Document for Shire HAE follow-up analysis project

1. Final delivery of new scoring
   1. Data
      * The updated 973-patients (ER is extracted only using Dx information)

* Location:

Kgxsapp101: F:\Jie\Shire\_follow\_up\01\_data\dat\_hae\_973\_dxER.csv

* The criteria for extracting the datasets

1. Remove the patients with \_FLAG=1
2. Remove the patients with AGE <= 12
3. Remove the patients with EVENT < 3
   * + The ~200K for training the model;

* Location: F:\Jie\Shire\_follow\_up\01\_data\newdata\_200K\_3M\nonhae\_200K\_v2.csv
* The criteria for extracting the datasets

1. Extracted the pure non-HAE dataset ~123M from original 123M which is extracted from ~300M by Zhenxing using the following criteria
2. excluding the 2.3M dataset above from the above pure non-HAE 123M
3. Sample 200K from pure non-HAE 123M using Dong’s sampling method (i.e. for each HAE patient, find the ~200 non-HAE patients who can match him/her using lookback\_days.).
   * + The ~2.3M for evaluating the model and setting the thresholds

* Location(the split 10 files): F:\Jie\Shire\_follow\_up\01\_data\newdata\_200K\_3M\split\_3M\_A1E1(C2)\nonhae\_3M\_split\_\*.csv
* The criteria for extracting the datasets

1. The ~2.3M subset needs to have the same distribution of the lookback period as that of the ~123M pure negative patients
2. The ~2.3M subset is randomly stratified from the ~123M pure negative patients.
   * + The ~110M new scoring data

* Location (kgxsapp101):

F:\Jie\Shire\_follow\_up\01\_data\newdata\_200K\_3M\ split\_newScoreMay27\_1to286\\*.csv

F:\Jie\Shire\_follow\_up\01\_data\newdata\_200K\_3M\ split\_newScoreMay27\_ 287to552\\*.csv

* The criteria for extracting the datasets
  + - * 1. Lookback>=24 months
        2. RX\_FLAG=1
        3. DX\_FLAG=1
        4. Lookback\_days > = 662
        5. Patient age > 12
        6. Patient gender <> ‘U’
        7. Index date >= 201404 (Index\_date >= to\_date('201404’, ‘YYYYMM’))
        8. Event > 0 (with at least one active predictor)
  1. Code

F:\Jie\Shire\_follow\_up\02\_Code\GoFromBaggingForest(HAECentral)

* + - Reproduce the result of C.2\_May27 which is saved in the following folder. There are several steps here.
* Splitting the data into simulations
* Building the model and get the performance on 200K test data
* Evaluating the model(get the performance on 2.3M dataset)
* Preparing the precision-recall pairs

To reproduce the experiment, for the above four steps, turn to kgxsapp101 and open the folder above and checkout to the branch “(JieGetPRPairsForC.2\_May27)” and open the following R file and click run.

(./scripts/main\_baggingRF.R)

* + - New scoring
* get the prediction value for the patients for new scoring

For the above step, turn to kgxsapp101 and open the folder below and checkout to the branch “(JieAddRecallBucketForNewScoringData)” open the following R file and click run.

(./scripts/main\_baggingRF.R)

* + - Calculating new buckets(one half of the scoring data)
* calculating new recall buckets using 2.3M dataset
* add recall bucket for each new scoring patients

For the above two steps, please checkout to the branch “JieAddRecallBucketForNewScoringData” and

* Please open the function “.\main\_otherFuns.R”
* Please select codes lines from line118 to line153 and click run
* After about half an hour, the program will complete
* Then please go to the folder “F:\Jie\Shire\_follow\_up\02\_Code\GoFromBaggingForest(HAECentral)\Results\nonhae\_200K\_v2(A1E1)&hae\_ptid\_973\2016-05-28 02.02.28\iters=20” to find the results.
* “recallBkt\_rangeCutoff\_allSimu.RDS”: table from the label-precision pairs of C.2\_May27
* “ptid\_resp\_pred.RDS”: table from the patient\_id, label, prediction of new scoring data after removing HAE patients and aggregate the prediction across all the 5 simulations
* “ptid\_recallBkt.RDS”: table of patient\_id, score, recall bucket for all the patient id from 3.ii below.
* “ptid\_recallBkt\_target.RDS”: table from iii, but only patients with recall bucket in [0, 30%].
* Then you can compare the result with that I did before in folder “. \AddBucketForNewScoringJie”
* Calculating new buckets(another half of the scoring data)

Please refer to the following folder in kgxsapp100 for the corresponding results

* F:\Jie\Shire\_follow\_up\02\_Code\GoFromBaggingForest\Results\nonhae\_200K\_v2(A1E1)&hae\_ptid\_973\2016-05-27 17.19.28\iters=20
  + - PR curves (Lichao)
  1. Results (all the delivery results have been collected together on kgxsapp101 and can be found in the folder: F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result)
     + The evaluation results on ~200K and ~2.3M(**please note that in the 973 patients there are 13 patients with only 2 predictors being positive**)
       1. Predictions
* ~200K:

“F:\Jie\Shire\_follow\_up\02\_Code\GoFromBaggingForest(HAECentral)\Results\nonhae\_200K\_v2(A1E1)&hae\_ptid\_973\2016-05-28 02.02.28\iters=20\resp\_pred.RDS”

* ~2.3M(kgxsapp100)

F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\predScore\_2.3M\simu\*\tst\_rf\_prob\_haeTs&3M.RDS

* ~110M

F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\predScore\_110M

(Please note that when we calculated the prediction score of the 110M patients, in order to speed up the running time, we split the data into two parts, one part was predicted on kgxsapp101 and the other part on kgxsapp100. That is the reason why there are two folders here.)

* + - 1. precision-recall pairs(kgxsapp101)

F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\PRCurve

* + - 1. PR curves (Lichao)
      2. Recall bucket of ~110M dataset

F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\recallBucket\_110M

* + - 1. Recall bucket (0 – 30%) of ~110M dataset

F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\recallBucketTarget\_110M

* + - 1. Important score

F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\importantScore

* + - How were the evaluation and predicting score done for ~200K and ~2.3M
* ~200K:

1. 5 models are returned with each model corresponding to an evaluation/simulation fold, and each of the 5 fold includes a number of forest (depending on the value of iters, for C.2, it should be 20.)
2. For each evaluation/simulation fold, apply the corresponding model to the corresponding test data (i.e. the left fold), that is to say 20% of the positive and 20% of the negative. With the predictions on 20% of the positives and 20% of the negatives, the prediction value based on each model (i.e. each simulation) can be obtained.
3. Aggregate the predicted score of all 5 evaluation/simulation fold together
4. Then get the performance metrics from the aggregated predictions. This performance shows the performance of models obtained from that particular experiment.
   * + - * ~2.3M:

In order to avoid the ppv-value-change introduced by the different ratios of positives vs negatives, we calculate the performance for 2.3M as the following steps:

1. For each experiment reported in the table above, 5 models are returned with each model corresponding to an evaluation / simulation fold, and each of the 5 model includes a number of forests (depending on the value of iters, for C.2, it should be 20.)
2. Apply each of the 5 models to the whole ~2.3M negatives. Note that the same model will also be applied to 20% of the positives. With the predictions on 20% of the positives and ~2.3M of the negatives, the prediction value based on each model (i.e. each simulation) can be obtained.
3. Aggregate the prediction values from the 5 models by patient id.
4. Then get the performance metrics from the aggregated predictions. This performance shows the performance of models obtained from that particular experiment.
   * + How was the predicting score done for the new scoring step?

* 5 models are returned with each model corresponding to each evaluation / simulation fold, and each of the 5 model includes a number of forests (depending on the value of iters, for C.2, it should be 20.)
* Apply each of the 5 models to the whole ~110M negatives. Then the prediction value based on each model (i.e. each simulation/evaluation) can be obtained (i.e. average the 20 predicted scores from 20 models as the predicted score of the very model).
* Aggregate the prediction values from the 5 models by patient id. Then get the performance metrics from the aggregated predictions.
  + - The extracted patients and recall buckets for recall = [0, 30%]
* Use the 5 models from Experiment c.2 to score all patients of the ~110M 5 times to get 5 scores for each patient;
* Average the 5 scores for each patient to get one prediction score (tb1);
* Take each of the unique predicted score as the threshold, the corresponding recall value can be obtained.
* For each of our targeting recall buckets, [0, 5%], (5%, 10%], (10%, 15%], (15%, 20%], (20%, 25%], (25%, 30%], it is easy to obtained the following columns, the minimum of predicted score at that range, the maximum of predicted score (tb2).
* Please note that the maximum value of predicted score should be revised using the minimum value of predicted score of the previous recall bucket.
* For each patient in tb1, to find out which recall bucket his/her predicted score falling into. (i.e. left join the two tables tb1, tb2 on tb2.maximum < tb1.score <= tb2.minimum)

1. Delivery in Apr.

Please note that there is a difference between the 973 HAE dataset for delivery in Apr and that for delivery in Jun.

* + - Delivery in Apr:

The variables, ER\_FLAG, ER\_FREQ, ER\_AFREQ are based on Dx and Rx data

* + - Delivery in Jun:

The variables, ER\_FLAG, ER\_FREQ, ER\_AFREQ are based on Dx only data.

1. Others in the follow-up analysis
   1. The current status of the follow-up analysis (Lichao)
      * Progress
      * Conclusions
   2. Other datasets
      * The ~165M, ~123M, ~95M and 1233 hae patients and the corresponding criteria

* ~165M:

1. lookback>=24 months
2. RX\_FLAG=1
3. DX\_FLAG=1

* ~123M:
  + - * 1. Lookback>=24 months
        2. RX\_FLAG=1
        3. DX\_FLAG=1
        4. Lookback\_days between 662 and 2051(+-30 days of 1233 HAE patients)
        5. Patient age > 12
        6. Patient gender <> ‘U’
        7. Event > 0 (with at least one active predictor)
* ~95M:
  + - * 1. Lookback>=24 months
        2. RX\_FLAG=1
        3. DX\_FLAG=1
        4. Lookback\_days between 662 and 2051(+-30 days of 1233 HAE patients)
        5. Patient age > 12
        6. Patient gender <> ‘U’
        7. event >= 3 (with at least three active predictors)
* 1233 HAE dataset:

It is Dong who extracted the 1233 HAE dataset and Zhenxing cannot reproduce the 1233 HAE dataset. So maybe we should turn to Dong for help

* 1. Other code
     + similarity modelling, cleaning, ModelStack, pipeline and those implemented by Jie etc. with their status (e.g., whether or not QCed)
* *Similarity modelling (kgxsapp101)(QCed by Jie and Zhiyu)*:

Script - F:\Jie\Shire\_follow\_up\02\_Code\Similarity\main.R

Functions - F:\Jie\Shire\_follow\_up\02\_Code\Similarity\funs\_similarity.R

* *Cleaning + feature selection(QCed by Jie and Zhiyu):*

Script - F:\Jie\Shire\_follow\_up\02\_Code\main.R

Functions - F:\Jie\Shire\_follow\_up\02\_Code\clean\_split\_featureSelection\_lasso.R

* *Get coefficient retained by lasso(QCed by Jie):*

F:\Jie\Shire\_follow\_up\02\_Code\get\_coef\_retained\_lasso\ get\_covar\_left\_lasso.R

* *Add easy positive element(without debug and without QC):*

F:\Jie\Shire\_follow\_up\02\_Code\Similarity\ funs\_similarity\_add\_easy\_positive.R

* + - Variable importance calculation (for RF and aggregating / normalizing the gini indices, as delivered in Apr.)(QCed by Jie)
* F:\Jie\Shire\_follow\_up\02\_Code\HAE\_R\_codes\_Dec15\ get\_importance\_score.R
  1. The method document sent by John recently (Lichao)