PHUSE SEND Data Factory Design

Goals:

A set of R scripts that will create a valid SEND dataset that contains reasonable looking, though made-up data for the needed SEND domains. The data should follow the SEND IG and pass the validation rules.

This is to allow people to demonstrate software or visualizations or SEND structure without using real or confidential data. It is not meant to supplant the need for CDISC to manually create examples for the SEND IG.

This can be useful for vendors, standards bodies, and other parties.

Is it a goal that the data set makes sense biologically? Scientifically – this might be unnecessary and difficult. Should the data tell a story? (Higher dose groups have – more findings, higher or lower blood test results? Lower body weights?)

Methods:

R-scripts will be created based on R version Rx64 3.4.4 and Shiny 1.0.5

The scripts are then shared in the GITUB

Open Questions

Set up with yml as designed by the clinical code sharing group to allow it to run with the "PHUSE" package GUI?

Priorities of development

Repeat dose study first and then the other study types will be considered.

SEND 3.1 version first.

Requirements:

Shiny is used as a user interface.

Shiny input selections include:

SEND version (3.0 or 3.1) and DART 1.1

Control terminology selection by date

Study type, choices are Single dose, repeat dose, carcogenicity, safety pharmacology, teratology

Number of dose groups

Number of animals per dose group for main study groups

Include TK or not (and number of animals per dose group)

Include recovery or not (and number of animals per dose group)

Output as XPT or CSV

Settable percentages for not done statuses

Setable percentages for positive findings for palpable masses, clin obs, macro and micro

% increase (or decrease) for each dose group over time for body weights, food consumption, clinical pathology, TK

Tissue selection for macro and gross (with a default set)

Clinical pathology test parameters (with a default set)

A reset button to set back to original defaults

A TSParm code should be included that states that this is FAKE data.

Dataset outputs

Repeat dose study output domains and endpoints are taken from the SEND 3.1 Codex

Data findings requirements for normal animals

All input selections should have defaults read from a set of configuration files and remember the last settings as well as retaining original defaults

Use control terminology for all output based upon a loaded control terminology file

Allow control terminology files to be added (or retrieved directly from the NCI website)

Reference range data should be used where available, for example for clinical pathology.

Is there public information on ranges by species? Is that critical?

If they are in configuration files, how can a user modify that?

Data findings requirements for treatment affects

Include some observations that are abnormal/dose responses that can be detected through analysis. Perhaps allow setting of which parameters should have treatment effects and specify +/-SD for outliers or dose groups.

Consideration of correlated abnormalities within specific animals.

Assumptions:

There already exists tools to create the define.xml from the datasets

Will use temporary location on the server to create files, then allow download as a zip when completed.