Project workflow

Step	Artifacts	Tools	Note
Collect requirements from clients.	Project Requirement	Teams files	Requirement may changes
Create POC/draft, give to client And collect feedback	Prototype code and result	R/Knime/python/sas	Optional
Plan project	Project execution plan	Teams files	Optional for small project
Collect data, SAP, table shell and etc from clients or partner	Save raw data, SAP, table shells and etc to shared project space	Team files?	Data are changing. We don't have all data be beginning.
Write table specs	Specs	Team files? Github?	Sometimes we can't create table spec since no data yet. Optional for small project
Work on tasks	Source code such as R, SAS, Python, Kime in Github. Table/List/Figure data(csv) and pdf version stored in shared project space	Github, team files?	
QC	R/SAS/Python/Knime/excel	Team files	
Final deliverable	Combined deliverable to client	Github, team files?	

Project requirement template

Study Background

short description of the study

Client Requirement

List of requirements from clients. This section should be a description of requirements from the client's point of view.

Data and Documents

List of data(including raw data in sas or csv format) and documents(including protocol, SAP, table shell and etc) we can get from clients or other partners.

Need a fixed folder structure for documents:
CRF
Protocol
SAP
TableShell

Document name	File server(SAS, Team files)	File path

Project execution plan template

Project Goal

Brief description of the study and the project's goal

Deliverables

List of deliverables to the client as the result of the project. This can be a list of tables, lists and figures. Only describe the name and content of the table/list/figure. There is no need to have table shells at this stage.

Execution plan

List steps and deliverables of each step.
E.g.

1. Write spec for processed raw data
2. Generate processed raw data
3. Write spec for derived data(analysis data)
4. Generate analysis data
5. Create summary data
6. Write spec for table/list/figure
7. Generate table/list/figure in csv format. (data type is string not float).
8. QC result. Potentially develop tool to do QC automatically
9. Use tool to generate table from summary data and configuration data(title, analysis-set)
10. Generate final deliverables.

Including team lead, supportting team members.

Tracking for tasks and intermediate deliverables. Team plan to track task progress and assignment

Process Raw data Spec template

Steps to create processed raw data spec:

1. Read all raw files and find all the variables from raw files and write all varialbes as the initial content of the spec.
2. Mark variables as common variable if the variable appear in almost all raw data files.
3. Next Add annotation number and full labels to the variables. The annotation number and full labels can be found from the CRF file. Some study has a ALS file with all the variable information incuding annotation number and full labels. Some study only has CRF file, we need to extract the annotatin number and full labels from the pdf then.
4. Put priority number the variables. Priority is a integer number such as 1, 2, -1 we assigned to each variable. 1 means very important and it is mos likely used in summary table. 2 means no that important, it may or may be used in summary. -1 means not important at all and they will not be needed in further operations.
   The rules to decide priority:
   a. If the variable has annotation number(which comes from CRF), priority is 1

    b. If the variable has a companion variable (such as AESEV and AESEV_STD), set the main variable priority to 1 and its companion variable (such as xxx_STD) priority to 2. Copy the label from the original variable to its companion variable.
    c. If the variable has a sets of raw variable(such as AEDAT and AEDAT_RAW), set the raw variable priority to 1 and other variables to 2. The reason is normally raw variable always has value and the original variable may not has value since either the raw value format does not fit the original variable or the raw value does not fit to the code list of the original variable. This rule may apply to variables besides date.

5. Upto this step, we have added priority and label to CRF related variables. There are still some other variables not defined in CRF. Next step is to manually add priority and label to those variables:
    a. For varialbes not command and not CRF variables (Dict coded variables), Set priorityh to 1 for those need in summary table. Others priority 2. Get label from a predefined variable , label mapping list. Need the predefined mapping

    b. For AE, MH, normally we use SOC and PT. (we don’t use _CODE).

    c. FOR Cm, normaly we use ACT4 and PT.
    d. For other variables, need use own judgement. Read raw data and CRF/protocol.
6. Create a new file with each raw dataset as a row.(mostly manual) Columns: data set name, description, RandOnly, and VisitRemoveName, comments
    a. Randonly column is used to indicate if the raw dataset only need keep randomized patients. The value is N for dataset such as DM, disposition, screen failures. The value is Y for other data set. (the reason being most table summary only need patient randomized). May be obselte. We will generate two sets of processed raw data. One with in-study patient only. One with all.
    b. Find varialbes in each data set which indicate visit.
       Set the VisitRemoveName as Y if we want to remove those variables. Rule:
       for those dataset only has one visit, remove such as demographic, (We need list of comman dataset as examples).
       For those dataset which visit has no meaning or we dont care such as AE(we only care date), remove.
       otherwise set as N.

7. Use the spec to generate processed raw data (by R, Knime, python) from raw data
   the rule to generate process raw data:
   Remove variables with priority value as –1

   To generate simplied processed raw data:
   a) Only keep variables with priority 1. Only Keep rows with randomized patient if the data set’s RandOnly is Y and the patient is randomized.

The patient is randomezed is decided by the Rand dataset, where the patient answered yes to Randomized?(whether patient is randomized or not) and has randomized date.
B) remove visit columsn if the VisteRemoveName variable is Y for the dataset.

As last step, add a sheet for each dataset to the spec. Each sheet has the all the variables in the dataset with the same row value as in the main spec. (This is just for convinienice, not used to generate data).

Processed raw data genration tool:
Input: all the raw dataset, spec excel(including each variable priority, label), and step 6 information.
Output:
Processed raw data set (including all patients)
Processed in-study enriched raw data set(only only including instudy patients. Has extra enriched columns)
Simplified raw data set(including all patents, with less columns)
Simplifeid in-study enriched raw data set(only including in-study patients with extra enriched columns. Not all columns).

Analysis(Derived) data spec template

Purpose: Get the derived dataset ready for Summary dataset/Listing and other types of analysis.

Workflow:
1, Generate a COMMON derived dataset, this one can include variables such as: First Dose date, first dose time, treatment group, population flag, necessary baseline variables, necessary demographics variables and others. In this dataset, the subject ID needs to be unique.
Rules:

- Some variables are always needed, especially for common tables.
- For efficacy tables, some efficacy variables always need a set of factor variables in the model such as MMRM. E.g.
  ANCOVA model in log scale with dose group, age, gender as factors and baseline titer as covariate. (dose group, age, gender should be in COMMON_X)
  MMRM modeling log NT-proBNP at Week 12 as a function of fixed effect terms for baseline log NT-proBNP, treatment, atrial fib status, visit, and treatment-by-visit interaction, as fixed independent variables. (treatment, atrial fib status, treatment-by-visit should be in COMMON_X, visit infor is not in COMMON_X since COMMON_X only incude variables common among all visites)
  MMRM modeling Change from Baseline in 6-minute walk distance as function of fixed

effect terms for baseline 6-minute walk distance, treatment, atrial fib status, visit, and treatment-by-visit interaction, as fixed independent variables. INL1 Pooled includes three INL1 groups together. *(treatment, atrial fib status, should be in COMMON_X)*

- *Some study tables need partient basic infor like demorgraphic, we can add those to COMMON_X*

2, Merge COMMON derived dataset with Processed Simplified (Unpivot) raw data.
Rules

- For each table, choose processed data set and merge(join) with COMMON_X, generate the derived data set.

3, Merge with other with Processed Simplified (Unpivot) raw data if needed.
Rule:

- Some derived data need to be generated by processed raw data and other derived data.

4, Derived new variables based on needs from Table/Graph/Listing or other analysis.
Rule:

- Split table shell in different type such as
  - AE, t_ae flag.
  - MH, CM, Eric Wang to add more details

Note: To simplify this process, a set of Processed Simplified (Unpivot) raw data would help. In addition, we can identify tables and dataset into different types, and predefine what is needed for each of those types. (Example: by visit Lab always want Baseline, Change from Baseline. AE related always want TEAE flag)

Summary data spec template

Purpose: Get the Summary Dataset ready for Tables and graphs.

Workflow:
1, Split Parameters into two types: Numeric and Categorical. We will have different kinds of summaries for them.
2, For each Parameter, run a predefined summary. (Need to predefine loop varaible/values)
3, Round up decimal place as predefined.

Summar Data to Table Generation

One summary data ssv for one table or on summary data for multiple tables.

Input:

- Summary Data (csv format)
- Table configure

Output

- Table in pdf format