Document for Shire HAE follow-up analysis project

1. Final delivery of new scoring (Jun 2016)
   1. Data
      1. The updated 973-patients (ER is extracted only using Dx information)
         1. Location: Kgxsapp100: “F:\Lichao\work\Projects\HAE\Documents\Documentation\Final Delivery of New Scoring\Data\973”
         2. The criteria for extracting the datasets from Dong’s 1233 HAE patients (defined in Section 3.2.4 below):
            1. Remove the patients with \_FLAG=1
            2. Remove the patients with AGE <= 12
            3. Remove the patients with EVENT < 3
      2. The ~200K for training the model;
         1. Location: F:\Lichao\work\Projects\HAE\Documents\Documentation\Final Delivery of New Scoring\Data\200k
         2. 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 below 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.).
      3. The ~2.3M for evaluating the model and setting the thresholds
         1. Location(the split 10 files): F:\Lichao\work\Projects\HAE\Documents\Documentation\Final Delivery of New Scoring\Data\2.3M
         2. 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 subjected to 1.1.3.2.1 above.
      4. The ~110M new scoring data
         1. Location (kgxsapp101):
            1. F:\Jie\Shire\_follow\_up\01\_data\newdata\_200K\_3M\ split\_newScoreMay27\_1to286\\*.csv
            2. F:\Jie\Shire\_follow\_up\01\_data\newdata\_200K\_3M\ split\_newScoreMay27\_ 287to552\\*.csv
         2. 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)
   2. Code
      1. Location: F:\Jie\Shire\_follow\_up\02\_Code\GoFromBaggingForest(HAECentral)
      2. Reproduce the result of C.2\_May27 which is saved in the following folder. To reproduce the experiment, turn to kgxsapp101, open the folder above, checkout to the branch “(JieGetPRPairsForC.2\_May27)” and open (./scripts/main\_baggingRF.R) and click run. This includes several steps:
         1. Splitting the data into simulations
         2. Building the model and get the performance on 200K test data
         3. Evaluating the model (get the performance on 2.3M dataset)
         4. Preparing the precision-recall pairs
      3. New scoring: get the prediction value for the patients for new scoring
         1. Turn to kgxsapp101, open the folder below, checkout to the branch “(JieAddRecallBucketForNewScoringData)” open (./scripts/main\_baggingRF.R) and click run.
      4. Calculating new buckets(one half of the scoring data)
         1. It does the following
            1. Calculating new recall buckets using 2.3M dataset
            2. Add the recall bucket for each new scoring patient
         2. For the above two steps, please checkout to the branch “JieAddRecallBucketForNewScoringData” and do the following
            1. Please open the function “.\main\_otherFuns.R”
            2. Please select codes lines from line118 to line153 and click run
            3. 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.
            4. “recallBkt\_rangeCutoff\_allSimu.RDS”: table from the label-precision pairs of the reproduced result
            5. “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
            6. “ptid\_recallBkt.RDS”: table of patient\_id, score, recall bucket for all the patient id from ii.
            7. “ptid\_recallBkt\_target.RDS”: table from iii, but only patients with recall bucket in [0, 30%].
            8. Then you can compare the result with that I did before in folder “. \AddBucketForNewScoringJie”
      5. Calculating new buckets(another half of the scoring data): Please refer to the following folder in kgxsapp100 for the corresponding results
         1. F:\Jie\Shire\_follow\_up\02\_Code\GoFromBaggingForest\Results\nonhae\_200K\_v2(A1E1)&hae\_ptid\_973\2016-05-27 17.19.28\iters=20
      6. PR curves (Lichao to add)
   3. 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)
      1. 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
            1. ~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. ~2.3M(kgxsapp100) F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\predScore\_2.3M\simu\*\tst\_rf\_prob\_haeTs&3M.RDS
            3. ~110M F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\predScore\_110M
         2. Precision-recall pairs F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\PRCurve
         3. PR curves (Lichao)
         4. Recall bucket of ~110M dataset F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\recallBucket\_110M
         5. Recall bucket (0 – 30%) of ~110M dataset F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\recallBucketTarget\_110M
         6. Important score F:\Jie\Shire\_follow\_up\Delivery\_Jun\02\_Result\importantScore
      2. How were the evaluation and predicting score done for ~200K and ~2.3M
         1. ~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-out fold), that is 20% of the positive and 20% of the negative.
            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. ~2.3M:
            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 models 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. This gives a positive to negative ratio a little less than 1:10,000. 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.
      3. How was the predicting score done for the new scoring step?
         1. 5 models are returned with each model corresponding to each evaluation / simulation fold, and each of the 5 models 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 ~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).
         3. Aggregate the prediction values from the 5 models by patient id. Then get the performance metrics from the aggregated predictions.
      4. The extracted patients and recall buckets for recall = [0, 30%]
         1. Use the 5 models from Experiment c.2 to score all patients of the ~110M 5 times to get 5 scores for each patient;
         2. Average the 5 scores for each patient to get one prediction score (tb1);
         3. Take each of the unique predicted score as the threshold, the corresponding recall value can be obtained.
         4. For each of our targeting recall buckets, [0, 5%], (5%, 10%], (10%, 15%], (15%, 20%], (20%, 25%], (25%, 30%], obtain the following columns, the minimum of predicted score at that range, the maximum of predicted score (tb2).
         5. Please note that the maximum value of predicted score should be replaced by the minimum value of predicted score of the previous recall bucket.
         6. For each patient in the ~110M, to find out which recall bucket his/her predicted score falls into.
2. Delivery in Apr. 2016
   1. The similarity and difference between the 973 HAE datasets for delivery in Apr and that in Jun.
      1. They are the same 973 patients
      2. Delivery in Apr: The variables, ER\_FLAG, ER\_FREQ, ER\_AFREQ are based on both Dx and Rx data
      3. Delivery in Jun: The variables, ER\_FLAG, ER\_FREQ, ER\_AFREQ are based only on Dx data. This leads 12 patients with less than 3 positive events.
3. Others in the follow-up analysis
   1. The current status of the follow-up analysis (Lichao)
      1. Progress
      2. Conclusions
   2. Other datasets
      1. ~165M:
         1. The patient has at least one Rx event and one Dx event during the lookback period (the lookback period for everyone inside the ~165M is defined as follows):
            1. The lookback period for everyone inside the ~165M is defined as from the look back date to the index date;
            2. The look back date is defined as the later date between the Rx\_lookback\_dt and the Dx\_lookback\_dt;
            3. The index date is defined as the latest Rx or Dx service date during 2010/1 - 2015/9;
            4. The Rx\_lookback\_dt is defined as the earliest Rx service date during 2010/1 - 2015/9;
            5. The Dx\_lookback\_dt is defined as the earliest Dx service date during 2010/1 - 2015/9.
         2. The lookback period is at least 24 months.
      2. ~123M:
         1. A subset from the ~165M above
         2. Lookback\_days between 662 and 2051(+-30 days of 1233 HAE patients)
         3. Patient age > 12
         4. Patient gender <> ‘U’
         5. Event > 0 (with at least one active predictor)
         6. No patients in the ~123M belongs to the 973 confirmed HAE cohort
      3. ~95M:
         1. A subset of the ~123M above
         2. event >= 3 (with at least three active predictors)
      4. 1233 HAE dataset:
         1. The patient satisfies at least one of the following three criteria within the period of 1/1/2012 – 7/31/2015:
            1. Shire’s OnePath® program (see Appendix 1);
            2. Patients with 277.6 ICD-9 and at least one HAE Treatment Rx/procedure (HAE Treatment Rx is defined in Appendix 2, and HAE Treatment procedure is defined in Appendix 3);
            3. Patients with at least one HAE Treatment Rx/procedure for a product that is not expected to be used for any other condition (Berinert, Cinryze, Kalbitor, Ruconest or Winstrol).
         2. And the patient has not had anynot only had TESTOSTERONE products;
         3. And the patient’s lookback period is greater than or equal to 24 months (the lookback period for confirmed HAE patients is defined below)
            1. The lookback period for a confirmed HAE patient is defined as from the look back date to the day before the index date;
            2. The index date is defined as the definite first exposure date if definite first exposure date is available, otherwise it is defined as the first exposure date;
            3. The look back date is defined as the earliest Dx and/or Rx date from 2010/1 to the day before index date;
            4. The definite first exposure date is defined as the earliest date of HAE diagnosis (277.6) OR HAE treatments Rx/Procedure (excluding TESTOSTERONE products) within selection window of 2012/1-2015/7;
            5. The first exposure date is defined as the earliest date of HAE diagnosis (277.6 or 995.1) or HAE treatment (Rx or procedure) within selection window of 2012/1-2015/7;
         4. The patient is naïve; being naïve is defined as
            1. If the definite first exposure date is available and the definite earliest exposure date is null, OR the earliest exposure date is null;
            2. The definite earliest exposure date is defined as the earliest date of HAE diagnosis (277.6) OR HAE treatments Rx/Procedure (excluding TESTOSTERONE products) within selection window of 2010/1 to 2011/12;
            3. The earliest exposure date is defined as the earliest date of HAE diagnosis (277.6 or 995.1) or HAE treatment (Rx or procedure) within selection window of 2010/1 to 2011/12.
4. Other code
   1. Similarity modelling (kgxsapp101) (QCed by Jie and Zhiyu):
      1. Script - F:\Jie\Shire\_follow\_up\02\_Code\Similarity\main.R
      2. Functions - F:\Jie\Shire\_follow\_up\02\_Code\Similarity\funs\_similarity.R
   2. Cleaning + feature selection(QCed by Jie and Zhiyu):
      1. Script - F:\Jie\Shire\_follow\_up\02\_Code\main.R
      2. Functions - F:\Jie\Shire\_follow\_up\02\_Code\clean\_split\_featureSelection\_lasso.R
      3. Get coefficient retained by lasso(QCed by Jie): F:\Jie\Shire\_follow\_up\02\_Code\get\_coef\_retained\_lasso\ get\_covar\_left\_lasso.R
      4. Add easy positive element(without debug and without QC): F:\Jie\Shire\_follow\_up\02\_Code\Similarity\ funs\_similarity\_add\_easy\_positive.R
      5. 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
5. The method document sent by John recently (Lichao)

**Appendix 1.** Shire’s OnePath program

* 2,832 Shire OnePath patients were sent to Experian for de-identification
* 2,831 were successfully de-identified
* 2,718 match to IMS database



**Appendix 2.** HAE treatment Rx



**Appendix 3.** HAE treatment procedures

