- The browser's Back button often did not work properly
- Links to external documents did not always go to the relevant sections within those documents (e.g. links to the annotated CRF did not go to the relevant page)
- Links/navigation features that would have been useful were missing:
  - A left-side "bookmarks" pane
  - Links from the analysis dataset data definition table (DDT) to the corresponding tabulation dataset DDT
  - Links from the DDT to the dataset (XPT file)

It was also noted that the reviewer's guide was helpful to those who used it. A hyperlink from the Define file to the reviewer's guide could facilitate more access to the document.

Refer to Section 4.5 for a brief description of the resolution of these issues.

5.3.2. Analysis dataset issues in the original pilot submission package

Key points made about the analysis datasets were:

- require transparency regarding how values from the SDTM data were handled for the efficacy analysis data (e.g. how the windowing and LOCF algorithms were applied must be clear)
- prefer that the analysis dataset for the primary efficacy variable contain sufficient data to facilitate the exploration of the sensitivity of certain algorithms such as LOCF and windowing as well as alternative methodologies

To address these issues, the two primary efficacy analysis datasets (ADQSADAS and ADQSCIBC) were re-structured for the revised pilot submission package, as described in Section 2.5.1. Ideally, the NPIX analysis dataset (ADQSNPIX) and the vital signs analysis dataset (ADVS) would also have been restructured. Due to time and resource constraints, this was not done, and therefore, these two datasets are not included in the public release of the pilot submission package.

5.3.3. Response to revised pilot submission package

- The navigation of the Define file was considered much improved. The bookmark pane in the left panel was especially helpful.
- It is preferred that links to external PDF documents open the documents to the specific part being cited, which was a problem found when using older versions of the Adobe software.
- The revised structure of the two primary efficacy datasets was considered much improved. The regulatory review team believes that the revised datasets are a good illustration of what information is critical to understanding the data lineage from CRF to analysis.

6. Conclusion

The goals of the CDISC SDTM/ADaM pilot project were met. It was established that the package submitted using CDISC standards met the needs and the expectations of both medical and statistical reviewers participating on the regulatory review team. The regulatory review team noted the importance of having both data in SDTM format that support the use of FDA review systems and interactive review, and data in ADaM format to support analytic review. The project demonstrated the importance of having documentation of the data (e.g., the metadata provided in the data definition file) that provides clear, unambiguous communication of the science and statistics of the trial.

The regulatory review team expressed a favorable impression of the pilot submission package. They were optimistic about the impact that data standards will have on the work associated with their review of new drug applications.

6.1. *Lessons Learned / Summary of key points*

This section provides a list of key points deemed worthy of emphasis by the pilot project team.

- Communication between sponsor and regulatory reviewers is key to a successful submission
  - Providing a "sample" submission would verify that the Define file renders as expected and that the level of detail in the content is appropriate
  - Need to agree which analysis results are "key", thus impacting the metadata to be provided
  - Need to agree on issues regarding analysis datasets to be included in submission package, including elements needed that will allow reviewers opportunity to explore the robustness of results
- Sequence followed in pilot project for creating datasets – legacy to SDTM-without-derived to analysis datasets to SDTM-with-derived (in that analysis logic/algorithms used in the analysis datasets were then also used to populate SDTM)
  - Difficult to provide links in Define file between the derived data in SDTM and analysis datasets
  - Challenging to determine how much and how to put derived data in SDTM
  - QC processes are required to maintain consistency between corresponding variables in SDTM and analysis datasets
- CRT-DDS provided in a Define.xml file
  - Required a style sheet be developed as no standard exists – needed to ensure the Define.xml file would render correctly on various system configurations
  - Style sheet used was intended for web browser viewing, not for printing
  - Define file included analysis datasets data definition tables and analysis results metadata in addition to tabulation data definition tables
  - Analysis results metadata involved extra effort both in terms of the technical aspects of the XML and style sheet and in terms of the content (documentation and links) for the Define file

- Attributes of the pilot submission package that addressed requests or expectations of the regulatory review team
  - Navigation made easier in the Define file through use of bookmark pane and table of contents
  - Reviewer's guide provided to orient reviewers to various aspects of the pilot submission package – link provided from annotated CRF and from Define file, as well as within the PDF file
  - Improved methods for annotating were used, which helped to facilitate search
  - Links provided in Define file to PDF files (e.g. annotated CRF, SAP, study report)
- Prescriptively using the information for the metadata in building the submission package resulted in significant efficiencies; the specifications for the datasets were entered once and then used not only as metadata but also to:
  - automatically generate the define file
  - support automation of the data set creation
  - support automation of order of variables in data sets to be the same as in the define
  - maintain consistency with datasets and support automation of data set validation
- Issues to be aware of in creating a submission package
  - Define file is crucial and must be accurate and consistent with the data
  - Use a consistent method of identifying the appropriate records used in the analysis
  - How to provide links between the derived data in SDTM and analysis datasets
  - Use of the "comment" and "purpose" columns in the Define file
  - Definition of the term "derived data"
  - Design and implementation of style sheet
  - Ordering of variables in the data is important, and must be consistent with ordering in Define file
  - Verify transparency regarding how data were derived and analyzed
  - Analysis datasets should be structured in such a way that reviewers can perform sensitivity analyses as well as verify analysis results
  - Confirm hyperlinks in the Define file perform as expected

6.2. *Outstanding issues*

There are several issues that were identified in the CDISC SDTM/ADaM Pilot Project that require more detailed exploration, either by the various CDISC work groups involved or in future CDISC projects.

The outstanding issues fall into two categories. The first set of outstanding issues involves the charge to the pilot project team to identify gaps in the current CDISC models. The gaps identified included issues around derived data in SDTM, value-level metadata in the Define file, and analysis metadata in the Define file (refer to Section 4.6). The second set of issues involves around the use of the Define file and includes the identification of a requirement to provide a Define file that can be printed (refer to Section 4.5).

As stated previously, style sheets used for viewing of the Define file do not facilitate printing the file in such a way as to produce a reasonably formatted document. Solutions to allow both easy viewing and printing of Define files have not been identified. This problem could be viewed as an implementation issue that sponsors will need to handle, after discussing the issue with their FDA reviewers. For example, a sponsor might choose to provide two versions of the style sheet – XML for viewing and PDF for printing. Ideally, a reminder of the issue would be included somewhere in the CRT-DDS guidance (e.g., a note that consideration be given to how the sponsor will respond to a request from reviewers for a print-friendly version of the style sheet). It should be noted that the regulatory review team for the pilot project emphasized that the ability to print the document would be essential for the future use of XML files.

6.3. *Acknowledgements*

The CDISC SDTM/ADaM pilot project team would like to acknowledge the contributions and support of the many people and organizations that helped to successfully complete this project.

Whatever technology and solutions were needed to get the job done was shared openly between FDA and industry, software and pharmaceutical companies, or services groups and individual contributors. This openness was a key factor in the success of the project.

It is not possible to overstate the value provided by the regulatory review team's interactions with the pilot project team. The guidance, feedback, and enhanced understanding of the others' processes were invaluable to both teams.

The members of the pilot project team want to express their appreciation for the enthusiasm and continuous support from others in CDISC, including the project sponsors, the CDISC boards, and, of course, the SDTM, ODM, and ADaM teams.

The entire project would not have been possible without the support of the employers of the various team members in allowing the participants to spend time and energy working on this pilot project. This exemplifies the CDISC spirit of working together for the common and greater good.

7. Appendixes

7.1. *Appendix: Project management*

The pilot project was managed by a team of co-leaders. Once the initial pilot project team was established and work officially started, no new team members were added. There were several face-to-face meetings: an initial planning meeting in January 2006 with sponsors and the pilot project team co-leaders, a kick-off meeting in February 2006 to discuss plans and work assignments for the pilot project, a working meeting in May 2006, and another meeting in September 2006 to hear the regulatory review team's comments. Regular teleconferences were held – bi-weekly initially, going to weekly during the peak workload periods. Minutes of all meetings were posted to the team's document repository.

Initially the pilot project team was divided into three sub-teams: analysis, data, and research. The data sub-team worked on mapping the CRF and creating the structure for the tabulation datasets. The analysis sub-team worked on writing the statistical analysis plan and designing the analysis datasets (including the writing of the metadata). As these tasks were completed, these sub-teams merged to perform the programming and QC of the analysis datasets and summary tables. The research sub-team focused on the creation of the Define file, including producing an XML file that would accommodate the requirements for analysis dataset metadata and analysis results metadata.

7.1.1. Team membership

There were team co-leaders representing the CDISC ADaM team (Cathy Barrows), the CDISC SDS team (Musa Nsereko), and FDA (Lonnie Smith, Project Specialist with FDA). After the original pilot submission package was sent to the regulatory review team, Lonnie Smith had to stop his work on the pilot project due to a change in his work responsibilities. At that time, Mina Hohlen (Regulatory Information Specialist with FDA) and Chris Holland (Mathematical Statistician with FDA) took over the lead of the regulatory review team.

The pilot project team consisted of volunteers from industry, representing pharmaceutical companies (large and small), contract research organizations (CROs), and vendors. The majority of the team members were also members of other CDISC teams. The pilot project team members were (in alphabetical order):

Greg Anglin, Eli Lilly and Company
Cathy Barrows, GlaxoSmithKline
Anthony Friebel, SAS Institute
John Gorden, Quintiles
Tom Guinter, Octagon
Edward Helton, SAS Institute
Joel Hoffman, Intrasphere Technologies then Insightful
Susan Kenny, SAS Institute then Inspire Pharmaceuticals
Sandy Lei, J&J

Richard Lewis, Octagon
Arline Nakanishi, Amgen
Musa Nsereko, Cephalon
Gregory Steffens, Eli Lilly and Company
Gary Walker, Quintiles
Aileen Yam, sanofi-aventis
Yuguang Zhao, sanofi-aventis then Eisai

The following people helped address issues with the Define file for the revised pilot submission package:

Neeru Bhardwaj, sanofi-aventis
Sally Cassells, Lincoln Technologies
Robert Dainton, sanofi-aventis
Chris Decker, SAS Institute
William Friggle, sanofi-aventis
Brent Jones, sanofi-aventis

The regulatory review team included a number of employees with a variety of review experience. This group of volunteers agreed to participate and to provide feedback and review of the packages, based on their areas of expertise and interest. The views expressed by these volunteers are based on their own opinions and experience and are not, necessarily, those of FDA.

The members of the regulatory review team and the areas of their contributions were:

Primary medical review/analysis:
- Chuck Cooper, CDER medical officer
- Steven Hirschfeld, CBER medical officer

Primary statistical review/analysis:
- Tristan Massie, CDER statistical reviewer
- Feng Zhou, CDER statistical reviewer
- Chris Holland, CDER statistical reviewer

Application of tools:
- Chuck Cooper, CDER medical officer
- Joy Mele, CDER statistical reviewer
- Mat Soukup, CDER statistical reviewer

Other FDA volunteers also provided input to the pilot project through attending meetings and discussion of design, processes, and procedures. These volunteers were:

Sue Bell, CDER statistical reviewer
Howard Chazin, CDER medical officer
Janet Gentry, OIT-CDER systems analyst
Mina Hohlen, CDER OBPS regulatory review support
Zei-Pao Huang, CDER OBPS regulatory review support

Cynthia Liu, CDER statistical reviewer
Armando Oliva, CDER medical officer
George Rochester, CDER statistical reviewer
Lonnie Smith, CDER project specialist
Bobbie Witczak, OIT-CDER project manager

Project sponsors, providing support and assistance as needed, included:
Sally Cassells, Lincoln Technologies, CDISC ODM Team Leader
Dave Christiansen, Christiansen Consulting, CDISC ADaM Team Leader
Gary Gensinger, CDER OBPS regulatory review support
Dave Hardison, SAIC, CDISC Board
Edward Helton, SAS Institute, CDISC Board
Susan Kenny, Inspire Pharmaceuticals, CDISC ADaM Team Leader
Wayne Kubick, Lincoln Technologies, CDISC SDS Team Leader
Rebecca Kush, CDISC President
Randy Levin, CDER medical officer
Stephen Ruberg, Eli Lilly and Company, CDISC Board
Norman Stockbridge, CDER medical officer
Ellis Unger, CDER Deputy Division Director
Steve Wilson, CDER Deputy Division Director
Diane Wold, GlaxoSmithKline, CDISC TDM Team Leader

7.2. *Appendix: Annotating the CRF*

The annotations on the blank CRF were made using PDF comments as a "layer" of information sitting on the sample CRF. This section of the report describes how the pilot project team used Adobe Acrobat 7 Professional to create these annotations. The pilot project team elected to use Acrobat 7 to annotate the CRF, in spite of the fact that this was a divergence from the specified submission standards at the time of the pilot project. The primary reason for this decision was to take advantage of specific features of the new version of Acrobat to enhance the deliverables for the project. Additional reasons that factored into the decision were that the pilot project did not involve an official submission and that FDA employees have Acrobat 7 available to them.

All annotations were made using the PDF "Text Box" comment type. All text-box annotations where made using Helvetica font in a red color (color = #FF0000) with a teal background color (color =#00FFFF). This combination of red and teal was selected to make the comments very readable, as these are complementary colors. Each variable was annotated where data were collected from the CRF. These text boxes were re-sized so that only the variable name was visible on the page. As noted later in this section, more text may have been added to describe the variable, but the initial view of the PDF page with comments only showed the variable name. The additional details would be visible in Acrobat using the "Comments" navigation pane. This pane (in Acrobat 6 and higher) displays as a list of the

comments, shows the complete contents of the text of any text-box comments and can be expanded to show additional details such as the “Author”, “Subject”, and modified “Date” and “Time” attributes of each comment.

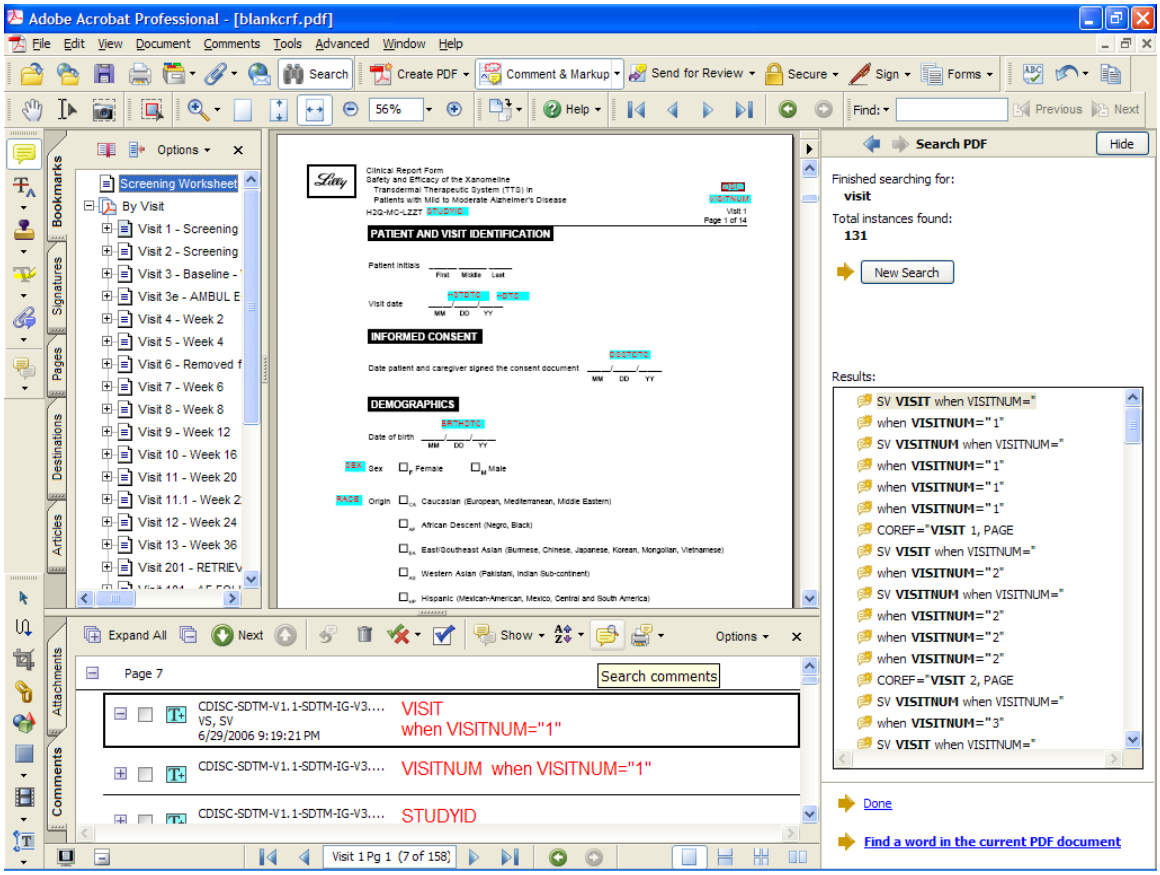


Figure 5 Illustration of comments navigation pane

The PDF comments, after being created, were modified so that all of the annotations for CDISC SDTM-defined variables (which were all variables) had the PDF comments' “Author” attribute set to *CDISC-SDTM-V1.1-SDTM-IG-V3.1.1*. This was done to differentiate these variables from non-CDISC-defined variables.

In addition to populating the Author attribute, the PDF comment “Subject” attribute was populated with the 2-letter domain prefix associated with the variable. In some cases, the annotated data corresponded to multiple variables, each in different domains. In these cases a comma-delimited list of the 2-letter domain prefixes were used to populate the “Subject” attribute field. An example of this is the visit date, which was collected on the first CRF page of each visit. All --DTC variables for data collected during that visit “inherited” this reported visit date. The annotation comment for this date collection field had its “Subject” attribute populated with “QS, VS, DS, AE, CO, CM, DM, SC, MH” because all of these domains inherited this --DTC (in Visit 1).

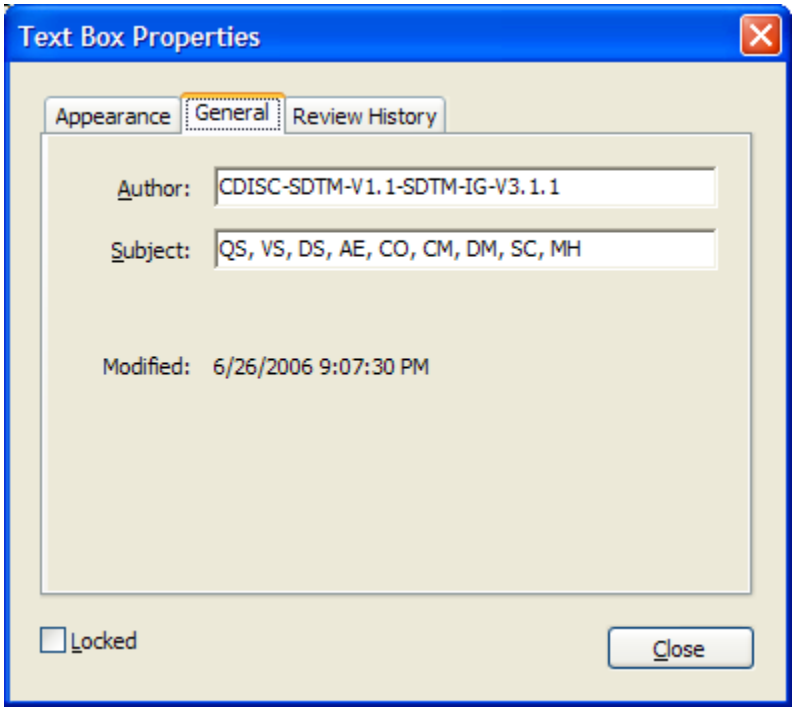


Figure 6 Illustration of annotation comment for the visit date collection field

The SDS draft Metadata Submission Guidelines recommended a process for annotating the ***text*** (in these "inherited" variable instances) by using the wildcard "--" to indicate that more than one 2-digit prefix may be applicable; this recommendation was followed in the pilot project. Another recommendation of the SDS metadata team that was followed was that a comma-separated list of applicable variables be included in square brackets, following the "wildcard" entry. An example is *--DTC [AEDTC, CODTC, CMDTC, DMDTC, SCDTC, QSDTC, VSDTC, DSDTC, MHDTC] when VISITNUM="1"*. Note that a qualifying statement was also included, indicating the instance or value of the VISITNUM associated with data collected for this variable (see next paragraph for more about this). The text box was then re-sized (as described earlier) to initially display (on the PDF page) only the text "*--DTC"* even though the brackets and additional text were part of the text in the comment/annotation.

In order to differentiate variables, many annotations had to have qualifying statements such as *VSTESTCD when VSTESTCD="PULSE"* so that this Vital Signs test code was differentiated from other test codes collected, such as blood pressure values or temperature. This became crucial in identifying different questionnaire questions and responses, especially once the comments were resized to display only the variable name such as "*VSTESTCD*". In fact, it became evident that both the topic variable (such as QSTESTCD) and the result (such as QSORRES) needed to be annotated, even though the topic variable was defined by the test or sponsor and applied to the CRF (thus the CRF was not really the source or origin of the topic variable). The result, however, was collected from the CRF and it was important to annotate the result in such a way that the result was related to the topic variable. This was achieved by using similar qualifying statements as described above, such as *VSORRES when VSTESTCD="PULSE"*. The result was this pairing of the topic variable annotation for the printed question and the result variable annotation for the CRF data "input" field.

As noted in Section 2.3, all pages where data were collected and reported were annotated. Instead of referencing other panels as an example of how a page should have been annotated, using Acrobat 7, comments were "cloned" and similar or identical pages and panels were commented from previously annotated pages and panels. This was done via a process where one annotated page or panel was created as the only page or panel in a template PDF file. The annotations/PDF comments were than exported via the Acrobat "Export Comments" feature. These were exported to an XFDF file, an Adobe XML file that describes Forms Data Fields (such as comments that are a subset of the Adobe Forms fields). The resulting XFDF files were then opened with a text editor and the page numbers were updated with the correct "target" page in the complete *blankCRF.pdf* file. The "target" page would be the first panel or page where the annotations needed to be applied. The *blankCRF.pdf* file was then opened and the XFDF file was used for importing comments to the appropriate page. This process was then repeated for the next and subsequent pages/panels until all were annotated. Fully populating the 158-page PDF file from the individual templates (approximately 13 visits cloned) took approximately 3 hours.

As described in Section 2.3, the pilot project team implemented a PDF Advanced Search that would provide a reviewer with an advanced search result including all "hits" for the searched values. The use of the Acrobat Advanced Search, using the "Search Comments" option, would let reviewers find annotations more efficiently. By combining "Search Comments" and "Whole Words", reviewers could find all variables for a particular domain using the 2-letter domain prefix which was placed in the "Subject" fields. Figure 7 depicts a screenshot of the PDF search comments option.

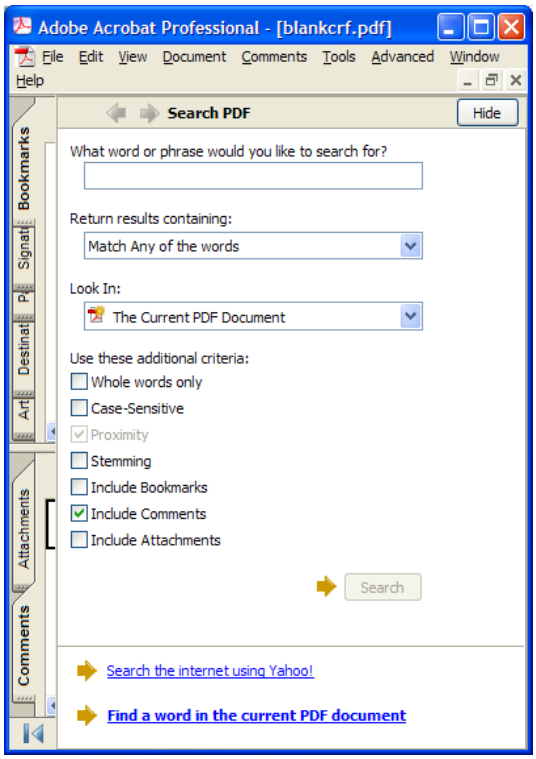


Figure 7 Illustration of including the "search comments" option

7.3. *Appendix: Analysis dataset changes*

The following illustrates the changes made to the ADAS-Cog analysis dataset (primary efficacy dataset ADQSADAS) for the revised pilot submission package. One subject has been selected for illustration.

A brief description of the data originally intended to be included in the analysis dataset is in order. ACTOT is the variable containing the total score for the questionnaire.

- Individual ADAS-Cog item scores were to be retained as additional columns on the ACTOT row.
- Two sets of data were to be used in reporting: observed cases–windowed (data as observed, but assigned to the correct analysis window based on the algorithm specified in the SAP) and LOCF (last observation carried forward to replace missing data within the observed cases–windowed dataset). Separate sets of analysis records were to be included in the analysis dataset for the data as observed and eligible for the analysis, the windowed data (i.e., observed with windowing algorithm applied), and the LOCF data (the windowed data with imputations made for missing data). These three sets of analysis records will be denoted by AVISFLG values of "Observed," "Windowed," and "LOCF," respectively.
- As described in the SAP, the last observation carried forward was to be based on the targeted assessments (i.e. those assigned to be the analyzable assessment based on the assessment windows).
- Only the rows relating to the visits being reported in the analysis summary would be retained in the analysis dataset (i.e., Baseline and Weeks 8, 16, and 24, also referred to as Visits 3, 8, 10, and 12, respectively).

Figure 8 provides three tables: a listing of the ADAS-Cog total score (ACTOT) records from the SDTM QS domain for the selected subject, a listing of the corresponding ACTOT data in the ADAS-Cog analysis dataset (ADQSADAS) as originally designed, and the metadata for the listed columns of the original version of ADQSADAS. Note that not all columns (variables) or rows (records) for the subject or dataset are included in the figure.

Illustration of changes in structure of analysis dataset ADQSADAS, specifically for the ADAS-Cog total score (ACTOT)

ACTOT Records from QS Domain

USUBJID	QSTESTCD	VISITNUM	VISITDY	QSSTRESN	QSBLFL	VISIT	QSDTC	QSDY
01-709-1259	ACTOT	3	0	15	Y	BASELINE	1/26/2013	1
01-709-1259	ACTOT	8	56	21		WEEK 8	3/23/2013	57
01-709-1259	ACTOT	10	112	19		WEEK 16	5/11/2013	106
01-709-1259	ACTOT	11	140	23		WEEK 20	6/13/2013	139

ACTOT data from ADQSADAS as originally submitted

USUBJID	AWEEK	AVISFLGN	ACTOT	ACTOTBL	ACTOTCH	ANLDY	AVISFLG	AVISIT	VISITDT
01-709-1259	0	1	15	15	0	1	Observed	BL	26-Jan-13
01-709-1259	8	1	21	15	6	57	Observed	Wk8	23-Mar-13
01-709-1259	16	1	19	15	4	106	Observed	Wk16	11-May-13
01-709-1259	0	2	15	15	0	1	Windowed	BL	26-Jan-13
01-709-1259	8	2	21	15	6	57	Windowed	Wk8W	23-Mar-13
01-709-1259	16	2	19	15	4	106	Windowed	Wk16W	11-May-13
01-709-1259	0	3	15	15	0	1	LOCF	BL	26-Jan-13
01-709-1259	8	3	21	15	6	57	LOCF	Wk8L	23-Mar-13
01-709-1259	16	3	19	15	4	106	LOCF	Wk16L	11-May-13
01-709-1259	24	3	19	15	4	106	LOCF	Wk24L	11-May-13

Metadata for ADQSADAS as originally submitted

Variable	Label	Computational Algorithm or Method
USUBJID	Unique Subject Identifier	
AWEEK	Analysis Visit Week	Week corresponding to AVISIT
AVISFLGN	Analysis Visit Type Flag, Numeric	1 when ACTOT is observed and visit is categorized as recorded on CRF; 2 when ACTOT is observed and visit is categorized as per visit windows; 3 when ACTOT is imputed using LOCF and visit is categorized as per visit windows
ACTOT	ADAS-COG(11) Subscore	Sum(ACITM01:ACITM02, ACITM04:ACITM08, ACITM11:ACITM14), see SAP section 14.1.1 for detailed scoring algorithm, adjusted for missing values
ACTOTBL	ADAS-COG(11) at Baseline	ACTOT when AVISIT='BL'
ACTOTCH	ADAS-COG(11) Change from Baseline	ACTOT - ACTOTBL
ANLDY	Analysis Day	ADQSADAS.VISITDT - ADSL.TRTSTDT + 1
AVISFLG	Analysis Visit Type Flag	decode of AVISFLGN
AVISIT	Analysis Visit	Week equivalent of QSAD.VISITNUM when AVISFLGN = 1; adqsadas.anldy categorized into windows when AVISFLGN=2 or 3
VISITDT	Visit Date	QSAD.QSDTC, converted to SAS date

Figure 8 Illustration of ADQSADAS changes - Original

The selected subject (01-709-1259) provides an illustration of the lack of transparency noted by the regulatory review team in the original pilot submission package. For this subject the ADAS-Cog total scores in the QS domain are 15, 21, 19, and 23 for Baseline and Weeks 8, 16, and 20, respectively. (The subject discontinued prior to Week 24.) The observed data and windowed data in ADQSADAS reflect only Baseline and Weeks 8 and 16. The LOCF data also include Week 24 value, carried forward from Week 16.

The regulatory review team questioned why the data for Week 20 were ignored for subject 01-709-1259 in terms of the LOCF data. The answer is that the subject's Week 20 record was outside of the window for Week 24. It actually fell into the window for Week 16 but was further away from the target day than the observed Week 16 record, so the data from Week 20 were not used. The Week 20 record did not show up in the LOCF records because the LOCF algorithm was to use the windowed visits, as specified in the SAP.

The metadata for the ACTOT variable indicates how the value was computed and refers to Section 14.1.1 in the SAP for the detailed scoring algorithm. However, there is no reference to the text specifying the windowing algorithm should be applied before the LOCF algorithm. There is also no reference to the reason for dropping the Week 20 data from the analysis dataset. This points out that though the algorithm followed was pre-specified in the SAP, determining the procedure followed was not clear in the metadata so required quite a bit of investigative work by reviewers.

In addition, the ADQSADAS dataset as originally submitted did not facilitate the testing of other strategies, such as including all data in the LOCF imputation rather than only the windowed visits.

Figure 9 shows the ADAS-Cog total score (ACTOT) data in the revised version of ADQSADAS. The dataset was revised as follows:

- There is one record per analysis parameter (i.e., outcome variable) per analysis visit per subject. In the case of the pilot submission package, the only analysis parameter in ADQSADAS is ACTOT. If there had been additional subtotals of interest for analysis, they would have been incorporated on separate rows in the dataset. The pilot project team elected not to include the individual item scores in the restructured analysis dataset.
- All ACTOT data found in the QS domain are included in the analysis dataset (i.e., no observations were "dropped").
- Because the ADaM Implementation Guide was still being developed at the time of the pilot project, the variable names and definitions do not correspond to recommendations by the ADaM team.
- In revising the dataset the pilot project team ensured that the identification of three types of data (observed, observed-windowed, and LOCF) was possible.
  - Observed data can be identified by rows where VISIT = AVISITC and ITYPE is blank
  - Data included in the analysis of observed-windowed data can be identified by ITTV = "Y" and ITYPE ≠ "LOCF"
  - Data included in the analysis of LOCF data can be identified by ITTV = "Y" (Note that ITYPE="LOCF" indicates the record added to hold the imputed value)
- The definition of "observed" data is different in the revised content because all data are included in the analysis dataset, so observed data are as found in the tabulation dataset regardless of whether they are eligible for analysis.
- Not shown in the illustration are four indicator variables included for the purposes of communicating how each record relates to the analyses. The variables are:
  - AFLNELIG (indicates whether record contains observed data eligible for analysis)
  - AFLNLOCF (indicates whether record contains data used for the LOCF analysis)
  - AFLNOBS (indicates whether record contains observed data from SDTM)
  - AFLNWIN (indicates whether record contains data included in the analysis of WINDOWED observations)

A detailed discussion of each variable added in the revision will not be included here, as the information is provided in the illustrated metadata.

Illustration of changes in structure of analysis dataset ADQSADAS, specifically for the ADAS-Cog total score (ACTOT)

ACTOT data from restructured ADQSADAS as in revised submission package

USUBJID	VISITNUM	AVISITN	AVISITC	VISIT	AVISITCD	ANLDY	QSDY	AWEEK	VISITDT	ITYPE	ITTV	VAL	BASE	CHG	ADD_REC
01-709-1259	3	3	BASELINE	BASELINE	BL	1	1	0	26-Jan-13		Y	15	15	0	
01-709-1259	8	8	WEEK 8	WEEK 8	Wk8	57	57	8	23-Mar-13		Y	21	15	6	
01-709-1259	10	10	WEEK 16	WEEK 16	Wk16	106	106	16	11-May-13		Y	19	15	4	
01-709-1259	10	12	WEEK 24	WEEK 16	Wk24	106	106	24	11-May-13	LOCF	Y	19	15	4	Y
01-709-1259	11	.	WEEK 20	WEEK 20		139	139	.	13-Jun-13		N	23	15	8	

Metadata for restructured ADQSADAS as in revised submission package

Variable	Label	Computational Algorithm or Method
USUBJID	Unique Subject Identifier	
VISITNUM	Visit Number	
AVISITN	Analysis Visit Number	If ITTV='Y', visit number corresponding to ADQSADAS.AVISITC; blank if ITTV='N'
AVISITC	Analysis Visit Description	If ITTV='Y' then AVISITC is the name of the window that ADQSADAS.ANLDY falls in; if ITTV='N' then AVISITC=QS.VISIT. Refer to Section 8.2 of the SAP for a detailed description of the windowing algorithm
VISIT	Visit Name	
AVISITCD	Analysis Visit Short Name	Short version of AVISITC if ITTV='Y'; blank if ITTV='N'
ANLDY	Analysis Day	ADQSADAS.VISITDT - ADSL.TRTSTDT + 1. ANLDY is equal to the SDTM variable QSSTDY for all study days greater than or equal to the treatment start date. The algorithm for ANLDY yields a value of 0 for the day prior to start of treatment whereas the value of QSSTDY equals -1 for the day prior to randomization. Thus ANLDY differs from QSSTDY by a value of 1 for all study days less than the treatment start date.
QSDY	Actual Study Day of finding or event	
AWEEK	Analysis Visit Week	if ITTV='Y', Week corresponding to AVISITC; blank if ITTV='N'
VISITDT	Visit Date	QS.QSDTC, converted to SAS date
ITYPE	Imputation Type	ITYPE='LOCF' if record was created to replace missing value
ITTV	Intent to Treat Visit Flag	If the observed data are eligible for analysis (i.e., QS.VISITNUM in 3,8,10,12,201) and if QS.VISIT = the name of the visit window containing ADQSADAS.ANLDY then ITTV='Y'; ITTV='N' otherwise
VAL	Numeric value of PARAM	Sum(ACITM01:ACITM02, ACITM04:ACITM08, ACITM11:ACITM14), see SAP section 14.1.1 for detailed scoring algorithm, adjusted for missing values; ACITMxx are the corresponding values of QS.QSSTRESN when QS.QSTESTCD=ACITMxx
BASE	Baseline value of VAL	VAL when AVISITCD='BL'
CHG	Change from baseline (VAL - BASE)	VAL - BASE
ADD_REC	Record created for analysis purposes?	ADD_REC='Y' if record created because of change in data due to windowing or because of missing data (LOCF) Regarding windowing - If the visit window is other than that noted in QS.VISIT, then another record is created containing a copy of the observed data. The first record has AVISITC=VISIT and ITTV='N' and ADD_REC=' . The new record has AVISITC=the name of the windowed visit and ITTV='Y' and ADD_REC='Y'.

Figure 9 Illustration of ADQSADAS changes - Revised

As stated in Section 5.3.3, the regulatory review team considered the content and structure of the revised datasets to be much improved. They believe that the new content, structure, and metadata provide a good illustration of what information is critical to understanding the data lineage from CRF to analysis.

7.4. *Appendix: Metadata creation*

Section 3 describes the creation of the metadata, as well as the use of the metadata to maintain consistency between the datasets and the Define file. This appendix provides screenshots of the Excel spreadsheets used to collect metadata components. Figure 10 illustrates the process.

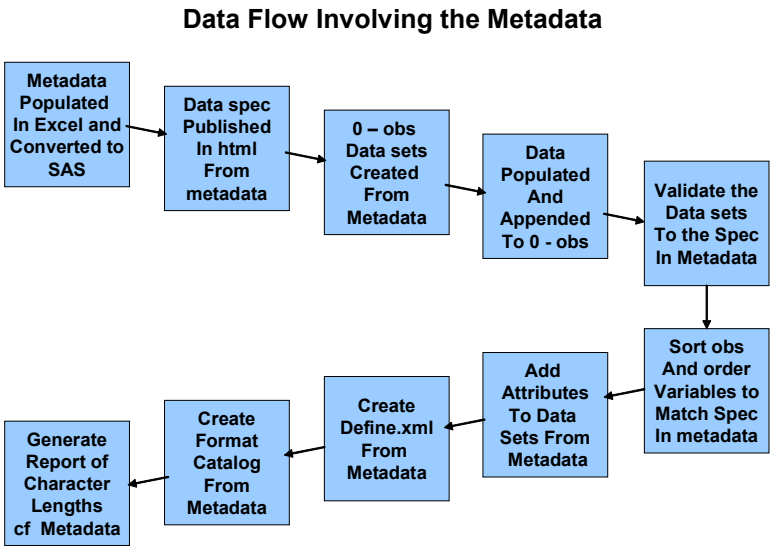
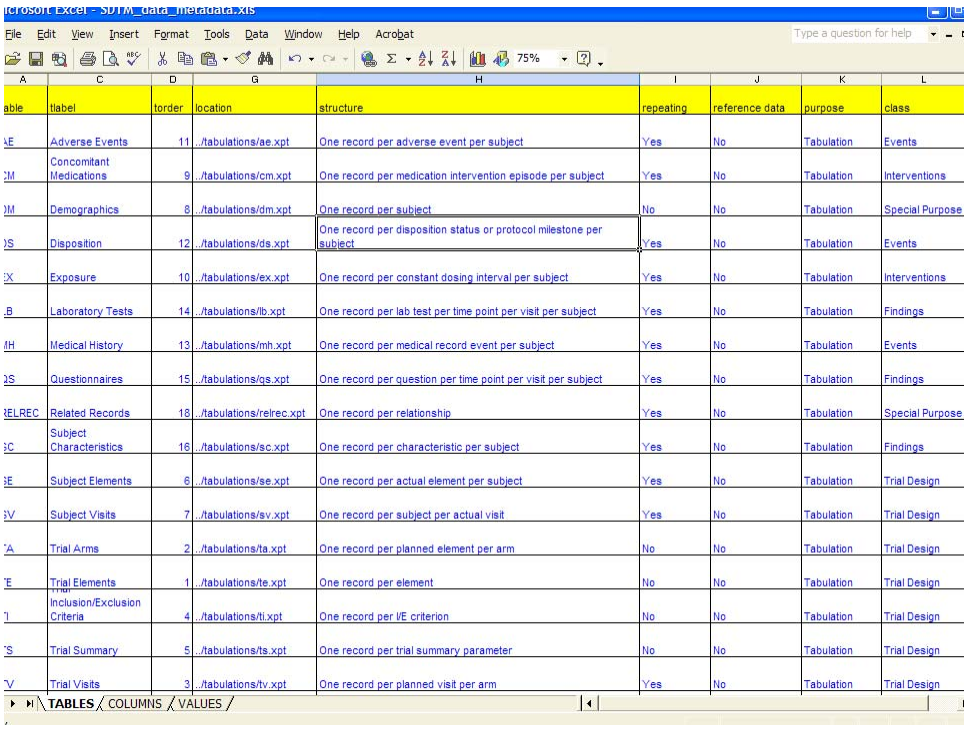


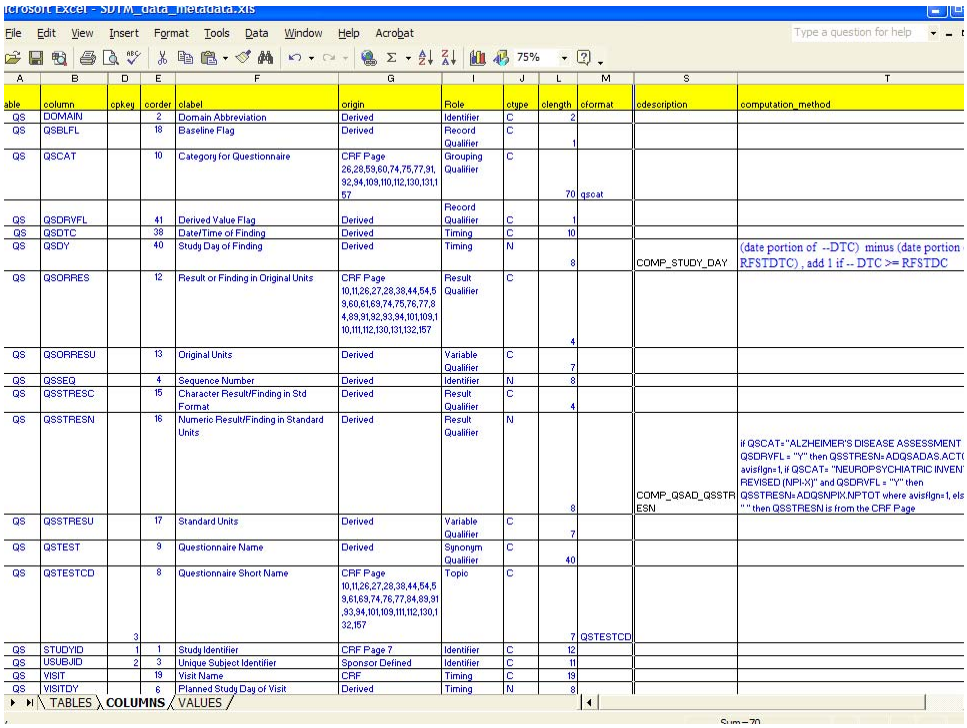
Figure 10 Illustration of steps followed in creation and use of metadata

Figure 11, Figure 12, and Figure 13 depict screenshots of selected columns of the Excel spreadsheets used to collect the metadata for the tabulation datasets. Those used for the analysis datasets were similar.



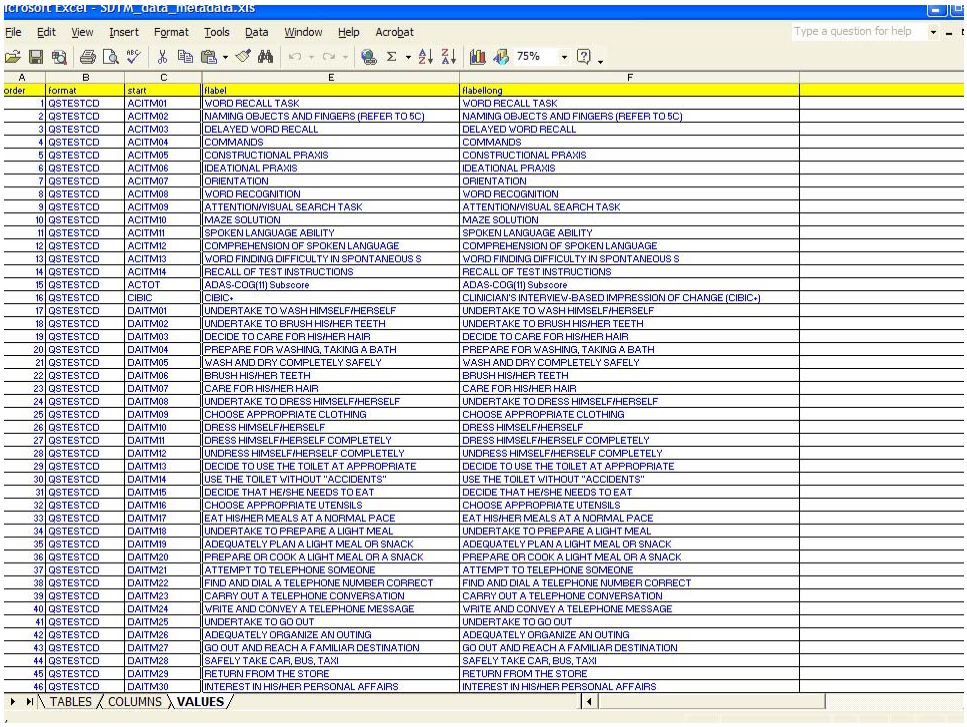
able	tlabel	torder	location	structure	repeating	reference data	purpose	class
AE	Adverse Events	11	../tabulations/ae.xpt	One record per adverse event per subject	Yes	No	Tabulation	Events
CM	Concomitant Medications	9	../tabulations/cm.xpt	One record per medication intervention episode per subject	Yes	No	Tabulation	Interventions
DM	Demographics	8	../tabulations/dm.xpt	One record per subject	No	No	Tabulation	Special Purpose
DS	Disposition	12	../tabulations/ds.xpt	One record per disposition status or protocol milestone per subject	Yes	No	Tabulation	Events
EX	Exposure	10	../tabulations/ex.xpt	One record per constant dosing interval per subject	Yes	No	Tabulation	Interventions
LB	Laboratory Tests	14	../tabulations/lb.xpt	One record per lab test per time point per visit per subject	Yes	No	Tabulation	Findings
MH	Medical History	13	../tabulations/mh.xpt	One record per medical record event per subject	Yes	No	Tabulation	Events
QS	Questionnaires	15	../tabulations/qs.xpt	One record per question per time point per visit per subject	Yes	No	Tabulation	Findings
RELREC	Related Records	18	../tabulations/relrec.xpt	One record per relationship	Yes	No	Tabulation	Special Purpose
SC	Subject Characteristics	16	../tabulations/sc.xpt	One record per characteristic per subject	Yes	No	Tabulation	Findings
SE	Subject Elements	6	../tabulations/se.xpt	One record per actual element per subject	Yes	No	Tabulation	Trial Design
SV	Subject Visits	7	../tabulations/sv.xpt	One record per subject per actual visit	Yes	No	Tabulation	Trial Design
TA	Trial Arms	2	../tabulations/ta.xpt	One record per planned element per arm	No	No	Tabulation	Trial Design
TE	Trial Elements	1	../tabulations/te.xpt	One record per element	No	No	Tabulation	Trial Design
TI	Inclusion/Exclusion Criteria	4	../tabulations/ti.xpt	One record per I/E criterion	No	No	Tabulation	Trial Design
TS	Trial Summary	5	../tabulations/ts.xpt	One record per trial summary parameter	No	No	Tabulation	Trial Design
TV	Trial Visits	3	../tabulations/tv.xpt	One record per planned visit per arm	Yes	No	Tabulation	Trial Design

Figure 11 Spreadsheet used to collect metadata regarding SDTM datasets



able	column	cpkey	corder	clabel	corigin	Role	ctype	clength	cformat	codescription	computation_method
QS	DOMAIN		2	Domain Abbreviation	Derived	Identifier	C	2			
QS	QSBLFL		18	Baseline Flag	Derived	Record Qualifier	C	1			
QS	QSCAT		10	Category for Questionnaire	CRF Page 26,28,59,60,74,75,77,91, 92,94,109,110,112,130,131,1 57	Grouping Qualifier	C	70	qscat		
QS	QSDRVFL		41	Derived Value Flag	Derived	Record Qualifier	C	1			
QS	QSDTC		38	Date/Time of Finding	Derived	Timing	C	10			
QS	QSDY		40	Study Day of Finding	Derived	Timing	N	8		COMP_STUDY_DAY	(date portion of --DTC) minus (date portion RFSTDTC) , add 1 if -- DTC >= RFSTDC
QS	QSORRES		12	Result or Finding in Original Units	CRF Page 10,11,26,27,28,38,44,54,5 9,60,61,69,74,75,76,77,8 4,89,91,92,93,94,101,109,1 10,111,112,130,131,132,157	Result Qualifier	C	4			
QS	QSORRESU		13	Original Units	Derived	Variable Qualifier	C	7			
QS	QSSEQ		4	Sequence Number	Derived	Identifier	N	8			
QS	QSSTRESC		15	Character Result/Finding in Std Format	Derived	Result Qualifier	C	4			
QS	QSSTRESN		16	Numeric Result/Finding in Standard Units	Derived	Result Qualifier	N	8		COMP_QSAD_QSSTRESN	if QSCAT="ALZHEIMER'S DISEASE ASSESSMENT QSDRVFL = "Y" then QSSTRESN= ADQSADAS.ACT avisitgn=1, if QSCAT= "NEUROPSYCHIATRIC INVEN REVISED (NPI-X)" and QSDRVFL = "Y" then QSSTRESN= ADQSNPIX.NPTOT where avisitgn=1, els "." then QSSTRESN is from the CRF Page
QS	QSSTRESU		17	Standard Units	Derived	Variable Qualifier	C	7			
QS	QSTEST		9	Questionnaire Name	Derived	Synonym Qualifier	C	40			
QS	QSTESTCD	3	8	Questionnaire Short Name	CRF Page 10,11,26,27,28,38,44,54,5 9,61,69,74,76,77,84,89,91 ,93,94,101,109,111,112,130,1 32,157	Topic	C	7	QSTESTCD		
QS	STUDYID	1	1	Study Identifier	CRF Page 7	Identifier	C	12			
QS	USUBJID	2	3	Unique Subject Identifier	Sponsor Defined	Identifier	C	11			
QS	VISIT		19	Visit Name	CRF	Timing	C	19			
QS	VISITDY		6	Planned Study Day of Visit	Derived	Timing	N	8			

Figure 12 Spreadsheet used to collect metadata regarding variables within SDTM datasets



Microsoft Excel - SDTM_data_metadata.xls

File Edit View Insert Format Tools Data Window Help Acrobat

order	format	start	label	flabellong
1	QSTESTCD	ACITM01	WORD RECALL TASK	WORD RECALL TASK
2	QSTESTCD	ACITM02	NAMING OBJECTS AND FINGERS (REFER TO 5C)	NAMING OBJECTS AND FINGERS (REFER TO 5C)
3	QSTESTCD	ACITM03	DELAYED WORD RECALL	DELAYED WORD RECALL
4	QSTESTCD	ACITM04	COMMANDS	COMMANDS
5	QSTESTCD	ACITM05	CONSTRUCTIONAL PRAXIS	CONSTRUCTIONAL PRAXIS
6	QSTESTCD	ACITM06	IDEATIONAL PRAXIS	IDEATIONAL PRAXIS
7	QSTESTCD	ACITM07	ORIENTATION	ORIENTATION
8	QSTESTCD	ACITM08	WORD RECOGNITION	WORD RECOGNITION
9	QSTESTCD	ACITM09	ATTENTION/VISUAL SEARCH TASK	ATTENTION/VISUAL SEARCH TASK
10	QSTESTCD	ACITM10	MAZE SOLUTION	MAZE SOLUTION
11	QSTESTCD	ACITM11	SPOKEN LANGUAGE ABILITY	SPOKEN LANGUAGE ABILITY
12	QSTESTCD	ACITM12	COMPREHENSION OF SPOKEN LANGUAGE	COMPREHENSION OF SPOKEN LANGUAGE
13	QSTESTCD	ACITM13	WORD FINDING DIFFICULTY IN SPONTANEOUS S	WORD FINDING DIFFICULTY IN SPONTANEOUS S
14	QSTESTCD	ACITM14	RECALL OF TEST INSTRUCTIONS	RECALL OF TEST INSTRUCTIONS
15	QSTESTCD	ACTOT	ADAS-COG(11) Subscore	ADAS-COG(11) Subscore
16	QSTESTCD	CIBIC	CIBIC+	CLINICIAN'S INTERVIEW-BASED IMPRESSION OF CHANGE (CIBIC+)
17	QSTESTCD	DAITM01	UNDERTAKE TO WASH HIMSELF/HERSELF	UNDERTAKE TO WASH HIMSELF/HERSELF
18	QSTESTCD	DAITM02	UNDERTAKE TO BRUSH HIS/HER TEETH	UNDERTAKE TO BRUSH HIS/HER TEETH
19	QSTESTCD	DAITM03	DECIDE TO CARE FOR HIS/HER HAIR	DECIDE TO CARE FOR HIS/HER HAIR
20	QSTESTCD	DAITM04	PREPARE FOR WASHING, TAKING A BATH	PREPARE FOR WASHING, TAKING A BATH
21	QSTESTCD	DAITM05	WASH AND DRY COMPLETELY SAFELY	WASH AND DRY COMPLETELY SAFELY
22	QSTESTCD	DAITM06	BRUSH HIS/HER TEETH	BRUSH HIS/HER TEETH
23	QSTESTCD	DAITM07	CARE FOR HIS/HER HAIR	CARE FOR HIS/HER HAIR
24	QSTESTCD	DAITM08	UNDERTAKE TO DRESS HIMSELF/HERSELF	UNDERTAKE TO DRESS HIMSELF/HERSELF
25	QSTESTCD	DAITM09	CHOOSE APPROPRIATE CLOTHING	CHOOSE APPROPRIATE CLOTHING
26	QSTESTCD	DAITM10	DRESS HIMSELF/HERSELF	DRESS HIMSELF/HERSELF
27	QSTESTCD	DAITM11	DRESS HIMSELF/HERSELF COMPLETELY	DRESS HIMSELF/HERSELF COMPLETELY
28	QSTESTCD	DAITM12	UNDRESS HIMSELF/HERSELF COMPLETELY	UNDRESS HIMSELF/HERSELF COMPLETELY
29	QSTESTCD	DAITM13	DECIDE TO USE THE TOILET AT APPROPRIATE	DECIDE TO USE THE TOILET AT APPROPRIATE
30	QSTESTCD	DAITM14	USE THE TOILET WITHOUT "ACCIDENTS"	USE THE TOILET WITHOUT "ACCIDENTS"
31	QSTESTCD	DAITM15	DECIDE THAT HE/SHE NEEDS TO EAT	DECIDE THAT HE/SHE NEEDS TO EAT
32	QSTESTCD	DAITM16	CHOOSE APPROPRIATE UTENSILS	CHOOSE APPROPRIATE UTENSILS
33	QSTESTCD	DAITM17	EAT HIS/HER MEALS AT A NORMAL PACE	EAT HIS/HER MEALS AT A NORMAL PACE
34	QSTESTCD	DAITM18	UNDERTAKE TO PREPARE A LIGHT MEAL	UNDERTAKE TO PREPARE A LIGHT MEAL
35	QSTESTCD	DAITM19	ADEQUATELY PLAN A LIGHT MEAL OR SNACK	ADEQUATELY PLAN A LIGHT MEAL OR SNACK
36	QSTESTCD	DAITM20	PREPARE OR COOK A LIGHT MEAL OR A SNACK	PREPARE OR COOK A LIGHT MEAL OR A SNACK
37	QSTESTCD	DAITM21	ATTEMPT TO TELEPHONE SOMEONE	ATTEMPT TO TELEPHONE SOMEONE
38	QSTESTCD	DAITM22	FIND AND DIAL A TELEPHONE NUMBER CORRECT	FIND AND DIAL A TELEPHONE NUMBER CORRECT
39	QSTESTCD	DAITM23	CARRY OUT A TELEPHONE CONVERSATION	CARRY OUT A TELEPHONE CONVERSATION
40	QSTESTCD	DAITM24	WRITE AND CONVEY A TELEPHONE MESSAGE	WRITE AND CONVEY A TELEPHONE MESSAGE
41	QSTESTCD	DAITM25	UNDERTAKE TO GO OUT	UNDERTAKE TO GO OUT
42	QSTESTCD	DAITM26	ADEQUATELY ORGANIZE AN OUTING	ADEQUATELY ORGANIZE AN OUTING
43	QSTESTCD	DAITM27	GO OUT AND REACH A FAMILIAR DESTINATION	GO OUT AND REACH A FAMILIAR DESTINATION
44	QSTESTCD	DAITM28	SAFELY TAKE CAR, BUS, TAXI	SAFELY TAKE CAR, BUS, TAXI
45	QSTESTCD	DAITM29	RETURN FROM THE STORE	RETURN FROM THE STORE
46	QSTESTCD	DAITM30	INTEREST IN HIS/HER PERSONAL AFFAIRS	INTEREST IN HIS/HER PERSONAL AFFAIRS

TABLES / COLUMNS / VALUES

Figure 13 Spreadsheet used to collect metadata regarding valid values for SDTM variables

7.5. *Appendix: the pilot project Define.xml*

In creating a Define.xml for the pilot project, the pilot project team had some lessons learned around the desired appearance and behavior of the file when rendered in a browser. The pilot project implementation of Define.xml involved extensions to the schema files that had not been vetted; therefore, specifics of the implementation will undoubtedly change.

7.5.1. Screenshots from the Define.xml

Section 4.2 described the appearance of the Define.xml as rendered in a browser by the pilot project's style sheet. This section provides some screenshots of portions of this rendering.

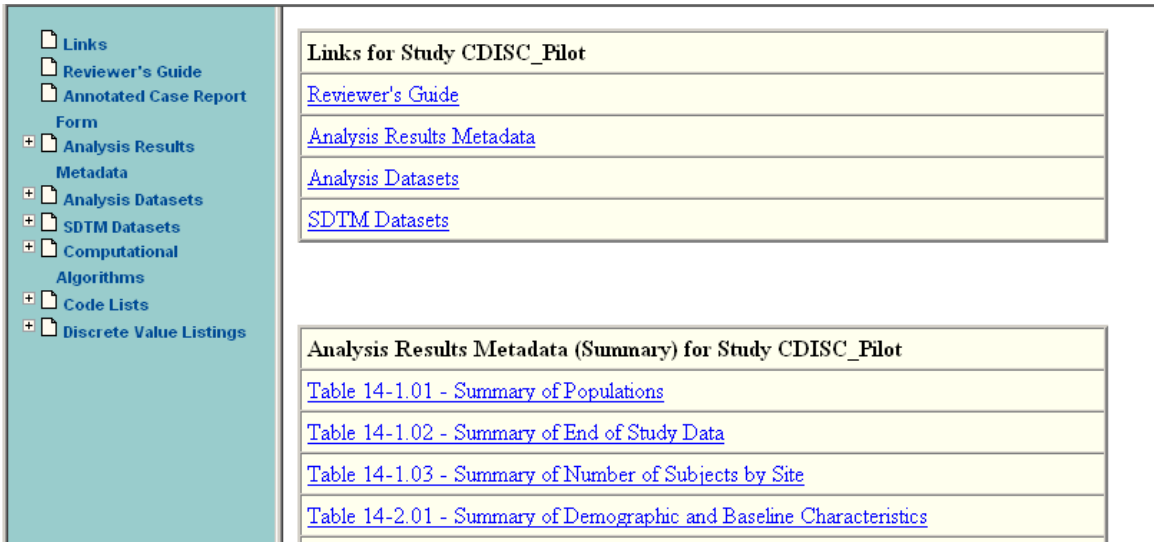


Figure 14 Example of the main-panel and left-hand pane tables of contents in the Define.xml file

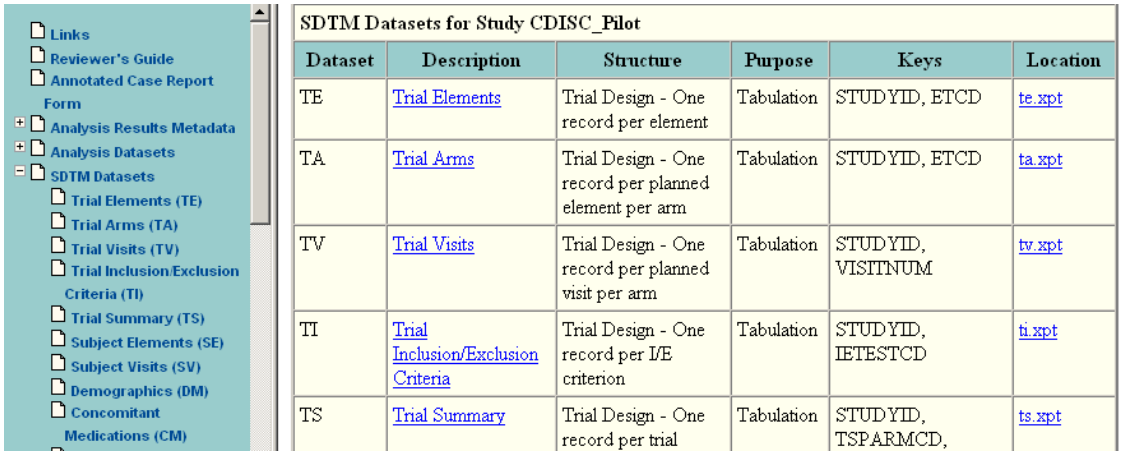
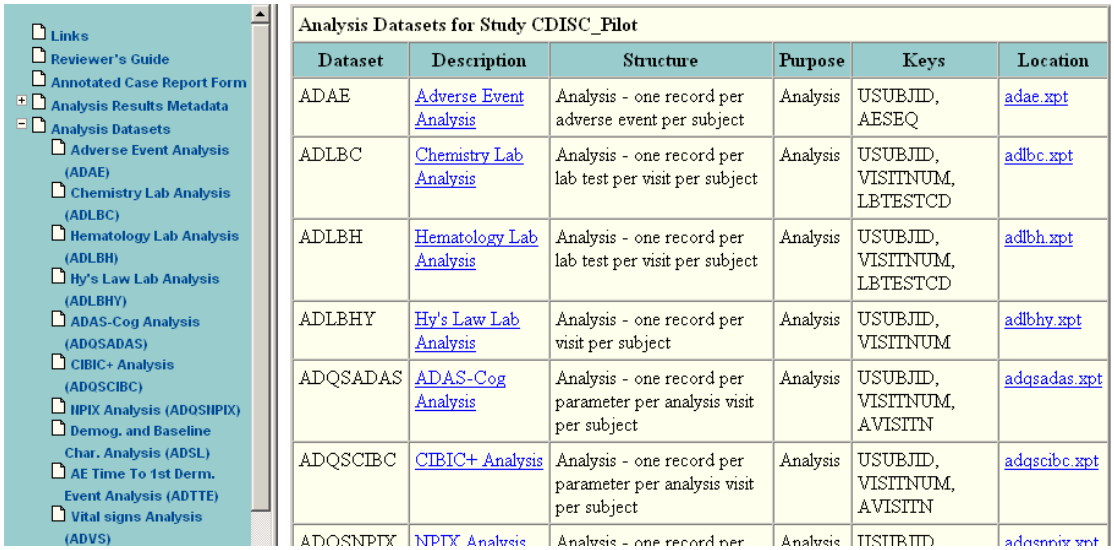


Figure 15 Example of the list of SDTM datasets in the Define.xml file



Analysis Datasets for Study CDISC_Pilot

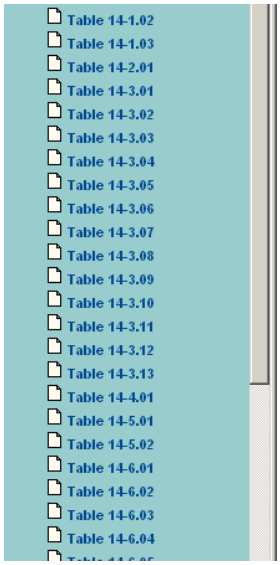
Dataset	Description	Structure	Purpose	Keys	Location
ADAE	Adverse Event Analysis	Analysis - one record per adverse event per subject	Analysis	USUBJID, AESEQ	adae.xpt
ADLBC	Chemistry Lab Analysis	Analysis - one record per lab test per visit per subject	Analysis	USUBJID, VISITNUM, LBTESTCD	adlbc.xpt
ADLBH	Hematology Lab Analysis	Analysis - one record per lab test per visit per subject	Analysis	USUBJID, VISITNUM, LBTESTCD	adlbh.xpt
ADLBHY	Hy's Law Lab Analysis	Analysis - one record per visit per subject	Analysis	USUBJID, VISITNUM	adlbhy.xpt
ADQSADAS	ADAS-Cog Analysis	Analysis - one record per parameter per analysis visit per subject	Analysis	USUBJID, VISITNUM, AVISITN	adqsadas.xpt
ADQSCIBC	CIBIC+ Analysis	Analysis - one record per parameter per analysis visit per subject	Analysis	USUBJID, VISITNUM, AVISITN	adqscibc.xpt
ADQSNPIX	NPIX Analysis	Analysis - one record per	Analysis	USUBJID	adqsnpix.xpt

Figure 16 Example of the list of analysis datasets in the Define.xml file

Analysis Results Metadata (Summary) for Study CDISC_Pilot

Table 14-1.01 - Summary of Populations
Table 14-1.02 - Summary of End of Study Data
Table 14-1.03 - Summary of Number of Subjects by Site
Table 14-2.01 - Summary of Demographic and Baseline Characteristics
Table 14-3.01 - Primary Endpoint Analysis: ADAS Cog (11) - Change from Baseline to Week 24 -- LOCF
Table 14-3.02 - Primary Endpoint Analysis: CIBIC+ - Summary at Week 24 -- LOCF
Table 14-3.03 - ADAS Cog (11) - Change from Baseline to Week 8 -- LOCF
Table 14-3.04 - CIBIC+ - Summary at Week 8 -- LOCF
Table 14-3.05 - ADAS Cog (11) - Change from Baseline to Week 16 -- LOCF
Table 14-3.06 - CIBIC+ - Summary at Week 16 -- LOCF
Table 14-3.07 - ADAS Cog (11) - Change from Baseline to Week 24 -- Completers at Wk 24 -- Observed Cases -- Windowed
Table 14-3.08 - ADAS Cog (11) - Change from Baseline to Week 24 in Male Subjects -- LOCF
Table 14-3.09 - ADAS Cog (11) - Change from Baseline to Week 24 in Female Subjects -- LOCF
Table 14-3.10 - ADAS Cog (11) - Mean and Mean Change from Baseline over Time

Figure 17 Example of the list of analysis results in the Define.xml file



Analysis	Table 14-1.01 - Summary of Populations
Description	Summary of number of subjects in each analysis population
Reason	pre-specified in SAP
Data References	Demog. and Baseline Char. Analysis (ADSL)
Documentation	SAP Section 9.1, SAP Template 1

Go to the Analysis Results Metadata Summary

Analysis	Table 14-1.02 - Summary of End of Study Data
Description	Summary of number of subjects completing/discontinuing, reasons for discontinuation, ITT population
Reason	pre-specified in SAP
Data References	Demog. and Baseline Char. Analysis (ADSL) [where ITT='Y']
Documentation	SAP Section 9.1, SAP Template 2, Reasons for discontinuation are summarized for subjects discontinuing before Week 24 (visit 12) (i.e. COMPLT24 ne 'Y' and VISNUMEN less than 12)

Figure 18 Example of the analysis results metadata for two tables in the Define.xml file

7.5.2. Placement of the Define file(s) in the pilot submission package

The directory structure for pilot submission package, following the eCTD guideline, is illustrated in Figure 19.

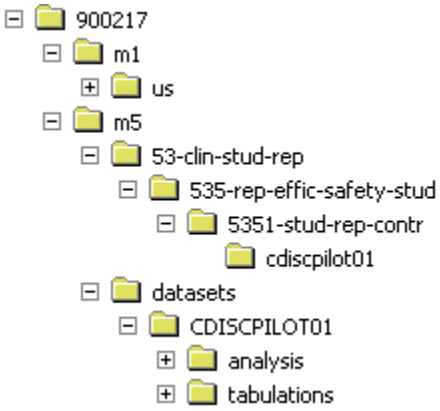


Figure 19 Directory structure for the pilot submission package

The Define file for a specific dataset should reside in the same directory as the dataset itself. As seen in the above figure, there are two distinct locations for datasets in the directory structure. The SDTM datasets reside in the tabulations directory, while the analysis datasets reside in the analysis directory; both the tabulations directory and the analysis directory reside inside the datasets directory.

Accordingly, expectation would be to find a Define file for SDTM data in the tabulations directory and a Define file for ADaM data in the analysis directory. However, the pilot project team wished to have a fully integrated Define.xml in which SDTM dataset metadata, analysis dataset metadata, and analysis results metadata were all present.

Combining these two needs above, the pilot project team decided to place an identical Define.xml in each of the tabulations and analysis locations, and arranged so that each Define file functioned the same way regardless of which one was opened. This also addressed a preference by the regulatory review team that they not have to open multiple Define files.

It turned out to be straightforward to devise a structure for the Define.xml so that not only did these two Define files *behave* the same way, but also that they were *identical files* that could reside in two different locations. How this was achieved is discussed later in this section.

7.5.3. Placement of schema and style sheet files in the pilot submission package

The supporting files for the Define.xml are located in a subdirectory UTIL present in both the tabulations and analysis folders. This keeps the supporting files in a subordinate location to each Define.xml, while keeping the tabulations and analysis directories relatively uncluttered. Within UTIL there is a folder called Foundation for schema files, and one called XSL for the style sheet. Figure 20 illustrates the location of the supporting files in the directory structure.

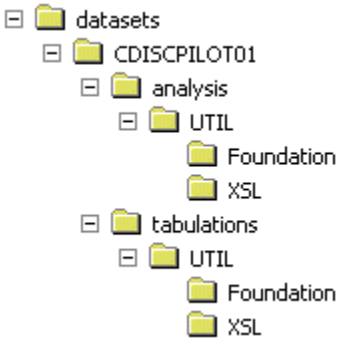


Figure 20 Directory structure showing location of supporting files for the Define.xml file

7.5.4. Schema used

At the time that the pilot project was initiated, work on ODM 1.3 was very close to completion, and it was known that this version of ODM would be fully reviewed and likely in production by the time pilot project materials were published. The pilot project team felt that an example produced in ODM 1.3 would be more consistent with ODM and Define practice moving forward than an example produced using earlier Define, and hence the pilot project submission Define.xml is based upon ODM 1.3. This had little impact on the concept of Define.xml production for the pilot submission package, though software tools used for processing the Define.xml files had to be capable of processing ODM version 1.3 files.

The schemas associated with ODM 1.3 incorporate many features directly in ODM that were part of the Define extension to earlier versions of ODM. However, there are still features of the Define specification – or, more precisely, the CRT-DDS – that are to be provided in the CDISC standards by way of an extension to ODM 1.3. The new CRT-DDS-3.1.1 – based on ODM 1.3, and replacing the previous Define-1.1.1.xml – is still being finalized, but a functional version of it was available for the pilot project work.

The structure of XML schemas that are envisioned for the CDISC Standards in the near future and of the XML schemas as used in the pilot project are illustrated in Figure 21 and Figure 22, respectively.

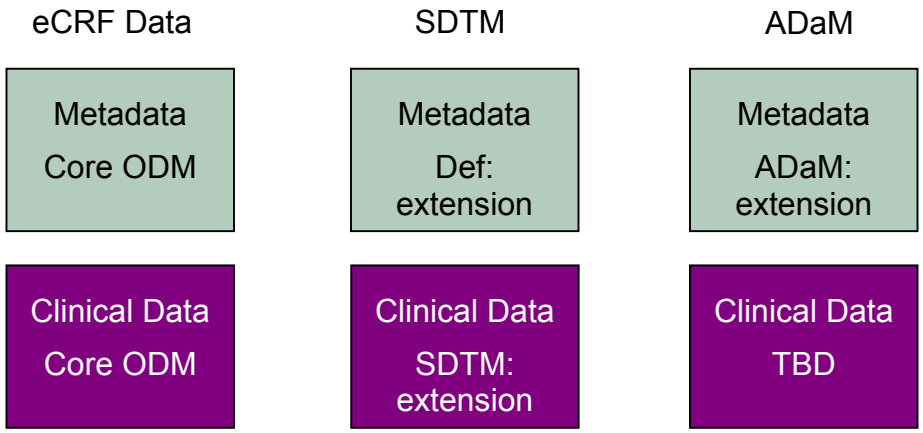


Figure 21 Illustration of the near-future structure of XML schemas

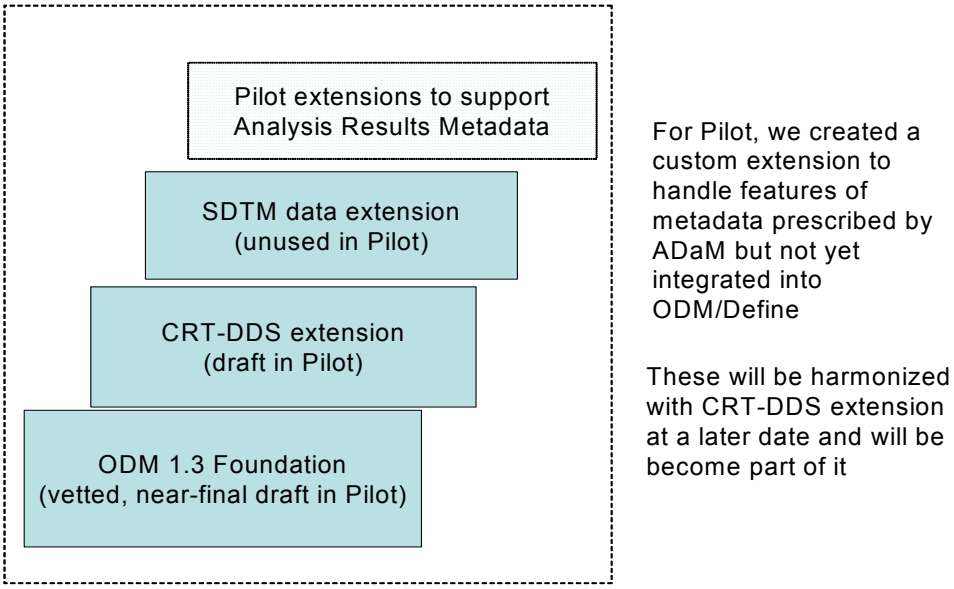


Figure 22 Illustration of structure of XML schemas used in the pilot submission package

7.5.5. Use of extension capability of ODM

The main content of the ODM extension built for the pilot project is to support the analysis results metadata specified by the ADaM team. For purposes of the pilot project, a schema was created that *functionally* supported the analysis results metadata. The fact that it was possible for the pilot project team to create an extension that met this need, within the extension capabilities of the standard and in a way that could be delivered fully functional, speaks to the underlying power of the ODM.

That said, in the longer term it is necessary that an official, vetted syntax be available that meets the need for analysis results metadata submitted as part of Define.xml. It is fully expected that, while the functionality of analysis results metadata will be supported in some future Define schema, the syntactic details will likely change. The example provided by the pilot project should be treated as one possible way to approach this need, suitable only in the short term and not as definitive of the syntax to be used in future editions of Define schema.

7.5.6. The style sheet used

An earlier style sheet was available that allowed for changes in ODM 1.3 and CRT-DDS-3.1.1; this was the basis for the style sheet used in the pilot project.

The CDISC Define team does not want to own style sheet implementation, as this has proved problematic in the past (e.g. with multiple different platforms and browsers). This likely reflects immaturity in implementations of the XML standards themselves, on those different platforms and browsers, and should become less of an issue over time. In any case, there are active discussions at CDISC regarding the *functionality* specification for style sheets, in discussion with FDA and other interested parties. This should provide guidance to style sheet implementers in the future.

As noted above, the pilot project effort did produce its own style sheet implementation. For this effort the pilot project participants, who are CDISC standards developers as well as sponsor/CRO implementers, were wearing their "sponsor/CRO implementer" hat. Following this model, sponsors would need access during submission development to their own, or hired, expertise on style sheets, in order to meet functionality expectations laid down by CDISC discussions

As the style sheet depends on extensions to the schema that have not been vetted, this style sheet code will need to evolve as the new functionality of the underlying schema is incorporated into the standard. It is unclear how much style sheet development will be formally owned by CDISC, although it is likely some "basic" version of the style sheet will be available for reference by those preparing submissions (though likely without warranty or support).

7.5.7. Creating the Define file

The pilot project team had distinct individuals working on different components of the process to produce the Define file. The SDTM metadata, the analysis dataset metadata, and the analysis results metadata were separately maintained in spreadsheets. The process by which these were combined into a single Define.xml file is outlined in Figure 23.

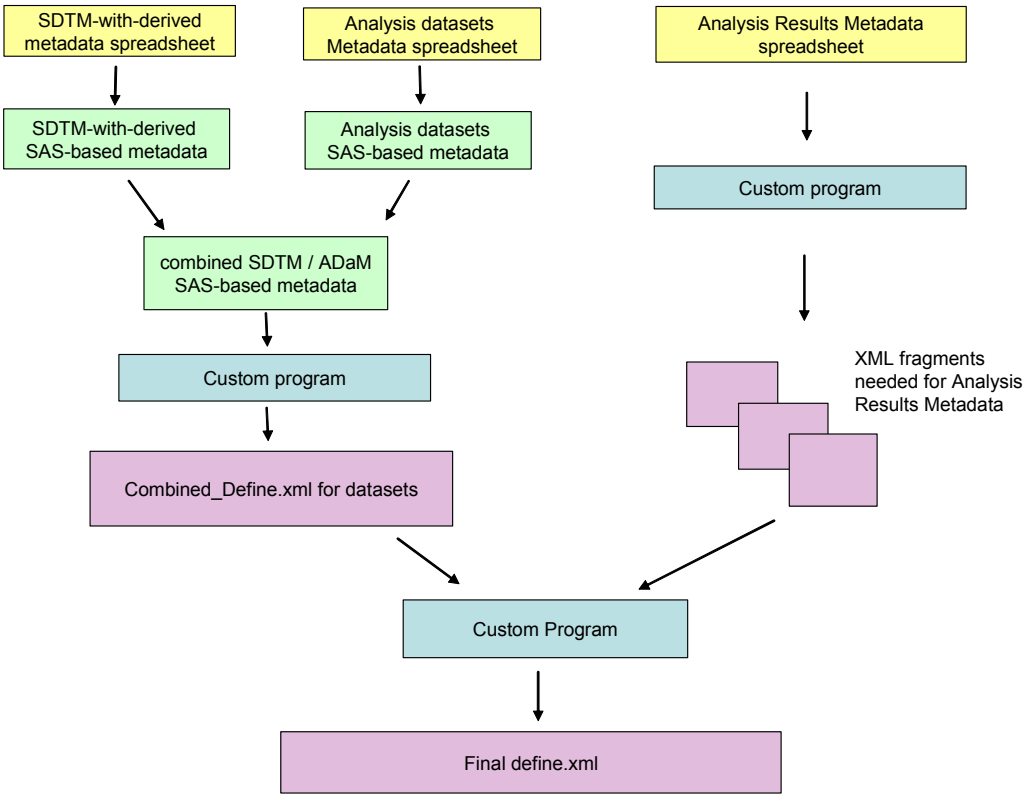


Figure 23 Illustration of the process by which metadata were combined into a single Define.xml file

7.5.8. Hyperlinking from the Define file

The CRT-DDS-3.1.1 extension to ODM 1.3 provides an element called crt:leaf that allows various elements in the Define file to make references to files external to the Define. These references are hyperlinks, and it was important and helpful to the pilot project that these references could use *relative paths*. The hyperlinks are easily rendered by the style sheet from the Define.xml to HTML.

For instance, the syntax "../tabulations" makes reference to a directory called tabulations, where tabulations is a subdirectory of the parent directory ("..") where the Define.xml resides. Importantly, this syntax refers to exactly the same file, whether the Define.xml file that was opened was the one in the analysis directory or in the tabulations directory.

The xlink:href in the following crt:leaf makes reference to the SDTM dm.xpt file by way of this mechanism. Whether the user has opened the tabulations/define.xml file or the analysis/define.xml file, the reference means the same thing.

```
<crt:leaf ID="Location.TABLE14" xlink:href="../tabulations/dm.xpt">
  <crt:title>dm.xpt</crt:title>
</crt:leaf>
```

A similar reference for the analysis dataset adae.xpt is made as follows:

```
<crt:leaf ID="Location.TABLE1" xlink:href="../analysis/adae.xpt">
```

```
    <crt:title>adae.xpt</crt:title>
</crt:leaf>
```

The syntax works generally on the hierarchy, with ".." meaning "up one directory". So, one can refer to the study report PDF file (as a whole) from either Define.xml file as

```
<crt:leaf ID="Study-Report"
xlink:href="../../../53-clin-stud-rep/535-rep-effic-safety-stud/
5351-stud-rep-contr/cdiscpilot01/cdiscpilot01.pdf">
   <crt:title>CDISC Pilot Study Report</crt:title>
</crt:leaf>
```

Extending this using PDF *named destinations* allows reference to individual sections and analysis displays in the study report:

```
<crt:leaf ID="ARM-Leaf0001" xlink:href="../../../53-clin-stud-rep/535-rep-effic-
safety-stud/
5351-stud-rep-contr/cdiscpilot01/cdiscpilot01.pdf#nameddest=OUT_TBL_14.1.01">
    <crt:title>Table 14-1.01</crt:title>
</crt:leaf>
```

7.5.9. Tools used

After the creation of the Define.xml, its integrity versus the schema was validated in three ways:

1. Using an XML authoring tool
2. Using a stand-alone Java program provided by Tony Friebel of SAS Institute
3. Using WebSDM™

The first two methods verified core syntactic correctness of the content, using the underlying schema files. The third method detected a couple of issues that were technically allowable in the XML, and so were not caught by the first two methods, but could have caused some problems if not detected. (It might be possible to build these additional checks into the schema files so that they would also be detected by the XML validation tools; this was beyond of the scope of this pilot project.)

7.5.10. Issues encountered in construction of Define.xml

Some technical issues with ODM and Define were uncovered during the pilot project will be addressed by the appropriate CDISC teams. None of these was a "showstopper" but illustrate areas where new functionality might be needed, or better understanding conveyed about available functionality.

There were some issues with "horizontal" and "vertical" representations of variables. When a variable occurs "vertically" as the --STRESN associated with a particular value for --TESTCD, it is then possible to specify a "horizontal" version of the same variable on an

analysis dataset. The fact that it is the very same variable can be made clear in the metadata because the variables would share a unique object identifier (OID). Using the Origin attribute, the horizontal structure (e.g. analysis dataset) can declare the vertical structure as the place where the variable was originally created.

In the pilot project metadata, it would have been desirable to do something similar, but in the other direction (i.e. state that some portion of a --STRESN variable was originally created on an analysis dataset).

To elaborate, the pilot project team placed some derived analysis variables for reviewer convenience on the vertical SDTM datasets, as additional rows with distinct --TESTCD. What this means is that on some SDTM datasets, the values of --STRESN for some --TESTCD are from the CRF, and for other --TESTCD are values selected and transposed from a particular column on some analysis dataset. The structures available at the time of the pilot project in ODM/Define could not readily make this relationship apparent. The best available resolution was to declare --STRESN as having Origin "Derived", that --STRESN was produced by a particular Computational Algorithm or Method, and then, in the description of that Comp Method, explain how the values are sometimes taken from CRF and sometimes based on the analysis dataset. The computational method (COMP_QSAD_QSSTRESN) for variable QSSTRESN in dataset QS is an example of this.

While the pilot project team found a way to express the nature of --STRESN in these cases, there may be some value in a more precise syntax for this operation; there is increasing alignment that at least major derived variables (e.g. questionnaire domain summary scores that serve as efficacy analysis variables) be present on the SDTM for convenience of reviewers.

7.6. *Appendix: Summary of February 2006 roundtable discussion*

The following is a summary of the notes taken by pilot project team members during the roundtable discussion with volunteers from FDA regarding expectations and requirements of FDA reviewers for a submission.

Disclaimer: All comments, statements, and opinions attributed in this document to the regulatory (FDA) review team reflect views of those individuals conveyed as informal feedback to the pilot project team, and must not be taken to represent guidance, policy, or evaluation from the Food and Drug Administration.

- The Define file was described as helpful, necessary, and crucial. Without it, reviews are much more difficult, because it provides a layout of database. Because reviewers rely on the Define file so heavily (some reviewers go straight to the Define file, assuming it will have everything they need in it), there are quite a few things to consider:
  - There tends to be a list of variables that does not always correspond to the populated variables. For example, it might include descriptions of unpopulated variables or descriptions that do not match what is in the variable. It needs to match exactly.
  - Hyperlinks are very helpful, providing links from the Define file to various documentation sources describing assumptions, derivations, etc. A good practice would be to have a short description in the Define file that is hyperlinked to a specific

documentation page providing more depth about what was done, as well as anything that would be critical to know when trying to understand the data file.
    - Providing links for tables would be very helpful. (Reviewers receive many tables, without really having a way to figure out where they came from.)
    - More focus should be placed on the Define file being consistent.
- The description of variables is sometimes not descriptive enough and there is ambiguity in the richness of the descriptions. Examples:
    - The list of controlled terms for the variable may be incomplete or may require descriptions of the terms.
    - A derived variable may contain the answer to a question (e.g. duration) but the variables in the database do not necessarily match those expected for the derivation; a clear identification of the variables was not included in the description of the derivation.
    - The description of a variable does not match the data populating the variable. For example, a variable might be described as a visit date, but the actual field is populated with the same one date for all subjects.
    - Descriptions of variables should not be ambiguous. For example, a date variable might be described as an "entry" date, without further clarification. This could be the subject's date of entry into study or it could be the date data were entered into database.
- Identifying the mapping between the protocol-specified analysis plan, the data, and the analyses performed will simplify the review process. Without this mapping, the reviewer will try to construct this mapping for the primary outcome variable and maybe a secondary variable. Many of the contacts with the sponsor are to identify how to get from the SAP to what the variables are, and then to reconstruct the analyses to see how the SAP was implemented.
- Regarding the providing of program code: If program code is requested, it is usually not to execute it, but rather to gain clarity for a variable's derivation that is missing from the define file. Because program code is not usually well documented, it takes a long time to decipher.
- Regarding data issues:
    - Two issues that present difficulties for reviewers are missing data and assumptions. Flags that would indicate if a field has missing data or if a field has assumptions applied to it would be very helpful.
    - Need to state clearly whether or not a variable could have missing values, and if so how to handle them.
    - There also needs to be a distinction made between missing data or data not collected.
    - One problem is that the variables are often named obscurely.
    - Eighty percent of the data having consistent names and attributes is essential.
    - Reviewers want to be able to manipulate data, if possible. For example, they might want to explore the impact of changing a cutoff for a lab parameter. The data should allow this flexibility (e.g., not include only the data above a specific cutoff, but instead flag the values above the cutoff and be sure to include what the cutoff is).

- Regarding tables:
  - Need to state clearly the different assumptions and rules used for a table.
  - The process is to generate a table and then draw conclusions based on the table. However, there is currently a deficiency in defining what the tables are. One person noted a desire to see blank tables submitted with the statistical analysis plan.
  - Footnotes could be added to tables to explain exceptions etc.
- Annotated CRFs:
  - Very helpful to help link the protocol and the Define files. However, reviewers tend not to look at the annotated CRF because they are often not accurate.
  - Ideally, the Define file would be set up so that the CRF is transparent, minimizing or eliminating references to the CRF. The CRF would be viewed more as a way to collect data instead of facilitating the review.
- Regarding the application used for the Define file: Define.pdf is useful, but it is inefficient to have data in one application and metadata in another. Using multiple applications is inefficient and frustrating, so they would prefer using Define.xml.
- Regarding types of data needed for the medical and safety reviewers
  - Both sets of reviewers need access to efficacy and safety analysis datasets, as well as SDTM.
  - For efficacy data, medical reviewers tend to get help from the statisticians. It would be hard to standardize, since efficacy is different for every submission and tends to have a different structure. There are predictable safety analyses asked for every time, and these can be performed and the links set up the same for every submission. A safety review guidance published in March 2005[6] gives a list of what reviewers are supposed to be looking at regarding safety data.
  - Flexibility is needed for the review of both efficacy and safety data. Though there is a set of standard analyses for every review, there are also always additional ones to be done. Safety data analyses are in general more exploratory than efficacy. However, some of the therapeutic areas (e.g. oncology) consider many exploratory efficacy analyses also, because of the different focus in life-threatening diseases. Therefore, both the efficacy and safety datasets need to be flexible in terms of being useful for additional analyses. It is also important to remember that a standard will cover things that every division needs to know and then have some 20% of the data that would need to be more flexible.
  - The safety analysis datasets are paramount.
  - Medical reviewers definitely need access to the analysis datasets as well as SDTM. They also need access to the analysis plan.
  - Statistical reviewers need access to both SDTM and analysis datasets, as they work to recreate the analysis datasets from the raw data.

[6] February 2005 FDA Reviewer Guidance: "Conducting a Clinical Safety Review of a New Product Application and Preparing a Report on the Review," a Good Review Practice by CDER. Refer to the following website: http://www.fda.gov/cder/guidance/3580fnl.pdf

- Regarding analysis datasets: The value of the analysis dataset is to tie the conclusions back to the raw data. In the past, analysis datasets have been supplied to some review divisions, but are generally lacking in adequate documentation. If the analysis datasets are not provided or are inadequate, the reviewer has to go back to the raw data and reconstruct the analysis. Generally, the reviewer can come to the same conclusion as the sponsor, but not the same numbers. The value of the analysis dataset is that they will be able to get the same numbers, as well as allowing the reviewer to see if the analysis datasets follow the analysis plan.
- The data specification for SDTM provides a well-defined structure for data. The most critical thing is to follow the naming structure defined in the specifications. If the data are in consistent locations, reviewers know where to look. In addition, automation is going to be very important in the future, so having a well-defined data structure will be crucial. Consistency and following specifications is very important.
- The FDA volunteers were asked what topics they would want covered in a conversation with an industry statistician to discuss plans for a submission.
    - These conversations should probably occur at least at the end of phase II meeting; the pre-NDA meeting is too late. Sponsors can request a pre-submission encounter meeting specifically to address data issues. Ideally, a formal plan of the data structure, variable naming, etc, and a mock of what the data will look like will be presented and all of the elements mentioned need to be specifically discussed.
    - A phase II study could actually be used as the "mock up", so have something substantive to look at.
    - Reviewers would also appreciate having a sponsor's statistician provide an orientation when the data are submitted. The focus would NOT be on showing what was done and why; instead it would be on how to find things in the data.
    - A reviewer's guide for the submission package would be very helpful. It could be put under the cover letter heading part of the eCTD. The text of the actual cover letter would say that a reviewer's guide has been included under the cover letter heading. Note that the reviewer's guide would cover more than the data. A benefit of the reviewer's guide would be that if the reviewer changes there would be a document available for orienting the new reviewer.
    - The reviewer's guide could also be linked to from multiple places in the submission, such as in the define file, because not everyone reads the cover letter.
- A very big issue is the lack of consistency within a submission.

7.7. *Appendix: Summary of April 2006 discussion with regulatory review team regarding specific content within the pilot submission package*

Disclaimer: All comments, statements, and opinions attributed in this document to the regulatory (FDA) review team reflect views of those individuals conveyed as informal feedback to the pilot project team, and must not be taken to represent guidance, policy, or evaluation from the Food and Drug Administration.

- Discussion was held as to whether or not programs would be included. One reviewer strongly encouraged the provision of programs as part of the documentation. ADaM v2 allows for this possibility, in cases where it would help describe what had been done. It was commented that the pilot project team hopes the metadata will be sufficient without providing programs, but understands programs or pseudo-code may be needed.
- It was requested that all levels of the MedDRA coding be included in the SDTM datasets. Currently, the only way to do this is to include the additional levels in SUPPQUAL. (AE domain only includes body system and preferred term.) This is a good opportunity to see how SDTM handles supplemental qualifiers and to see if this method works effectively. If there are secondary mappings, those should also be included.
- It was agreed that no patient listings would be provided. This includes listings such as SAEs and deaths. If the data are such that these subjects can easily be identified, there is no need for patient listings to be produced. There was a strong preference that patient listings be avoided.
- Regarding the analysis of lab data and Hy's Law, it was agreed that a summary be presented using modified Hy's Law. In addition, a table, or at least a flag in the data, would be included to indicate the full Hy's Law criteria.
- In addition to the above agreements, the following agreements were reached:
  - An additional efficacy analysis was requested.
  - Definitions for specific terms (e.g. dermatological event, treatment-emergent adverse event) were agreed.
  - The MedDRA Preferred Term will be used in summaries rather than the Lower Level Term.
  - P-values would be included in lab analyses.
  - It was requested that the pilot submission package include normal shift tables for labs.
  - Both the normal and the reference ranges would be included in the data. SI units will be used, but original units will also be provided.
  - All population flags will be included in the analysis datasets.
  - Dermatological adverse events will be flagged in the analysis dataset, as will the subject's first occurrence of a dermatological AE.
  - Both date of last dose & date patient was last observed would be included in all analysis datasets.

7.8. *Appendix: Key revisions to the pilot submission package*

Key revisions made to the pilot submission package are noted here, to indicate issues of interest to the regulatory review team.

Disclaimer: All comments, statements, and opinions attributed in this document to the regulatory (FDA) review team reflect views of those individuals conveyed as informal feedback to the pilot project team, and must not be taken to represent guidance, policy, or evaluation from the Food and Drug Administration.

Modifications to the DEFINE file

A new style sheet was used for the revised DEFINE file in order to include additional features and easier navigation. Key new features included are:

- A left-side "bookmarks" pane
- Links from each variable in each analysis dataset DDT to the dataset DDT(s) from which the variable was created
- Links from each individual DDT to the dataset (XPT file)
- Links from the Origin column in each SDTM data definition table to the specific page in the annotated CRF document
- Link to the reviewer's guide in the table of contents at the top of the file

Modifications to the annotated CRF

The changes made to the annotated CRF were:

- Added annotations to indicate what fields collected on the CRF were not provided in the SDTM data
- Removed annotations for fields that were not provided in the SDTM data
- Included a link to the reviewer's guide at the top of the document

Modifications to the primary efficacy analysis datasets

Comments from the regulatory review team regarding the first pilot submission package clearly pointed out the lack of transparency about how values from the SDTM data were handled for the efficacy analysis data (ADQSADAS and ADQSCIBC). Changes to these two datasets included:

- Used a structure of one record per parameter per analysis visit per subject
- Included all relevant observations from the QS tabulation datasets, instead of only observations pertaining to the subjects and visits being analyzed
- Added flag variables to convey information regarding records selected for the LOCF and windowing algorithms and for analysis
- Ensured ability to identify "observed" records (i.e., included as is from the tabulation dataset) as distinct from "derived" records (i.e., records derived as a result of an analysis algorithm)

Modifications made to the analysis datasets and relevant metadata

In addition to the changes to ADQSADAS and ADQSCIBC described previously, the following changes were made to the analysis datasets and metadata.

- Modified the description of "structure" in the list of analysis datasets to be consistent with the ordering used in SDTM
- Population flag variables modified to contain no blank values
- Include dosing start and stop dates in all appropriate analysis datasets
- Added a flag variable to indicate whether the observation occurs while the subject is on-treatment or off-treatment

- Use of a more logical ordering of in the analysis datasets (rather than alphabetically)

Modifications to the tabulation datasets and relevant metadata

The changes made to the SDTM (tabulation) datasets and metadata include:

- Included the dictionary names and versions for the AE and CM coded fields in the TS dataset
- Provided one QS domain rather than splitting it by questionnaire
- Included variable SESEQ in the SE dataset

Other modifications to the pilot submission package, as requested by the regulatory review team

- Patient narratives provided in the CSR and in a separate ASCII text file
- Raw statistical output from the primary efficacy analyses and from the repeated measures analysis provided as a subsection to Appendix 9 of the CSR (Documentation of Statistical Methods)

7.9. *Appendix: List of abbreviations and acronyms*

The following is a list of abbreviations and acronyms used in this pilot project report. Not included here are explanations of the various SDTM domains (e.g., QS, DM). Also not included is a description of the variables referenced.

Term	Definition
0-obs	Zero observation dataset
aCRF	Annotated case report form
ADaM	Analysis Data Model
ADaM v2	Analysis Data Model Version 2.0
ADAS-Cog	Alzheimer's Disease Assessment Scale - Cognitive Subscale
ADSL	Subject level analysis dataset
AE	Adverse Event
CBER	Center for Biologics Evaluation and Research
CDER	Center For Drug Evaluation And Research
CDISC	Clinical Data Interchange Standards Consortium
CRF	Case Report Form
CRO	Contract Research Organization
CRT-DDS	Case Report Tabulation Data Definition Specification, also known as the Define.xml
CSR	Clinical Study Report
DDT	Data Definition Table
eCTD	Electronic Common Technical Document
eNDA	Electronic New Drug Application
ETL	Extract, Transform, and Load
FDA	Food and Drug Administration
HLGT	Higher Level Group Term
HLT	Higher Level Term

Term	Definition
HTML	Hypertext Markup Language
ICH	International Conference on Harmonisation
LLT	Lower Level Term
LOCF	Last Observation Carried Forward
MedDRA	Medical Dictionary for Regulatory Activities
MSSO	Maintenance and Support Services Organization
OBPS	Office of Business Process Support
ODM	Operational Data Model
OIT	Office of Information Technology
PDF	Portable Document Format
QC	Quality Control
SAP	Statistical Analysis Plan
SDS	Submission Data Standards
SDTM	Standard Data Tabulation Model
TDM	Trial Design Model
TOC	Table of Contents
WHODD	WHO Drug Dictionary
XFDF	XML Forms Data Format
XML	Extensible Markup Language
XPT	extension for a SAS Transport File