Implementation Plan for CDISC SDTM & ADaM Standards at MedImmune

Alan Meier, MedImmune, LLC., Gaithersburg, MD

Abstract:

The Clinical Informatics group within MedImmune, LLC., has embarked on an implementation project for CDISC SDTM and ADaM data submission standards, known as CDISC-Lite. The main aspect of this implementation plan is to produce submission and analysis SAS data sets for use within the organization that can easily be converted to submission ready, CDISC compliant data formats.

While both STDM and ADaM call for data to be stored in a vertical structure, most people in the Clinical Development process are more "at-home" with a horizontal structure. In a simplified view, CDISC-Lite's goal is to produce SDTM and ADaM data sets that are approximately "one PROC TRANSPOSE away" from the standards' requirements.

This paper will provide details on the implementation plan and provide lessons-learned from the on-going pilot project.

Project Summary:

There are a number of reasons for moving forward with the implementation of CDISC's (Clinical Data Interchange Standards Consortium) SDTM (Study Data Tabulation Model) and ADaM (Analysis Data Model) data models. These include the pending FDA rule proposal and increases in data analysis efficiencies. For MedImmune, we made the decision to move forward with an implementation at this time. In order to create an implementation plan, the most popular implementation methodologies were examined and determined to be lacking. The main area of concern centered on the vertical structure of the models which clouded the temporal and logical relationship of the records for the observer.

The proposed final implementation will use a CDISC "Lite" data set structure. "Lite" implying that it is not quite the real model. This structure will keep the data sets in a horizontal format, maintaining data associated with a single visit together, while using CDISC names and terminologies and be approximately one "TRANSPOSE" away from a full CDISC compliant SDTM/ADaM implementation.

After consideration, we decided the implementation would be conducted using a "phased-in" approach. The benefits to this approach include limiting the liabilities if issues arise, being able to gauge the effectiveness against existing practice, determining the impact on the existing standards, policies and culture and determining the impact on the various functional areas beyond the Programming

group. The pilot will begin with a single study. If successful, it will be followed up with implementation of all new studies and eventual conversion of the other on-going studies in the pilot drug project. Finally, it will be rolled out to all new studies. Conversion of other on-going studies will be determined on a study-by-study basis.

In addition to the traditional method, the pilot will create a second set of "raw" and analysis data sets using the CDISC-Lite data set structures. CDISC terminologies will be implemented for the CDISC-lite data sets where appropriate. A second set of Clinical Study Report (CSR) tables and listings will be programmed using the "ADaM-Lite" set of data sets. This will ensure timely completion of study project deliverables, limiting our liability if issues arise, while testing the feasibility of this implementation and determining the extent of cross-project macro modifications needed. Associated documentation for each set will be created and maintained.

At the pilot's completion, submission ready data sets and documentation from the "Lite" data sets will be created to determine the effort necessary to complete a submission-ready CDISC compliant package. Also, the affected Standard Operating Procedures (SOPs) and Work Practice Documents (WPDs) will be updated appropriately.

This pilot will not include the Trial Design Model data sets. It will include the creation of the definition metadata file (DEFINE.DOC), but that will remain in MS Word / Rich-Text Format (RTF).

Background - What and Where is CDISC:

Per the CDISC mission statement:

> "CDISC is a global, open, multidisciplinary, non-profit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare."

Below is a list of CDISC projects that we will be using and a brief summary:

- ***Study Data Tabulation Model (SDTM)*** – Standardized content of CRF tabulations used for clinical submissions and data exchange. The current model is version 1.1 with an implementation guide of version 3.1.1. This is

the version that the FDA is currently accepting. In July of 2007, SDTM version 1.2 was released along with an implementation guide 3.1.2 for public comment. Per discussions at the CDISC DC Area local implementation networking group and the 2008 DIA Annual Conference, the finalized version of these two documents was to be released before the end of 2008. Most of the changes were additions and clarifications to the previous standards.

- ***Analysis Data Model (ADaM)*** - Standardized content of analysis data sets used for clinical submissions and data exchange. The current model is version 2.0. There are no FDA requirements for ADaM data sets at this time. In May of 2008, version 2.1 and an initial implementation guide were released for public comment. Comments were due back by early September.
- ***Case Report Tabulation Data Definition Specifications (CRTDDS) – (define.xml)*** – An XML-based format and content standard for SDTM metadata – extension of ODM.
- ***Terminology*** – Standard vocabularies used in all models. This is an on-going project. The latest release was March of 2008.

Project Justification:

MedImmune has a number of on-going initiatives aimed at improving efficiencies. One area being examined is data standards. After reviewing our situation with respect to standards and system implementations, management determined the time was right to move forward with a CDISC implementation. One of the factors influencing management was the FDA's position on CDISC, and specifically SDTM.

On July 21st, 2004, the FDA announced that as part of their Critical Path initiative, they were accepting electronic data submissions in the CDISC SDTM format. At the same time, it was announced that they were "exploring approaches to require the use of SDTM standards for regulatory submissions." (FDA Press Release P04-73 http://www.fda.gov/bbs/topics/news/2004/new01095.html) At about the same time, the eCTD specifications were updated to create a folder specifically for SDTM-based data sets.

At a public meeting in February 2005, Randy Levin, M.D. then Director for Health and Regulatory Data Standards Food and Drug Administration, provided the agency's position on the use of standards in the industry. He outlined the follow goals:

Improve patient safety and reduced costs by reducing time to market for safe and effective treatments

–Improve efficiency of evaluation of safety and efficacy of investigational treatments

•Facilitate communication between regulatory authority and applicant

- Facilitate development of efficient review environment (e.g., access to data, orientation, redundancy, training, analysis tools)

– Improve efficiency for clinical research
  - Facilitate design and conduct of clinical trials
  - Facilitate communication between researchers and study sponsor (e.g., between CRO and drug company)
  - Integration with the electronic health record

On December 11, 2006 the Department of Health and Human Services issued a notice of proposed rulemaking in the Federal Registry (Vol. 71, No. 237, The Regulatory Plan Page 72783 http://edocket.access.gpo.gov/ua061211/pdf/ua061002.pdf) of a proposed rule that would require clinical study data be provided in an electronic format using standard data structures. It went on to state in the abstract (page 72784) that this requirement would be for all NDAs, BLAs, and ANDAs and specifically name CDISC's SDTM as the required format. See the proposed rule here (http://www.reginfo.gov/public/do/eAgendaViewRule?ruleID=279292).

In May of 2008, the FDA published the final version of the PDUFA IV Information Technology Plan that indicated that the anticipated publication date for the new proposed rule will be September of this year (http://www.fda.gov/OHRMS/DOCKETS/98fr/FDA-2008-N-0352-bkg.pdf).

FDA staff members are currently being trained in the SDTM file formats and their associated tools and MedImmune has already received requests to submit electronic study data in SDTM format. In recent interactions with FDA staff members, sponsor companies were repeatedly encouraged to not wait for the final rule.

In addition to meeting the regulatory requirements, there are additional efficiencies that will be gained by using an international standard for data storing and analysis. Some of these include more efficient use of standardized analysis programming and reporting, more efficient data sharing between partner organizations and regulatory agencies, and more familiarity of new employees and internal customers to a common data structure. These standards will also align with the data standardization efforts outlined in a recent eClinical initiative undertaken at MedImmune.

Project Implementation Alternatives:

When planning for the implementation, we reviewed the most popular implementation strategies. Susan Kenny's description of these strategies is often quoted in CDISC presentations and discussions.

At the 2005 PharmaSUG conference, Susan J. Kenny, of Maximum Likelihood Solutions, Inc and Octagon Research Solutions, Inc., identified four basic SDTM implementation strategies (Paper FC03 - http://www.lexjansen.com/pharmasug/2005/fdacompliance/fc03.pdf). Below is her graphical summary along with the pros and cons she identified for the HYBRID method.

PARALLEL METHOD

This development path is illustrated as:

DBMS Extract → *SDTM Domains*

DBMS Extract → *Analysis Datasets*

RETROSPECTIVE DEVELOPMENT

This development path is illustrated as:

DBMS Extract → *Analysis Datasets* → *SDTM Domains*

LINEAR METHOD

This development method is illustrated as:

DBMS Extract → *SDTM Domains* → *Analysis Datasets*

HYBRID METHOD

This development method can be illustrated as:

DBMS Extract → *SDTM Draft Domains* → *Analysis Datasets* → *SDTM Final Domains*

With the HYBRID method, the differences between SDTM Draft domains and SDTM Final domains are envisioned to be small. The SDTM Final domains contain the subset of variables or records that are optimally created during the analysis or at the final stage of submission preparation. An example is the creation of USUBJID. This variable is required in the SDTM and provides a unique key identifier for a given subject. In some situations, however, the creation of USUBJID cannot be

defined until all studies are complete since a given subject may participate in multiple trials. Other examples include the creation of expected variable that is present in all Findings domains that indicate the data record considered to be the baseline value (e.g. 'EGBLFL'). Since these indicator flags likely would be derived in the AD's, creating the Final SDTM domains retrospectively from the AD's prevents redundant derivation and eliminates the possibility of discord between SDTM domain and the analysis dataset. Finally, population indicator variables, such as those for intent-to-treat or per protocol status, can be optimally created in the AD and then placed in the supplemental qualifier domain.

Project Implementation "Lite" Method:

While the proposed methodology for MedImmune was not one of those that Susan identified, the closest of her methods was the HYBRID. The major difference is that some of the MedImmune CDISC-Lite data sets will remain in a non-normalized or horizontal structure. The reason for this is that when working with these data, it is easier to visualize and manipulate data when temporally related observations remain in the same record. In fact, FDA staff members have commented that the first thing they do with CDISC-compliant SDTM and /or ADaM data sets is transpose them into a non-normalized structure for analysis. Also all supplemental data will remain on the record as opposed to being stored in a separate data set.

The data sets not normalized would include such domains as Vital Signs, questionnaires and other efficacy type domains. Data sets that would remain normalized would include those domains that are currently in a normalized structure like Laboratory data, Medical History and Physical Exams.

An example of how this methodology will vary is to look at a Vital Signs data set. For a given subject visit, the CDISC-compliant SDTM data set will have multiple rows and a few generic variables whereas the SDTM-Lite data set will have a single row and many test specific variables. See the tables below:

CDISC-compliant SDTM Data Set		
VSTEST	**VSORRES**	**VSORRESU**
TEMP	34.7	C
SYSBP	153	MMHG
DIABP	98	MMHG

MedImmune SDTM-Lite Data Set					
TEMP	**TEMP_U**	**SYSBP**	**SYSBP_U**	**DIABP**	**DIABP_U**
34.7	C	153	MMHG	98	MMHG

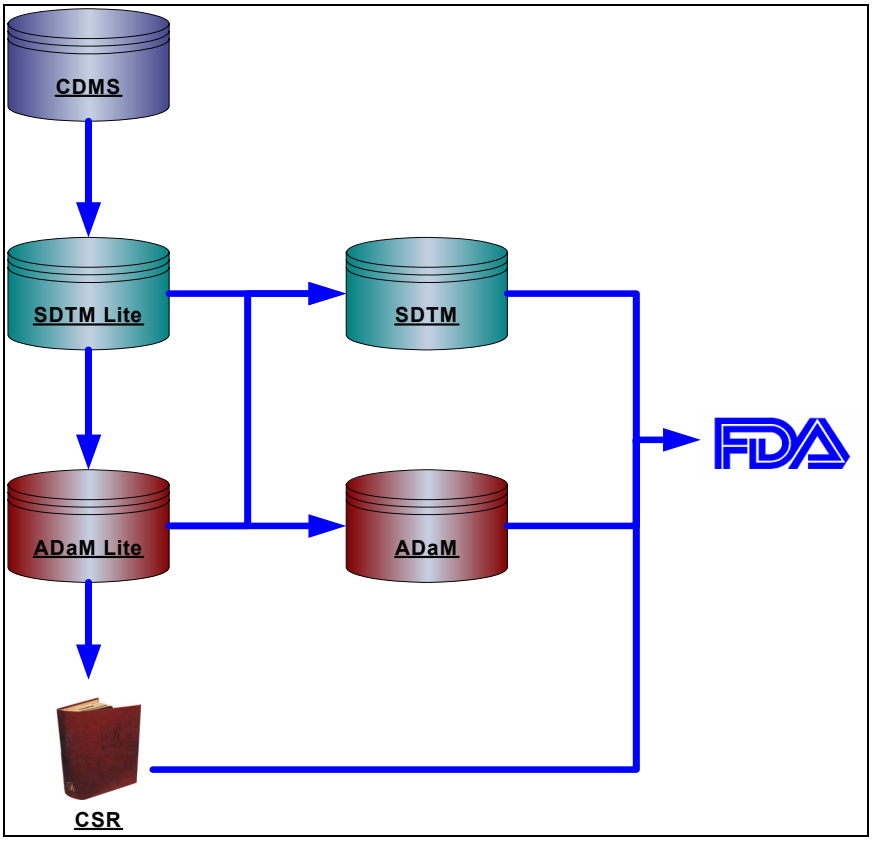
A second difference is that all variables that are not defined in the core CDISC-compliant data sets that would normally go into the Supplemental Qualifiers (SUPPQUAL) data sets will be included in the MedImmune SDTM-Lite data sets. These additional variables will be moved to the appropriate SUPPQUAL data sets when the final CDISC-compliant SDTM data sets are created.

As with the HYBRID method, the analysis defined variables (e.g. timing, baseline, flagging, etc.) will be created with the analysis data sets and copied into the final CDISC-compliant SDTM data sets when the SDTM-Lite data sets are normalized just prior to submission. All of the other variables that can be optimally created directly from the Clinical Database Management System (CDMS) will be created as part of the SDTM-Lite data sets.

As for variable names within the SDTM-Lite and ADaM-Lite data sets, CDISC-compliant SDTM and ADaM variables should be used when possible (identification and timing variables) – but only if the attributes and contents conform to the CDISC approved specifications and terminologies. For the non-normalized fields, the variable names should follow the approved CDISC vocabularies. Any qualifying fields (units, standardized reporting value, standardized reporting units, etc.) would be easily identified by using an underscore '_' between the variable name and a MedImmune defined vocabulary of suffixes (e.g. 'U' for units and 'ST' for standardized reporting).

This development method is illustrated as:

Figure 1



It is expected that a standardized transformation program will be created for converting CDMS data domains to MedImmune SDTM-Lite domains using a spreadsheet specification. Standardized domains will be defined once and the specifications can be reused from study to study. Similarly, if a standardized naming convention is adapted for the non-normalized variables, a similar transformation program can be created to transform the MedImmune "Lite" data sets into CDISC-compliant SDTM and ADaM data sets.

The advantages to this method are:

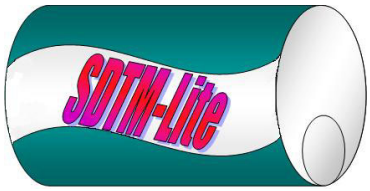
• Using SDTM-Lite domains as input to analysis data sets allows for standardization of analysis data set structures and programming methods to produced study report summaries.
• The variables in the SDTM-Lite that need biostatistical input, such as indication of baseline records or creation of population flags, are done in harmony with analysis data sets to avoid the possibility of discrepancies.
• If important, derived fields can be added to the SDTM domains thus providing the reviewer with both CRF and analysis variables.
• Final completion of the SDTM and ADaM compliant domains can be done at the time of submission.

The disadvantages of this method are:

• An extra set of data transformations is needed to create the SDTM-Lite data sets from the CDMS data sets.
•The development of the analysis data sets relies on the completion of the SDTM-Lite domains.
• The SDTM-Lite domains are created for all clinical trials regardless of whether they will be part of a submission.
• This may be potentially more difficult to manage if the data management and/or the biostatistics functions are outsourced.

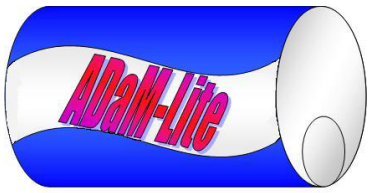
Project Pilot:

We decided on a pilot project to test this approach and to determine areas requiring addition enhancement. The pilot will involve one study. It will include creating a set of SDTM-Lite and ADaM-Lite data sets for that study in addition to the conventional "raw" and analysis data sets.



The creation of the SDTM-Lite data sets will include the enhancement of a current TRANSFORM.SAS program that is used to convert eDC SAS data sets to MedImmune traditional structures prior to creating analysis data sets. Included in the enhancement will be field definitions that can be used to assist in creating the

DEFINE.DOC files for validation and submission. The implementation will be based on the draft SDTM Implementation Guide 3.1.2. While version 3.1.1 is the current approved version, the final 3.1.2 document is due out shortly.



From the SDTM-Lite data sets, the ADaM-Lite data sets will be created with the specifications provided in the Statistical Analysis Plan (SAP) and the Statistical Programming Plan (SPP). Where possible, the TRANSFORM.SAS program will be used and where not, standardized SAS macros will be created to complete as many of the remaining conversions as possible. The current draft ADaM Implementation Guide version 1.0 will be used. This is the first Implementation Guide to be released for the ADaM data sets.

The associated DEFINE.DOC files will be created for both sets of data sets. While these data sets are not intended for regulatory submission, the accompanying DEFINE.DOC files are needed for validation.

Once the "Lite" data sets are created, a subset of the CSR tables, listings and figures will be programmed to test the robustness of current SAS standard reporting macros. Any deficiencies will be noted and turned over to the Standard SAS Programming committee as requests for updates.

Also upon completion of the "Lite" data sets, a separate effort will be made to perform the final transformations to CDISC-compliant SDTM and ADaM data sets and their associated DEFINE.DOC files.

Validation of the data sets and CSR data will include only a code review and comparison of the duplicated outputs.

All programs and macros created or enhanced during this pilot will be reusable for future studies and conversion. Additional flexibility and enhancements will be added to these as necessary. Documentation on these programs and macros will be completed as part of the post-pilot activities.

Post-Pilot Activities:

Following the pilot, a document will be created to summarize the results. We will assess the effort expended and estimate the future effort needed to implement the "Lite" method. It will outline the necessary changes to the current processes and the current standard SAS macros. It will also indicate areas where additional process review will be necessary for efficiency.

The summary will identify the changes necessary to any existing SOPs and WPDs. These will need to be updated and approved prior to moving forward with any study using only CDISC-Lite data sets.

Validation and documentation of any programs and/or macros that will be created in the pilot and will be used for other studies will be required. Training on these specific programs and/or macros will also be required in addition to CDISC general training.

A review of MedImmune's terminologies will be made against the CDISC approved terminologies. Any discrepancies identified will be addressed where possible.

Finally, an implementation plan will be needed to identify which other projects and/or studies will be migrated to the CDISC-Lite structures.

Inter-Functional Area Activities:

Database Programming:

The eDC vendor MedImmune is using has availed us a unique opportunity to gain efficiencies by implementing some CDISC standards at the source, where the data is collected. It would greatly facilitate the implementation if CDISC standard terminology and variable attributes can be built into the new global library.

Because we are implementing both processes simultaneously and because of the timelines associated with getting new studies designed, we will not always be able to implement all of the CDISC standards as eCRF panels are created. This might require changes in our standard eCRF panels as we move forward.

Clinical Data Management:

As we implement these standards, variable names and attributes will change. There will also be changes to the values stored in these variables as we incorporate CDISC terminologies.

Biostatistics:

Implementation of SDTM-Lite and ADaM-Lite data set structures will also affect the Biostatistics group as the format and content of the "raw" and "analysis" data set will be changing.

Medical Writing:

Implementation of the standard terminologies may require changes to protocol and CSR templates. These terminologies are often taken directly from the protocol and entered on to the eCRF screens. While the terminologies may be synonymous (e.g. Death versus Fatal for AE severity), using the approved CDISC terminologies will remove the necessity to map the results to the approved terms and document the change.

Clinical:

By accepting CDISC data standards, the FDA has also indicated their preferred terminologies. The terminologies are actually stored and maintained at the National Cancer Institute. Using the approved terminologies will facilitate the creation of the CSR and review by the FDA.

Every attempt is being made through cooperative interactions between groups to implement these standards as quickly as possible. We can't be prevented from making changes "because it was done a certain way in the past studies" or this "won't work in this therapeutic area".

Training:

Training for the SAS programming group will be provided through a series of presentations at team meetings. The presentations will begin with an overview of CDISC and the various standards and eventually drill down to the specific implementations of SDTM and ADaM standards. Training on the new programs and macros will be provided as they come on-line.

Training for other functional areas will be available upon request. The training can be tailored to the level of detail required for each group.

Current Status of Pilot as of February 2009:

The project proposal was accepted. A review of terminologies was made and changes implemented. Protocol templates were updated to incorporate new terminologies. A review of new eDC standard eCRFs was made and where possible CDISC compliant variable names and content were incorporated. Staff training on general CDISC principals is on-going. The pilot study was identified and traditional analysis data sets, tables and listing production is currently underway.