Hi Jie,

The following is the list of output tables we’ll need to produce in these two months. Let’s talk through it as well as the protocol tomorrow during our meeting.

1. Descriptive statistics (30 studies x 2 = **60 tables**)

[Jie:]5 cohorts \* 6 outcomes = 30

* 1. Includes both the univariate and bivariate statistics, which can be combined into one table for each study
  2. For each of the 30 studies, there will be two tables: one table before continuous / numerical variables being broken into categoricals and the other after (where all variables are categorical)
  3. For a table involving continuous variables, instead of breaking a table into two sub-tables, make a long table whose columns include statistics for both the continuous and categorical variables, i.e., for a continuous variable, the values will be NA in columns specific for categorical variables, and vice versa.
  4. Includes statistics for the outcome variables and the predictors
  5. Each row is a variable (if continuous) or a variable category (if categorical).
  6. Columns
     1. For continuous variables (these columns will only exist if there are continuous variables in this cohort)
        1. The number and percentage of the missing values;
        2. The percentiles: min, 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 99%, max and the mean (all excluding missing values)
     2. For categorical variables
        1. Whether or not it's a reference category

[Jie:] how to choose the reference category?

Female0 male1 missing-reference

EDSS – 0 1 2 3 4 : 0 as reference

分得均匀。

Suit logical

Meaningful (contact with medical expert)

* + - 1. The number of instances and the percentage for that variable,

[Jie:] Age0, Age1, Age2?

* + - 1. How many instances in this category has a positive response (output) and the corresponding percentage (these are the bivariate descriptive stats)
      2. There'll be one category for the missing value (if there are missing values for this variable). Detailed methods for treating missing values to follow.
      3. For example for gender,
      4. Three categories: female, male and missing, corresponding to three rows, one for each category;
      5. The columns are:
         * Whether this is a reference category,
         * How many instances are there falling into this category and the corresponding percentage.
         * How many instances in this category has a positive response (output) and the corresponding percentage

1. Cohort contingency table (this is included in the single-variable descriptive statistics but for clarity it can be a stand-alone table): the cohort name, the outcome name, number of patients and number and percentage of patients with the positive outcome in every cohort;(**1 table**)
2. Model performance table: 6 cohorts by 5 outcomes of AUC values and the corresponding confidence intervals. Each row is an outcome and each column is a cohort. (2 x 2 = **4 tables**)
   1. For each study, there'll be two results for sensitivity analysis (the mean AUC + standard deviation) and two results for AUC and the confidence interval. The reason that there are two results for each is that one is for the elastic-net and the other is for the logistic regression without regularisation.
   2. The ‘sensitivity analysis’ is repeating the modelling 10 times by changing the random selection of patients with multiple appearances in one cohort. This will lead to a mean AUC and a standard deviation.
3. Variable importance for each study (a study is defined by a cohort name and an outcome) (**30 tables**)
   1. Choose only one from the 10 sets experiments above for each study to measure the importance (the same one for computing the AUC confidence interval)

[Jie:] what do you mean by saying the same one for computing the AUC confidence interval?

* 1. Every row is an important variable (depending on the sample size of the cohort and modelling result, e.g., top 10 most important variables can be included in one study)
  2. Two methods for measuring the importance:
     1. Average magnitude of the coefficient from every elastic-net model used for that study. Columns:
        1. Variable name
        2. Variable description
        3. Percentage of that Boolean predictor being positive (by now all variables are binarized)
        4. Coefficient value in the final logistic regression without regularisation plus the confidence interval

[Jie:] what is the final logistic regression obtained?

* + - 1. Odds ratio and the confidence interval corresponding to the coefficient above
      2. Importance rank (because the protocol proposes to use two methods for calculating the importance, there will be two values, one from using the coefficient magnitude and the other from accuracy loss after removing that variable)
      3. Number of times the variable is selected in all elastic-net models during cross evaluation (optional)
      4. Average magnitude of the corresponding coefficient in the elastic-net models during cross evaluation (optional)
      5. Average odds ratio from all the elastic-net models during cross evaluation (optional)
    1. AUC decrease after removing a specific variable. Columns:
       1. AUC decrease after removing that variable
       2. Importance rank

1. Table of probability deciles: (**30 tables**)
   1. Each row corresponds to a decile in the predicted risk / probabilities;
   2. Three columns
      1. Decile index
      2. Average predicted probability of having the positive outcome in the deciles
      3. Actual probability (empirical frequency divided by the total number of patients) of having the positive outcomes in the deciles.
2. A number of other analytical approaches (e.g. SVM and RF depending on the resources availability and time) as sensitivity analyses
   1. For every different modelling method, similar tables in Steps III, IV and V above will needed to be produces, which is **125 tables**.