**For step (1) and (2), only those naïve HAE patients from criteria 1-3 with minimum look back window of 24 months will be used as HAE patients.**

**(1) ML-0: feature selection: bagging LASSO/random under sampling random forest**

We have 12 Rx indications, 8 procedure indications and 58 diagnosis indications. Each indication has 3 variables (total frequencies, average frequencies over the look back period and binary flags). 12+8+58=78 and 78x3=234. Also there are 3 ER variables, age, gender, region, total another 6.

Therefore, we have 240 variables in total.

This step is to use bagging LASSO or random under sampling random forest for feature selection among total 240. Average selection frequencies from bagging LASSO and average importance from random under sampling random forest will be output for selection decision. (number of features to be selected is based on the frequency/importance distribution)

**(2) To build predictive model**

**ML-1A: bagging biased SVM**

Each HAE patient has 200 non-HAE patients matched with similar lookback period, those 200 non-HAE pats will be divided into 20 mutually exclusive folds (10 HAE pats each fold). We run 20 iterations, for each iteration: 1 HAE + 10 non-HAE pats are used for training, biased SVM is applied to this sample, 5-fold cross validation will be used for grid search of cost parameters (specifically, with given cost parameters, we use 80% of the sample to build biased SVM model and calculate Area under precision-recall curve (AUP) (here precision=PPV, recall=sensitivity, R package ‘PRROC’ is available) , AUP is averaged over 5 folds and used for pick optimal cost parameters).

**ML-1B: random under sampling random forest**

At each iteration, sample same number of non-HAE patients (without replacement) and combine it with all HAE patients (thus non-HAE/HAE=1:1), then we build a random forest with this new sample. Final classifier is defined as aggregation of all random forests.

**ML-1C: bagging biased SVM with stratified sample**

Repeat **ML-1A** with stratified sample.

(Optional: ML-1D) I will try some other algorithms such cost sensitive boosting.

**(3) Model evaluation and pick best model**

All models from step (2) will be evaluated based on cross-validation results (new random split), a best model is defined as with highest average AUP.

**(4) Add extra highly scored patients (all patients from selection criteria 1 to 4)**

Originally, we have 9692 patients from criteria 1-4, we will apply model selected from step (3) to score the remaining patients (those among 9692 patients yet not used for model training in step (2) and (3)), those highly scored HAE patients will be added to the training sample of step (2) as our final HAE training sample, and the picked best model is retrained with this final HAE training sample with their matched non-HAE patients.

**(5) Scoring all patients**

**Scoring sample**

Scoring sample is defined as patients who have at least one positive rx/proc/dx predictor during 2010/1-2015/7.

The model output from step (4) is applied to the scoring sample.

**(6) HCP targeting**

Pre-defined HCP association rule is applied to highly scored patients. A HCP list will be derived then with number of highly scored patients.