CDISC, ADaM, and TTE: What Are These Acronyms and How Can They Help Me with My Filing?

Sandra Minjoe, Genentech, Inc., South San Francisco, California
Judy Phelps, Pfizer, Inc., New London, Connecticut

ABSTRACT

If you're planning a filing and know you'll need to submit analysis data to the FDA, but you don't know where to begin, then you've come to the right place. This paper presents a set of Time-to-Event (TTE) analysis datasets developed for the Analysis Dataset Modeling Team (ADaM) working group of the Clinical Data Interchange Consortium (CDISC). It shows how datasets of this type can be developed by your site and then used by both you and FDA reviewers in running your analyses. It also provides references to other ADaM models and working groups of CDISC that can be useful in planning for your submission.

KEYWORDS
CDISC, ADaM, time-to-event, filing, submission, FDA, model, dataset, analysis

INTRODUCTION

This paper will first introduce CDISC and give a little history of the organization. It continues by describing ADaM's purpose and how it fits in with CDISC. Included are some ADaM accomplishments and future projects. The specific ADaM model for Time-to-Event analysis is then described, and each dataset is outlined with examples.

We hope that by educating our fellow statistical programmers about CDISC, we can encourage you to use and maybe even get involved in the development of industry standards.

THE CDISC APPROACH

CDISC is a volunteer, non-profit organization open to everyone. Work is performed in teams who meet regularly in telecons. ADaM is one working group of CDISC. The other working groups are the Operational Data Model Team (ODM), the Laboratory Data Team (LAB), and the Submission Data Standards Team (SDS).

As stated on the CDISC website (http://www.cdisc.org/), CDISC is an open, multidisciplinary, non-profit organization committed to the development of industry standards to support the electronic acquisition, exchange, submission and archiving of clinical trials data and metadata for medical and pharmaceutical product development. The mission of CDISC is to lead the development of global, vendor-neutral, platform independent standards to improve data quality and accelerate product development in our industry.

CDISC was established to address the problems within clinical trial research. FDA-regulated products account for about 25 cents on every consumer dollar spent in the United States, yet each pharmaceutical company establishes their own clinical trials content standards independently of other companies in the industry. This presented a challenge for reviewers trying reconcile data from multiple sources and incompatible systems.

BACKGROUND AND OVERVIEW

The challenges leading up to the adoption of standards are numerous. Firstly, requirements in our industry are not clearly defined and articulated. Secondly, many organizations have focused on their own internal standards and not industry-wide standards and there is sometimes a reluctance to share those internal standards. Thirdly is that industry-wide standards may require a change in process for the organization.

Before CDISC, the industry and the FDA were faced with multiple organizations with their own shifting standards. Adding to that was the pressure on the FDA scientists to make the 'right' decision in a timely fashion. FDA was also under pressure to provide some solutions to this problem. The FDA SMART initiative took initial steps to help develop electronic review tools in response to multiple CANDA/CAPLAs submitted from various companies. In 1999, the FDA published its "Guidance for Industry on Electronic Submissions – General Considerations" which called for electronic submission of CRFs/CRTs.

This lead to a perfect setting for the development and acceptance of clinical trial content standards, and for industry acceptance and participation in standards development. Time was ripe as well for regulatory participation in the standards development process. CDISC was started in 1997 in response to the need to establish standards.

ADAM

CDISC has drafted an Operational Database model (dealing with the structure and content of clinical trail data from the perspective of how it was collected) and a Submission Metadata Model (SMM) that addresses the structure and content of clinical data from the perspective of how the data are used. There are two components of SMM: Submission Data Standards (SDS) which deals with CRT/Domain datasets, and Analysis Data Models (ADaM) that deals with datasets used for statistical review and analysis.

The main objective of both SDS and ADaM is to provide regulatory submission reviewers with clear descriptions of the usage, structure, contents and attributes of all submitted datasets and variables. This will allow reviewers to replicate most analyses, tables, graphs and listings with minimal or no transformations ("one proc away"), and to enable them to easily view and subset the data used to generate any analysis, table, graph or listing without complex programming ("more think time").

The ADaM objectives, as outlined on their website (http://www.cdisc.org/about/about_adam.html), are to:

- Provide metadata models and examples for analysis datasets used to generate the statistical results for a regulatory submission.
- Build on metadata models developed for safety domains, adding attributes and examples specific to statistical analysis.
- Acknowledge that clinical trials are unique and that the design of analysis datasets is driven by the scientific and medical objectives of the study.

Analysis datasets, although not required in the Guidelines, are strongly suggested by CDISC. These are generated from the data values in the CRT (Domain) datasets. Analysis datasets are used primarily by statistical reviewers who need the data to replicate analyses, test assumptions and perform alternative analyses. Using the CDISC submission domain standards, an electronic submission from various companies will have the same look and feel.

ADAM DATA MODELS

The ADaM dataset models are built upon the metadata of the CRT models developed for safety domains, adding attributes and examples specific to statistical analyses as you will see later in the TTE model. Sample statistical results are used as a guide for developing the models and initially we just focused on primary and secondary efficacy endpoints. The ADaM approach also acknowledges that clinical trials are unique and that the designs of analysis datasets are driven by the scientific and medical objectives of the study.

Most trials are powered to test significance for the primary endpoint. Usually safety analysis is mainly descriptive with little formal hypothesis testing, although this may change. The ADaM dataset models depend on the statistical methodology more than the actual endpoint. Describing all statistical models with all endpoints from every indication would be impossible. We must understand the principles and concepts to apply the models to different indications and trial designs. ***The CDISC dataset models are unlikely to contain all the dataset and variable metadata needed to define substantive clinical trial datasets*** (CRT or ADaM). As Dave Christiansen (current chair of the ADaM team) likes to point out " this is science, not a bank transaction". We need to standardize what we can and then clearly define the rest. To do this requires a good understanding of the principles and concepts behind the metadata models.

ADaM emphasizes the importance of early and frequent communication with FDA reviewers. ***There are no efficacy domains or analysis datasets listed in the current guidelines***. Statistical and medical reviewers have tremendous leeway in this area. Submission of programs with datasets is another area where the guidelines refer to the reviewing division for guidance. The CDISC metadata can be used as a language to communicate more effectively with the FDA reviewers.

Progress to date of the ADaM team includes publishing "Guidelines for Analysis Files" and "Criteria for Data Structures" documents. We have also completed draft data models for "Change from Baseline" and "Time to Event (Survival Analysis)"; the latter is presented here in this paper. We have begun to address a categorical analysis model and we are collaborating with the FDA on guidelines for study design, data analysis and regulatory applications for exposure-response relationships.

NEXT STEPS FOR ADAM

There are some sticky issues that ADaM is still working on. One is the issue of how to identify analysis populations in the datasets. There are a number of options such as using status flags, separate datasets, separate variables, plus others. The choice depends upon the statistical analysis, the study design, the structure of the dataset plus the "ease-of-use' vs. 'ease-to-create' dilemma. ADaM is also currently developing approaches for documenting imputation methods such as partial dates and 'last observation carried forward'.

Another issue that ADaM will address in the future is the submission of SAS® programs. There are many things to consider under this topic, such as which programs to submit and how will the programs be used (will they be executed in the FDA environment?). There is also the issue that standardized report programs are usually pretty complicated and macros are difficult to transport and understand. This dialogue must be initiated with the FDA statisticians.

Numerous technologies are emerging to leverage these new standards. There are tools being developed in XML with the advantage that XML can handle both documents and data (so the entire submission could be done in XML!). Progress is being made on subject profile software, based on the SDS metadata and also on SAS-based standard reports and algorithms, by companies such as Trialex, Inc. As these tools are developed and moved into production use, they will help to facilitate review and development of new treatments.

TTE

As mentioned earlier, ADaM has developed a draft TTE model. Sandra Minjoe, co-author of this paper, worked with Alex Bajamonde, a Genentech biostatistician and long-term ADaM member, to develop the TTE model datasets for ADaM. The goal was to create a set of datasets general enough that they could be created by any company, regardless of their in-house data structures, yet specific enough that the FDA (CDER, CBER, and CDRH) could easily find the information they would be looking for when analyzing time-to-event data.

The final analysis datasets in the model were designed to be "one proc away" from an analysis table or graph, and could plug directly into a statistical procedure, such as SAS® PROC LIFETEST. Note that FDA reviewers are not statistical programmers, and may use SAS® Analyst application, JMP®, S-Plus®, or other statistical analysis products to independently perform statistical tests. Thus datasets must be generic enough to work in many different packages.

DATASETS IN THE TTE MODEL

A total of 4 datasets were developed for the TTE model, one source dataset (which was called SOURCE1) and 3 analysis datasets (which were called ANAL1, ANAL2, and ANAL3).

- SOURCE1 contains one record per result per subject, and provides a starting point on which to build. It was developed to represent the incoming "raw" data.
- ANAL1 is an intermediate analysis file that may be useful to the reviewers, but is not used directly by any statistical procedures. It is instead used to create ANAL2 and ANAL3.
- ANAL2 and ANAL3 are both analysis files that are "one proc away" from generating statistical analysis output. They each contain the same data but in different structures: ANAL2 has one record per analysis variable within subject, and ANAL3 has one record per subject with sets of columns for each analysis variable. The two different layouts will allow FDA reviewers to use the data structure most comfortable for them, and a simple transposition (e.g., PROC TRANSPOSE) is all that is required to switch between them.

Each of the datasets in this TTE model contain the variables from the CDISC Metadata Model:

- DRUGID (sponsor-defined variable name)
- STUDYID
- SITEID (Center or Site ID)
- INVID (Investigator ID)
- USUBJID (Unique Subject ID)
- SUBJID (Subject ID that may be non-unique)
- S_PP (sponsor-defined variable name for subject per-protocol flag, such as evaluable for efficacy)
- AGE (in years at baseline)
- SEX (character version)
- SEXCD (numeric coded version)
- RACE (character or numeric)
- TRTCD (numeric treatment code)
- TRTGRP (character treatment group)
- DMREFDT (sponsor-defined data used as reference, such as screening or randomization date)

Specifics for SOURCE1:

The SOURCE1 dataset represents the incoming "raw" data. It contains multiple records for a subject, specifically one record per subject event. In addition to the CDISC metadata variables, it also contains the following:

- Visit information (visit name and number)
- Derived study day (based on visit date and DMREFDT)
- Response (a numeric grade, such as
  - 1 = complete response
  - 2 = partial response
  - 3 = stable disease
  - 4 = progressive disease
  - 5 = unable to evaluate response)

- Censoring information (one censor flag, 0=not censored and 1=censored, for each event, such as progression.)

An example of some of the variables in a few records of one subject in a SOURCE1 dataset might look like:

VISIT	RSPACTDY	RSP	V_CNRTTP
Visit 1	51	2	1
Visit 2	117	2	1
Visit 3	213	4	0

In this example, Subject 1001 has 3 events, partial responses (RSP=2) at day 51and day 117 (RSPACTDY), and progression (RSP=4) at day 213. The partial response records are flagged as censored for time to progression (V_CNRTTP=1), but the progression record is not censored (V_CNRTTP=0).

Specifics for ANAL1:

The ANAL1 dataset contains information about the events. Recall that this is not a "one proc away" analysis file. Instead, this intermediate file contains all the specific information about each relevant event for the subject, and how this data is used to determine the analysis event(s).

The actual events used for analysis might be a compound of several of the events. For example, time to disease progression might be determined by whichever occurs first: death, onset of other treatment, or discontinuation due to toxicity. All of these constituent events are contained in this intermediate file. This file includes the following variables, in addition to those mentioned previously as common to all datasets:

- Information for each event, such as disease progression, death, or study discontinuation. Variable names are sponsor-defined, but might look something like:
  - EV1ACTDT (actual date of event 1)
  - EV1ACTDY (relative study day of event 1)
  - EV1CNSR (censoring indicator for event 1)

  … and so on, for all constituent events…
- Information for each analysis event (which may contain combinations of the above constituent events) are also sponsor-defined, but might look something like:

  Time to Disease Progression variables
  - CNRDTTTP (censoring date)
  - CNRDYTTP (relative study day)
  - CNRRTTP (reason for censoring)
  - TTPTRIG (triggering constituent event)

  Time to Treatment Failure variables
  - CNRDTTTF (censoring date)
  - CNRDYTTF (relative study day)
  - CNRRTTF (reason for censoring)
  - TTFTRIG (triggering constituent event)

  Survival variables
  - CNRDTSRV (censoring date)
  - CNRDYSRV (relative study day)
  - CNRRSRV (reason for censoring)
  - SRVTRIG (triggering constituent event)

The following example shows how, for 4 different subjects, constituent events 1 (Disease Progression) and 2 (Death) are used in the ANAL1 dataset to determine the trigger for the compound event TTP (Time to Progression):

EV1ACTDY	EV1CNSR	EV2ACTDY	EV2CNSR	TTPTRIG
213	0	.	1	1
.	1	.	1	.
.	1	59	0	2
1	0	43	0	109

The first subject progressed at day 213 and did not die, so the TTP triggering event is progression (TTPTRIG=1). The second subject did not progress and did not die, so there is no TTP trigger (TTPTRIG=.). The third subject did not progress but died at day 59, so the TTP triggering event is death (TTPTRIG=2). The last subject progressed at day 43 and later died at day 109, so the TTP triggering event is the first event, progression (TTPTRIG=1).

Specifics for ANAL2:

The ANAL2 dataset is one of the two files that are "one proc away" from an analysis PROC. ANAL2 is structured with one record for each analysis event variable within subject. This file includes the following variables, in addition to those mentioned previously as common to all datasets:

- TTEVAR (Time-to-event variable, such as TTP, TTF, Survival)
- TTEVALUE (Time-to-event days)
- TTECNSR (Time-to-event censoring indicator, where 0=not censored, 1=censored)

The following example shows the analysis variables of the three TTE records for one subject:

TTEVAR	TTEVALUE	TTECNSR
Time to Disease Progression (days)	213	0
Time to Treatment Failure (days)	213	0
Duration of Survival (days)	235	1

Here the subject has an actual time to progression and time to treatment failure, both at day 213, and thus no censoring for these events. Also, the subject did not die as of day 235, and so is censored at that day.

Specifics for ANAL3:

The ANAL3 dataset is the other file that is "one proc away" from an analysis PROC. ANAL3 is structured with one record per subject, and a set of columns for each analysis event. Thus, in addition to the variables common to all datasets, this file contains sets of time to event analysis variables similar to ANAL2. Variable names are sponsor-defined, and likely based on the type of analysis performed, but might look something like:

Time to Disease Progression variables
- TTP (days)
- TTPCNSR (censoring indicator, where 0=not censored, 1=censored)

Time to Treatment Failure variables
- TTF (days)
- TTFCNSR (censoring indicator, where 0=not censored, 1=censored)

Survival variables
- SURV (days)
- SURVCNSR (censoring indicator, where 0=not censored, 1=censored)

The following example shows the analysis variables in the TTE records for two subjects, one record per subject:

TTP	TTPCNSR	TTF	TTFCNSR	SURV	SURVCNSR
213	0	213	0	235	1
10	0	10	0	54	0

Here the first subject is the same one shown above for the ANAL2 dataset, with time to progression and time to treatment failure at day 213 and a censored survival at day 235. The second subject has no censoring for any events, an actual time to progression and time to treatment failure at day 10, and died at day 54.

USING THE TTE MODEL

The ANAL2 or ANAL3 datasets are designed to get you “one proc away” from your analysis results. Using one of these datasets, statistical estimation and comparison of survival curves (such as those from Kaplan-Meier, logrank and Wilcoxon tests) can be generated for analysis variables with the SAS®/STAT procedures LIFETEST and LIFEREG, or with other statistical analysis software.

More information about how to implement the TTE model is described in the document “ADaM Model for Time to Event Analysis Files“. A 37-page draft, which describes the model and methodology, including analysis mockups and data structures, is currently in review. It will be made available on the CDISC website when it is complete, hopefully within the year.

In the mean time, as your protocol and statistical analysis plan are developed for the study, pay particular attention to any time-to-event analysis (such as time to disease progression, time to treatment failure, and survival time). If you have primary or secondary objects that use these or other similar time-to-event calculations, consider developing a dataset model such as the one described here.

It is always recommended that you discuss your analysis approach with the FDA stat reviewer before submission, regardless of whether you intend to use a CDISC model.

CONCLUSIONS

The primary criterion for the design of datasets submitted to the FDA is to provide clear communication of the scientific results to the regulatory reviewers. Both CRT and ADaM datasets can be used.

There are no ‘cookbook’ answers to what datasets and variables should be submitted. A clear understanding of the CDISC models will allow sponsors to design the proper submission for each indication. Standards and CDISC technology tools, such as those being developed by Trialex Inc., have the potential to cut development and review time and speed approval of new treatments.

If you are interested in learning more about CDISC, visit their fact-filled website. They are always looking for people who can contribute to the development of standards for use throughout the industry.

REFERENCES

The CDISC website is www.cdisc.org. The ADaM TTE and other models, when finalized, will be available at this site. From the CDISC website you can also access information other working groups within CDISC and their products.

Information on Trialex, Inc. can be found at www.meta-x.com.

ACKNOWLEDGEMENTS

Judy would like to acknowledge Dave Christiansen for allowing the liberal use of his slides and thoughts in this paper. Thanks also to Sy Truong for information on the CDISC SAS® tools. Also, many thanks to Mary Lenzen of Pfizer for her continued support of CDISC of and its mission.

Sandra would like to thank Alex Bajamonde at Genentech for his work in developing the TTE model, and the entire ADaM team (including FDA statisticians) who are reviewing the model.

CONTACT INFORMATION

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

Sandra Minjoe
Genentech, Inc.
1 DNA Way
South San Francisco, CA 94080-4990
Phone: (650) 225-4733
Fax: (650) 225-3034
E-mail: sminjoe@gene.com

Judy Phelps
Pfizer, Inc.
P.O. Box 159
Chelan, WA 98816
Phone: (509) 687-2432
Fax: (860) 715-8464
E-mail: judy_a_phelps@groton.pfizer.com