# Standards for Analyzable Submission Datasets for Directed Safety Analyses

Michael Nessly
Director, Clinical Biostatistics
Merck Research Labs
CDISC Analysis Dataset Models Team (ADaM)

nesslym@merck.com

# Outline

- Establishing the Problem/Motivation
- The new submission data standard
- Data Elements and Issues
  - Events
  - Time at Risk/exposure
  - Analysis Populations
  - Risk Factors/Subgroups
- Summary

# Definitions

- **Data Tabulation Datasets (DTD)** - Datasets in which each record is a single observation for a subject.

  *Study Data Specifications, Version 1.0*, http://www.fda.gov/cder/regulatory/ersr/5640studydata-v1.pdf
- **Analysis Datasets –** datasets used for statistical analysis and reporting by the sponsor; submitted in addition to the DTD.

# Study Data Tabulations Model (SDTM)

- SDTM is the future submission data input format to the FDA
- Having a data standard across sponsors, academia and governmental researchers (NIH, NCI) is a tremendously good idea with great potential for enhanced interactions
- This common data structure could be very useful in joint ventures and in-licensing.
- SDTM could assume the role of data source.

# Study Data Tabulations Model (SDTM)

http://www.cdisc.org/models/sds/v3.1.1/index.html

- Very normalized, tall & skinny
- One record per patient per visit per outcome
- Optimized as input for the FDA data repository
- Driven by technological specifications rather than reviewer needs
- Some rudimentary back end tools exist for Clinical Review, but none exist presently for Statistical Review

# Study Data Tabulations Model (SDTM)

- SDTM is dynamic and still changing
- Action is underway to address at least some of the concerns expressed here.
- However, it is unlikely that all of the concerns expressed here can be addressed in SDTM in its current form.

# Data Elements Absent from SDTM *per se*

- Study Phase designators
  - Prestudy, Baseline, Double Blind, Open Label Extension, Randomized Withdrawal, Titration, etc.
- Event and Censoring Flags
- Analysis Population flags
- Exposure Duration flags (censoring structure)
- Patients without events (AE, Med Hist)

# Can SDTM supplant Analysis Submission Datasets?

- The data are there in raw form
- The standard is not optimized for many types of statistical analysis, models or study designs
- The key issues surround finding the variables needed, putting them in the appropriate form for analysis and doing the needed selection, merges, transposes and other programming tasks

# Future Visions On Alternatives to Analyzable Datasets

- Proposes submission of metadata which describe the derivations/logic/analyses which create or define analyzeable datasets from SDTM.
- These metadata describe the programs for data preparation and are an inferior tool as additional work is required.
- Assumes that a ‘back-end’ tool using the datasets and metadata exists

# Statistical review at the FDA

- Access to the right data in a timely manner is key
- Reviewers have had to spend 80% of time becoming familiar with data and structure in application, leaving only 20% of time for review
- Statisticians need more ‘think’ time

# Statistical review at the FDA

- No programming support across 18+ review divisions
  - Compare to Merck: ~75 statisticians, ~70 programmers covering 11 broad therapeutic areas
- 1 statistical reviewer for 4-5 medical reviewers
- Prescription Drug User Fee Act (PDUFA) deadlines provide tight timelines
- Each stat reviewer supports multiple project responsibilities

# Motivating Example

Imagine you have just received the following request from the FDA:

In reference to your submission please provide updated combined analyses from studies 1, 2 and 3 for the following events A, B, C & D

Please provide crude rates, rates per 100 patient-years and Kaplan Meier cumulative rates as well as the datasets supporting these results.

# Directed Safety Analyses Scenarios

- Study report for a trial with defined prespecified safety primary endpoints.
- Program with an emergent adverse experience of special interest.
- Data explorations in past and ongoing trials across programs and sponsors for a potential emergent safety issue for a program or across a class of drugs

# Safety vs Efficacy Differences

**Efficacy Analyses**

- Inference based on primary outcome(s)
- Type I error strictly controlled
- *a priori* specified in detail
- Powered adequated for endpoint(s)

**Safety Analyses**

- Many potential safety endpoints
- No adjustments for multiplicity
- Usually *ad hoc* and/or *post hoc*
- Generally limited power

# Safety vs Efficacy Differences

**Efficacy Analyses**

- Designed for strong inference
- Subgroup analyses for hypothesis generation, not strong inference

**Safety Analyses**

- Exploratory in nature
- Subgroup/risk factor analyses important in inferences regarding risk management

# Safety vs Efficacy Differences

**Efficacy Analyses**

- Generally analyzed as randomized
- If studies are combined usually done as formal meta-analysis

**Safety Analyses**

- All Patients As Treated
- If studies are combined, data are generally puddled

# What is needed?

- FDA Statistical Reviewers would like to have a single file that records all needed safety-related information during the trial.
- Ideally all of this information for a particular patient should be present on one record.
- Additional data at the study level are also frequently needed.

# Data Structures: Issues and Discussion

- Think about the "natural form" of the data
  - Demographics – per subject
  - AEs – per event (adverse event)
  - Longitudinal data – per visit
  - Lab – per measure
  - Physical Exam – usually per abnormality

# Safety Data Analysis Plan

- Although largely exploratory there is structure to good safety data generation and analysis.
- Even if *post-hoc*, there is value in writing down the details of data handling and analysis logic.
- It deserves the same level of rigor and detail that is often involved with the demonstration of efficacy.
- Many of the salient details are buried in copious data coding and entry guidelines

# Safety Analysis Dataset Descriptor File

- The safety analysis dataset should be submitted with a safety analysis dataset descriptor file.
- This should include clear, accurate and precise description of how the elements were derived.

# Key Data Elements for Safety Analysis

- On the surface, the data elements seem simple and basic:
  - Observed Events of Interest
  - Observation Period/Time at risk
  - Population At Risk
  - Patient Status/Discontinuations

# Difficult Issues

- Finding the events
- Linking of time at risk and events to DTD datasets
- Dictionary changes
- Adjudicated endpoints
  - Reclassification of events
- Derivation of Exposure
- Assessing onset and resolution times
- Applying logic for Analysis populations

# Need case definition for events of interest

- Need clear, concise, logical, pre-specified and mappable definitions
- For many analyses need time of onset of emergent events and need to determine the first onset of a constellation of event types.
- Listing of dictionary terms (and any qualifiers should be included).

# Term Hierarchy

Facilitates tracking of the events back to source

- Adjudicated - Other event / fatal haemorrhagic stroke - APTC only -not a confirmed thrombotic event
- Verbatim\Reported - hemorrhage
- Broader - stroke
- Dictionary – cerebrovascular accident
- Body System – nervous system disorders

# Nephrolithiasis Example

- 43 distinct preferred terms across different System Organ Classes
- Many non-specific terms: abdominal pain
- For non-specific terms from 40 to 90% were not associated with nephrolithiasis
- Flag variable makes for easy selection
- Selection/reduction task very difficult without flag

# Example of Risk Factor Definition

This group was defined post hoc as those with either ≥2 major risk factors for coronary artery disease

(current smoker, history of hypertension, diabetes, or hypercholesterolemia)

OR with a prior history of a cardiovascular thrombotic event.

List of MedDRA terms identify each risk factor

# Diabetes Set of Terms

- insulin-dependent diabetes mellitus
- diabetes mellitus
- diabetic ketoacidosis
- diabetic nephropathy
- diabetic neuropathy
- diabetic vascular disease
- type 2 diabetes mellitus
- hyperglycemia
- diabetes retinopathy

# Issues with Terms

- Determined *post hoc* by searching what was observed in particular study/studies
- Dictionary terms change over time
- Variability across studies, programs, sponsors
- Ongoing continual task until all studies are done
- Need to use common dictionary version

# Time at risk example

At least 6 variables are needed to calculate the time at risk by this definition:

- Date of randomization
- Date of first dose
- Date of last dose
- Date of event
- Date of study discontinuation
- Date of study therapy discontinuation

These are patient level meta-data (phasing calendar)

# Analysis Population Flags

- What level is flag best applied?
  - Patient level
    - ITT –randomized='in analysis'
  - Patient by visit level
    - Patient takes forbidden drug and is a protocol violator subsequently
  - Patient by visit by outcome
    - May use All Patients As Treated for AEs but ITT for Mortality

# Issues: Identification of Analysis Populations

- Options
  - Status flags
  - Separate variables
  - Separate records
  - Separate datasets
- Choice depends on:
  - Statistical analysis
  - Study design
  - Dataset structure
  - Priority of ease-of-use vs ease-to-create

# All Patients As Treated Example

- This target population consists of all randomized patients who received one or more doses of test drug therapy
- Patients who are documented to have never received study medication are excluded from analyses
- The patients are counted in the treatment group for the drug they actually received, rather than the treatment group to which they were randomized

# The near future – What should be done in submissions for statistical review

- An analysis dataset in a non-SDTM ADaM model format facilitates analysis in **a standard fashion across protocols**, across development programs, across sponsors OR
- Analysis dataset creation programs using SDTM as source should be submitted as metadata

# The near future – Some Options :

- Optimized analysis datasets in Non-SDTM structure
- SDTM compliant files with additional **rows** including the requisite data and metadata not in SDTM but needed for analysis
- SDTM+ using the SDTM standard with additional **columns** of necessary data
- Sending analysis programs that use SDTM as source that can create the analysis datasets

# Summary

- Most, but not all, of the needed raw data exist in the SDTM datasets.
- Enormous work is needed to find, collect, reduce, merge, derive and process the specific data needed for directed safety analyses.
- The sponsor has already done this work, why make the statistical reviewer do re-work?

# Summary -2

- The SDTM standard as analysis file 'mandates' submission of either the data processing code or all of the logic contained within the analytic programs
- Non-SDTM safety analysis datasets would be useable at the FDA right now whereas a metadata stream (in XML) on how to transform the Events SDTM file would not.

# CDISC ADaM

- Check us out

http://www.cdisc.org/models/adam/V1.0/index.html

- Be aware of the issues
- Actively participating members needed to develop future models
- Learn more about your peers in Industry and at the FDA