BI IPF Modelling Document

Contents

[Data Preparation and QC 2](#_Toc461115409)

[Project summary folder 2](#_Toc461115410)

[Cohorts 2](#_Toc461115411)

[Code for reading in data 3](#_Toc461115412)

[Variables 3](#_Toc461115413)

[Demographics 3](#_Toc461115414)

[Event flags and frequencies 3](#_Toc461115415)

[Code for descriptive statistics 5](#_Toc461115416)

[Interaction terms 6](#_Toc461115417)

[Code for interaction term 6](#_Toc461115418)

[Sequence variables 6](#_Toc461115419)

[Code for sequence variable 7](#_Toc461115420)

[Exclusion of Variables 7](#_Toc461115421)

[Excluded variable lists 8](#_Toc461115422)

[Removal of extreme values 8](#_Toc461115423)

[Code for capping extreme values 8](#_Toc461115424)

[Predictive Modelling 9](#_Toc461115425)

[Methods 9](#_Toc461115426)

[Logistic regression 9](#_Toc461115427)

[Random forests 9](#_Toc461115428)

[Model Implementation 10](#_Toc461115429)

[Single-stage modeling 10](#_Toc461115430)

[Code for one-stage modelling 10](#_Toc461115431)

[Two-stage modeling 11](#_Toc461115432)

[Code for two-stage modelling 12](#_Toc461115433)

[Application of the Two-stage Model to the Scoring Sample 14](#_Toc461115434)

[Code for applying two-stage model on scoring sample 15](#_Toc461115435)

[Code for extract subset patients from original scoring sample 16](#_Toc461115436)

[Performance Metrics 16](#_Toc461115437)

[Model Interpretation 16](#_Toc461115438)

[Binary variables 16](#_Toc461115439)

[Age 16](#_Toc461115440)

[Frequency variables 16](#_Toc461115441)

[Appendix 1 17](#_Toc461115442)

[Appendix 2 19](#_Toc461115443)

# Data Preparation and QC

The data were sourced from the IMS LRx and Dx databases. Briefly, the LRx database contains information about prescription claims data and the Dx database contains information on medical claims as well as patient demographics. Patients contained in both databases can be linked via encrypted identifiers.

# Project summary folder

The project summary folder is “F:\Hui\Project\_2016\BI\_IPF\_2016\04\_Summary” on server 101(kgxsapp101). It will be referred as “.\” in the following. There are 3 subfolders in the summary folder, which are:

1. *001\_Reference\_file*: contains the reference file, such as excluded variable list, included variable list, sampling strategy and sequence variable mapping file
2. *002\_Code*: contains all the codes in BI project
3. *003\_Results*: contains the model results.

# Cohorts

* IPF cohort (positive patients; N=8574).
  + Treated for IPF with either Esbriet or OFEV
  + Have Asthma or COPD (as evidenced by diagnosis or treatment claim to first IPF exposure)
  + First exposure in or after Oct 2014
  + 24+ months of available medical history
  + Greater than 40 years old
  + Patients should exist in both Dx and LRx

The IPF cohort is restricted to patients that received treatment as this provides a conservation approach for patient selection for modelling. However, note that there are several reasons why an IPF patient might not have received treatment such as (i) being newly diagnosed, (ii) cost of treatment, (iii) free trials that are not captured in the data source and (iv) the need to balance side effects with treatment effectiveness. Therefore, when estimating the prevalence of IPF in the data source we then remove the criteria that a patient needs to have received a prescription for either OFEV or Esbriet. This results in a total of 30,074 IPF patients with a diagnosis of Asthma or COPD over the age of 40 with at least 24 months of patient history with IPF exposure from Oct 2014 onwards. There are 25,664,173 Asthma or COPD patients who are also over the age of 40 and with at least 24 months of patient history with first exposure for Asthma or COPD after Oct 2014. Therefore the representative ratio of IPF to Asthma/COPD is 1 to 853.

* Non-IPF cohort (853 negative patients are selected to match each patient in the IPF cohort; 200 matched patients will be used for training and the full 853 patients will be used for testing)
  + Not part of the IPF cohort
  + Have Asthma or COPD (as evidenced by diagnosis (1 claim) or treatment(at least 2 Rxs))
  + Greater than 40 years old
  + Similar lookback period and Rx/Dx activity to an IPF patient.
  + Patients should exist in both Dx and LRx

Code for reading in data

The pre-processing code is stored in “.\002\_Code\SAS\_code\01 read raw data and subset.sas”. The code would use the following input files:

1. Included variable list: “.\001\_Reference\_file\newpredictorList1.csv”
2. IPF cohort data: “F:\Hui\Project\_2016\BI\_IPF\_2016\02\_data\IPF\_Cohort.csv\IPF\_Cohort.csv”
3. Non-IPF Asthma/COPD cohort data (1:200): “F:\Hui\Project\_2016\BI\_IPF\_2016\02\_data\Non\_IPF\_Asthma\_COPD\_Cohort.csv\Non\_IPF\_Asthma\_COPD\_Cohort.csv”
4. Representative sample (1:653): “F:\Hui\Project\_2016\BI\_IPF\_2016\02\_data\Representative\_Sample\_Asthma\_COPD\_Cohort.csv\Representative\_Sample\_Asthma\_COPD\_Cohort.csv”
5. Scoring sample: “F:\Hui\Project\_2016\BI\_IPF\_2016\02\_data\Scoring\_Sample\_V2.csv\Scoring\_Sample\_V2.csv”

The outputs of the code are stored in “F:\Hui\Project\_2016\BI\_IPF\_2016\02\_data\SAS\_dataset”:

1. “*ipf\_columns\_new.sas7bdat*”: included variable list
2. “*ipf\_cohort\_subset.sas7bdat*”: original IPF cohort data
3. *“nonipf\_ac\_cohort\_subset.sas7bdat”*: original Non-IPF Asthma/COPD data
4. *“score\_sample\_cohort.sas7bdat”*: original representative sample
5. *“score\_sample\_26m.sas7bdat”*: original scoring sample

# Variables

The following section describes the variables that were included in modelling.

Demographics

* Age
* Gender (converted to 0 for female and 1 for male)

Event flags and frequencies

Variables capturing symptoms, procedures, co-morbidities, treatments and risk factors that are relevant to IPF will be quantified using a binary flag to indicate at least one event in the lookback period (var\_name\_FLAG) and a frequency variable computed as the number of events in the lookback period divided by the length of the lookback period (var\_name\_AVG\_RXDX). A total of 69 events were considered and these are listed by category below:

* Symptom events
  + ABNORMAL CHEST SOUNDS
  + ABNORMAL CHEST XRAY
  + ABNORMAL RESPIRATORY FUNCTION
  + ABNORMAL STRESS TEST
  + ANOREXIA/LOSS OF APPETITE
  + CHEST PAIN
  + CLUBBING OF FINGERS
  + COUGH
  + DECREASED EXERCISE TOLERANCE
  + MALAISE/FATIGUE
  + MENTAL CONFUSION
  + POSTNASAL DRIP
  + SOB/DYSPNEA
  + TACHYCARDIA
  + TREMORS
  + WEIGHT LOSS
* Procedures events
  + 6 MINUTE WALK
  + BLOOD GAS EVALUATIONS
  + BRONCHOALVEOLAR LAVAGE
  + CHEST ULTRASOUND
  + CHEST X-RAY
  + CORONARY ANGIOGRAM
  + CORONARY BYPASS
  + GASTROSCOPY
  + LOW DOSE CT SCAN
  + OTHER PULMONARY FUNCTION TESTING
  + OXYGEN EVALUATION
  + OXYGEN SATURATION TESTING
  + THYROID ULTRASOUND
* Diagnosis of respiratory co-morbidities event
  + ACUTE RESPIRATORY INFECTIONS
  + HISTORY OF RESPIRATORY DISEASE
  + HYPERSENSITIVITY PNEUMONITIS
  + HYPOXEMIA
  + INFLUENZA
  + OBSTRUCTIVE SLEEP APNEA
  + OTHER LUNG INFECTIONS
  + PNEUMONIA
  + RESPIRATORY ARREST
  + RESPIRATORY FAILURE
  + TUBERCULOSIS
* Diagnosis of non-respiratory co-morbidities event
  + ACS/MI
  + ANXIETY/DEPRESSION
  + CHRONIC ISCHMIC HEART DISEASE
  + CONGESTIVE HEART FAILURE
  + CROHNS DISEASE
  + DIABETES
  + DVT
  + GERD
  + HYPERLIPIDEMIA
  + HYPERTENSION
  + HYPERTHYROIDISM
  + HYPOTHYROIDISM
  + LEG EDEMA
  + LUNG CANCER
  + OSTEOARTHRITIS
  + PSORIATIC ARTHRITIS
  + PULMONARY HEART DISEASE
  + STROKE
* Treatment events

|  |
| --- |
| * + ANTIARTHRITICS Dx |
| * + ANTIARTHRITICS LRx |
| * + ANTIBIOTICS Dx |
| * + ANTIBIOTICS LRx |
| * + ANTICOAGULANTS Dx |
| * + ANTICOAGULANTS LRx |
| * + ANTIDIABETICS LRx |
| * + ANTIHYPERLIPIDEMICS LRx |
| * + BREATHING EXERCISES |
| * + CARDIOVASCULAR AGENTS Dx |
| * + CARDIOVASCULAR AGENTS LRx |
| * + COUGH SUPPRESSANTS LRx |
| * + CROHNS Dx |
| * + CROHNS TREATMENT LRx |
| * + GERD Dx |
| * + GERD TREATMENT LRx |
| * + LUNG CA ANTINEOPLASTICS Dx |
| * + LUNG CA ANTINEOPLASTICS LRx |
| * + OXYGEN THERAPY |
| * + OXYGEN USE |
| * + PREDNISONE LRx |
| * + PULMONARY REHABILITATION |
| * + THYROID AGENTS LRx |

* Risk factors
  + ASBESTOS EXPOSURE
  + FAMILY HISTORY RESP DISEASE
  + RADIATION EXPOSURE
  + SMOKER

Four additional count variables were created that summarise the number of events across four of these categories: a symptom count variable (SYMP\_CNT), a procedure count variable (PROC\_CNT), a respiratory co-morbidity count (COMOR\_CNT) and a non-respiratory comorbidity count (NRCOMOR\_CNT).

#### Code for descriptive statistics

The descriptive statistics code is “.\002\_Code\SAS\_code\02\_descriptive\_statistics.sas”, the inputs data for the code is:

1. “*ipf\_columns\_new.sas7bdat*”: included variable list
2. “*ipf\_cohort\_subset.sas7bdat*”: original IPF cohort data
3. *“nonipf\_ac\_cohort\_subset.sas7bdat”*: original Non-IPF Asthma/COPD data
4. *“score\_sample\_cohort.sas7bdat”*: original representative sample
5. *“score\_sample\_26m.sas7bdat”*: original scoring sample

The descriptive statistics were computed considering all observations (with imputation) and considering only observations that have a positive flag for the event (without imputation). The following labels were used to describe the cohorts:

1. ‘ipf\_cohort’ – the 8574 positive IPF patients
2. ‘nonipf\_ac\_cohort’ – the 200\*8574 negative IPF patients with Asthma/COPD activity; each patient in the ipf cohort has 200 negative matches.
3. ‘ss\_cohort’ – the 653\*8574 negative IPF patients with Asthma/COPD activity; each patient in the ipf cohort has 653 negative matches.
4. ‘ss\_26m\_cohort’ – the scoring sample of 26m negative ipf patients with Asthma/COPD activity.

The descriptive statistics are stored in.\04\_Summary\003\_Results\Descriptive\_stats\_results:

1. *“Des\_ipf\_cohort\_newpred\_with\_imputation.csv”*
2. *“Des\_ipf\_cohort\_newpred\_without\_imputation.csv”*
3. *“Des\_nonipf\_ac\_cohort\_newpred\_with\_imputation.csv”*
4. *“Des\_ nonipf\_ac \_cohort\_newpred\_without\_imputation.csv”*
5. *“Des\_ss\_cohort\_newpred\_with\_imputation.csv”*
6. *“Des\_ss\_cohort\_newpred\_without\_imputation.csv”*
7. *“Des\_ss\_26m\_cohort\_newpred\_with\_imputation.csv”*
8. *“Des\_ss\_26m\_cohort\_newpred\_without\_imputation.csv”*

Interaction terms

A decision tree was used to identify potential interactions terms to aid visualisation. The tree was built using the ‘party’ package in R (Hothorn & Zeileis 2013). This implementation is a recursive partitioning method that is designed to be less prone to overfitting (Hothorn et al. 2006). The tree was built using the top ten rated variables with a max tree depth of 5 and based on one third of the data. The output of the tree is displayed in Appendix 1. The tree revealed two potentially useful interaction terms (i) the frequency of other pulmonary function testing and the frequency of O2 therapy and (ii) the frequency of other pulmonary function testing and the frequency at which the 6 minute walk procedure was administered. Both of these interaction terms were presented as binary variables. First, the frequency variables were dichotomised using the node split thresholds from the decision tree and the product of the pair of dichotomised variables were used to represent the interaction term.

Sequence variables

Variables quantifying the sequence in which events occur were developed to investigate the potential of taking the order of events into account. The final dataset contains information on 69 events which results in 2346 unique pairs using the formula (N\*(N-1)/2) where N is 69. For each of these unique pairs, patient counts for the positive and negative cohorts were developed using a sub-sample of the data (1607 positive patients and 321,400 negative patients). *Figure 1* illustrates how the events pairs were reduced to 28 pairs of events that were subsequently included in the final model scheme. Please see the Appendix 2 for the list of sequence variables selected for modelling.

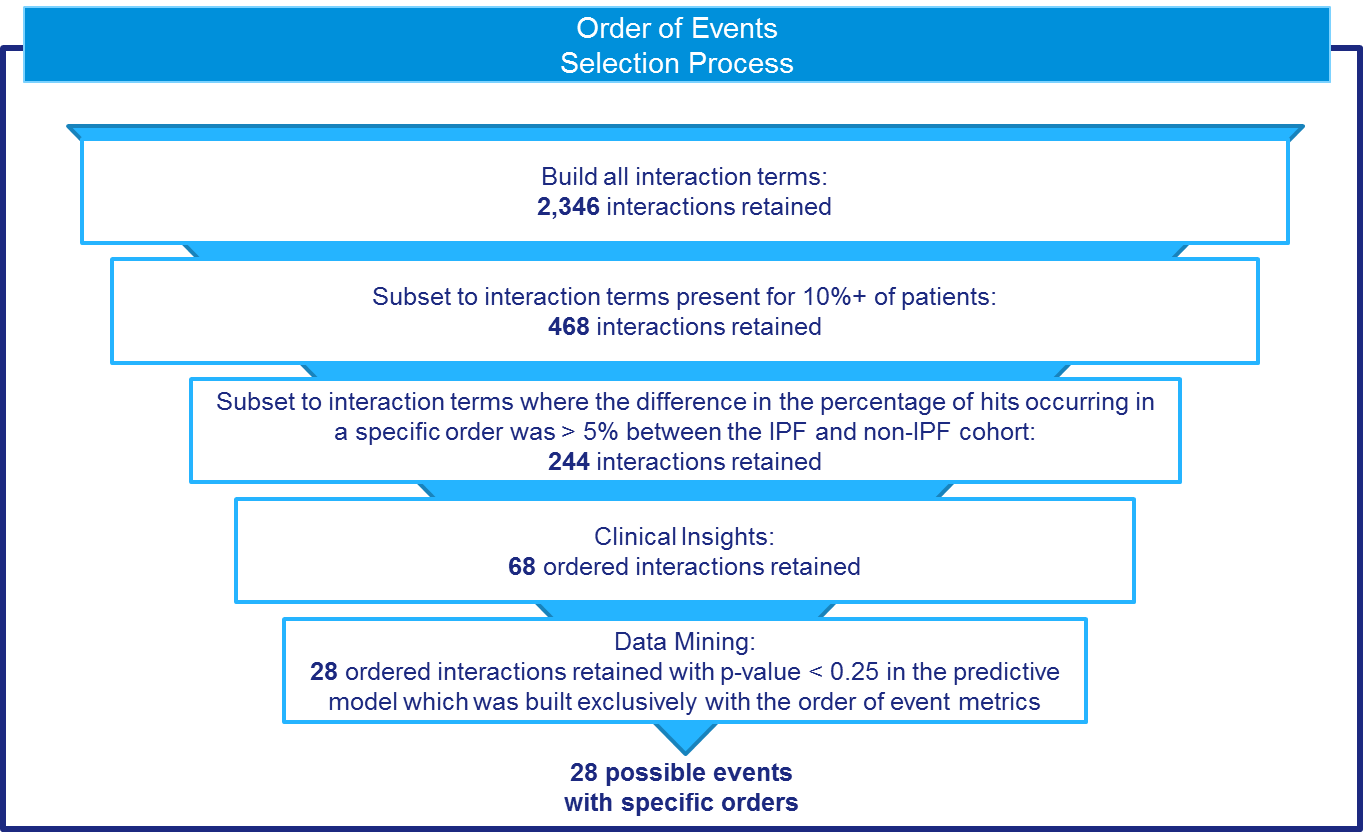


Figure 1 Criteria for selected event pairs to include in the final model

#### Code for interaction terms and temporal sequence variables

The code for the interaction terms is stored in “./002\_Code\Code from Zi”.

1. ‘event\_sequence\_0829\_28pairs\_26m scoring sample(NEW METHOD).sas’ - used to create temporal sequence for 26 million scoring sample.
2. ‘interaction term\_0829.sas’ - Code to create interaction terms for ipf\_cohort, nonipf\_ac\_cohort, ss\_cohort and ss\_26m\_cohort scoring samples.
3. ‘merge all 175 features.r’ – R code used to merge raw data (after extreme values replaced), temporal sequence and interaction terms, for for ipf\_cohort, nonipf\_ac\_cohort, ss\_cohort and ss\_26m\_cohort scoring samples.
4. ‘merge\_all\_features\_0829.sas’ – Code to merge all 218 features for 26 million scoring sample, including raw data (after extreme values replaced), temporal sequence and interaction terms.

# Exclusion of Variables

The following 16 variables were removed as positive flags for these patients were found in no more than 50 patients (<0.6%). Given the low frequency of events these variables will be removed as they are not likely to well-tolerated by models such as logistic regression.

* LVL3\_CROHNS\_DX (CROHNS Dx)
* LVL3\_BRTHNG\_EX (BREATHING EXERCISES)
* LVL3\_LNG\_CA\_AO\_LRX (LUNG CA ANTINEOPLASTICS LRx)
* LVL3\_LD\_CT\_SCAN (LOW DOSE CT SCAN)
* LVL3\_F\_H\_RESP\_DIS (FAMILY HISTORY RESP DISEASE)
* LVL3\_DSC\_EXER\_TOL (DECREASED EXERCISE TOLERANCE)
* LVL3\_RESP\_ARRST (RESPIRATORY ARREST)
* LVL3\_GERD\_DX (GERD)
* LVL3\_RAD\_EXP (RADIATION EXPOSURE)
* LVL3\_CLUB\_FNGRS (CLUBBING OF FINGERS)
* LVL3\_ANTI\_COAG\_DX (ANTICOAGULANTS Dx)
* LVL3\_PSOR\_ARTH (PSORIATIC ARTHRITIS)
* LVL3\_CHEST\_ULTRA (CHEST ULTRASOUND)
* LVL3\_TUBERCUL (TUBERCULOSIS)
* LVL3\_OTHR\_LUNG\_INF (OTHER LUNG INFECTIONS)
* LVL3\_O2\_EVAL (OXYGEN EVALUATION)

Variables pairs that have a correlation coefficient of greater than 0.9 will be identified and one of the variables will be removed. In the case of a flag and frequency variable pair having a high level of correlation then the frequency variable will be excluded.

### Excluded variable lists

The list is “.\001\_Reference\_file\vars\_to\_exclude\_list1.csv” which contains the variables listed above and four variables related to Asthma/COPD. Additionaly, the variable “LVL3\_GASTROSC\_AVG\_RXDX” was removed as it is highly correlated with its equivalent flag variable.

When come to the two stage, there is an additional excluded variable list which is “.\001\_Reference\_file\vars\_to\_exclude\_list2.csv”. This accounts for additional checks of variables that have less than 50 observations in a cell.

## Removal of extreme values

For the variables that are count or continuous (e.g. the frequency variables that are computed as the count of events divided by the lookback) extreme values will be removed using the following method.

1. Compute the 99th percentiles of a variable for both the IPF and non-IPF cohorts. Note that this computation should not include data from patients have a count of zero for the particular variable. As an illustrative example, consider a variable that counts the frequency with which a medication was prescribed. In the disease cohort there are 100 patients, 55 of these patients have a non-zero frequency of prescriptions. These 55 patients are then used to calculate the 99th percentile of the frequencies.
2. The threshold for finding extreme values is the maximum of the 99th percentiles computed for each cohort. Values greater than this threshold are replaced by the threshold value.

### Code for capping extreme values

The capping code is “.\002\_Code\SAS\_code\04\_data replace extremes and DS.sas”, the inputs of the code are:

1. “*ipf\_columns\_new.sas7bdat*”: included variable list
2. *“Ipf\_cohort\_subset.sas7bdat”*: IPF cohort data
3. *“nonIpf\_ac\_cohort\_subset.sas7bdat”*: Non-IPF cohort data
4. *“score\_sample\_cohort\_subset.sas7bdat”*: representative sample
5. *“ss\_26m\_cohort\_subset.sas7bdat”*: scoring sample

The outputs contain model data and new descriptive statistics:

Model data are stored in “/fs219.3/hjin/BI\_IPF\_2016/02\_data/model\_data”, which are:

1. *“Ipf\_cohort\_replace\_extremes.csv*
2. *“non\_Ipf\_cohort\_replace\_extremes.csv”*
3. *“score\_sample\_model\_data\_cap\_ext.csv”*
4. *“score\_sample\_26m\_model\_data\_cap\_ext2.csv”*

The new descriptive statistics are stored in “/fs219.3/hjin/BI\_IPF\_2016/03\_result”, which are

1. *“Des\_ipf\_cohort\_newpred\_with\_imp\_cap\_ext.csv”*
2. *“Des\_ipf\_cohort\_newpred\_without\_ imp\_cap\_ext.csv”*
3. *“Des\_nonipf\_ac\_cohort\_newpred\_with\_ imp\_cap\_ext.csv”*
4. *“Des\_ nonipf\_ac \_cohort\_newpred\_without\_ imp\_cap\_ext.csv”*
5. *“Des\_ss\_cohort\_newpred\_with\_ imp\_cap\_ext.csv”*
6. *“Des\_ss\_cohort\_newpred\_without\_ imp\_cap\_ext.csv”*
7. *“Des\_ss\_26m\_cohort\_newpred\_with\_ imp\_cap\_ext.csv”*
8. *“Des\_ss\_26m\_cohort\_newpred\_without\_ imp\_cap\_ext.csv”*

# Predictive Modelling

## Methods

The performance of different predictive models will be investigated in this study. The benchmark model will be conventional logistic regression whereby the association between a set of independent variables a binary dependent variables is modelled.

Logistic regression

Logistic regression is a widely used tool in biostatistics for cases where the outcome variable (or label) is binary. The logistic regression framework models a transformation of the outcome variable, namely the log of odds of the outcome:

The terms represent the regression coefficients where represents the intercept and are associated with the variables included in the model. The probability or risk of a certain outcome is represented by and this is transformed into the log odds via the logit function:

.

In this study, logistic regression will be used to assess the association between the independent variables ( and the outcome variable () as well as computing the probability of belonging to the IPF cohort for independent test data.

Random forests

Predictive modeling also will be implemented using the random forest paradigm. The random forest is an ensemble approach whereby the outputs of multiple decision trees are combined to form a prediction about a test case. The central idea is that each decision tree represents a “weak learner” and their combination forms a “strong learner”. Additionally, combining a collection of trees will help alleviate overfitting.

Specifically, the individual decision trees comprising the forest are grown using the following approach:

1. Sample cases from the training set with replacement. The number of cases sampled is a hyperparameter of the model.
2. Select a subsample of the variables to take forward to training. If P is the number of variables then M variables will be selected where M < P. M is held constant across all trees and the value of M is a hyperparameter of the model.
3. The extent to which each tree grows can be controlled using a hyperparameter relating to the depth of the tree.
4. This process (steps 1 to 3) is repeated for each decision tree in the forest. The number of trees that form the random forest is a hyperparameter of the model.
5. To make predictions for a test case the predictions across all decision trees are aggregated to form a consensus prediction for the test case.

The hyperparameters listed above can be set to default values or alternatively some or all of these can be optimized in a data-driven fashion while preserving the independence of the test data. For this study, the number of trees and depth of the trees will be optimising during training.

For random forest the gini importance measure will be computed to assess variable importance.

# Model Implementation

The positive to negative ratio of IPF to non-IPF patients is 1 to 853. To reduce the computational complexity, a lower proportion of negative patients will be selected. The ratio was varied from 1 to 10, 1 to 50, 1 to 100 and 1 to 200 and using a subset of the data the 1 to 200 ratio produced superior performance and therefore this ratio will be taken forward as the training ratio. Therefore models will be trained on a positive to negative ratio of 1 to 200 and subsequently tested using a test set with a positive to negative ratio of 1 to 853.

Single-stage modeling

In this study, 5-fold cross validation will be used to estimate model performance. This will involve randomly splitting the positive patient cohort into 5 partitions. The negative patients will be subsequently partitioned into 5-folds stratified by the positive patient to which they are a match (in terms of lookback period). For example if a positive patient is assigned to the first fold then all negative patients who are matched to this positive patient will also be assigned to the first fold. For the training folds, 200 of these matched negative patients will be taken forward for training. For a testing fold, all 853 matched negative patients will be taken forward for testing.

Logistic regression and random forests were investigated within the single stage modelling paradigm. Logistic regression was found to outperform random forests. As logistic regression can be prone to overfitting, the predictions on the training set and test set were assessed. A close agreement between the performance on the training and test data was observed indicating that overfitting is not an issue.

#### Code for one-stage modelling

* Logistic regression:
  + **Code**: “.\002\_Code\Spark\_code\Conventional\_LR\03\_Standard\_LR\_CV\_withpatid.py”
  + **Data files** are in S3 bucket “s3://emr-rwes-pa-spark-dev-datastore/BI\_IPF\_2016/01\_data/”:
    - *"Ipf\_cohort\_replace\_extremes.csv"*: IPF cohort
    - *"non\_Ipf\_cohort\_replace\_extremes.csv"*: Non-IPF cohort
    - *'score\_sample\_model\_data\_cap\_ext.csv'*: Representative sample
  + **Excluded variable list**: “/home/hjin/BI\_IPF\_2016/01\_data/vars\_to\_exclude\_list1.csv” on master node
  + **Sample strategy**: “.\001\_Reference\_file\ Sample\_strategy\_Conventional\_LR.xlsx”
  + **PR curve codes**:
    - “.\002\_Code\R\_code\PR\_curve\Conventional\_LR\PR\_curve\_lr.R”
    - “.\002\_Code\R\_code\PR\_curve\Conventional\_LR\plot\_pr\_curve.R”
  + **Results** are in “.\003\_Results\Conventional\_LR\_results”
    - *“\pred\_score\_alldata\”* : original prediction on cohort data
    - *“\pred\_socre\_allss\*”: original prediction on representative sample
    - *“\PR\_curve\cohort\_pred\_data.txt”*: combined prediction on cohort data
    - *“\PR\_curve\ss\_pred\_data.txt”*: combined prediction on representative sample
    - *“\PR\_curve\pr\_curve\_cohort.csv”:* grouped PR curve on cohort data
    - *“\PR\_curve\pr\_ss\_grp\_re.csv”:* grouped PR curve on representative sample
    - *“\PR\_curve\pr\_f\_0.3.jpeg”:* plot of PR curve for representative sample
    - *“\PR\_curve\pr\_f\_0.5.jpeg”:* plot of PR curve for representative sample
* Random Forest
  + **Code**: “.\002\_Code\Spark\_code\Random\_forest\_withGS\ 05\_RandomForest\_job**\***.py”
  + **Data files** are in S3 bucket “s3://emr-rwes-pa-spark-dev-datastore/BI\_IPF\_2016/01\_data/”:
    - *"Ipf\_cohort\_replace\_extremes.csv"*: IPF cohort
    - *"non\_Ipf\_cohort\_replace\_extremes.csv"*: Non-IPF cohort
    - *'score\_sample\_model\_data\_cap\_ext.csv'*: Representative sample
  + **Excluded variable list**: “/home/hjin/BI\_IPF\_2016/01\_data/vars\_to\_exclude\_list1.csv” on master node
  + **PR curve codes**:“.\002\_Code\R\_code\PR\_curve\ Random\_forest\PR\_curve\_RF.R”
  + **Results** are in “.\003\_Results\Random\_forest\_results”
    - *“\RF\_job1\_db8\_20160815\_015615\”* : results for job 1
    - *“\RF\_job2\_db4\_20160812\_115851\”* : results for job 2, **delivery**
    - *“\RF\_job3\_db9\_20160816\_021056\”* : results for job 3

Two-stage modeling

The rationale behind a two-stage model is to use the first stage to help define IPF and non-IPF that have the highest degree of similarity and thus focus the second-stage of the model of patients are more likely to be misdiagnosed. This is achieve by employing two cross-validation schemes – the first is an outer loop of 5-fold cross validation and the second is a nested cross validation that uses the training data of the outer cross validation loop as the input data that is subsequently partitioned into nested folds. In Figure 2, a single loop of the outer cross validation scheme is illustrated. For this single outer loop, the data have been partitioned into training and test folds where 80% of the data are contained in the TRAIN partition and 20% of the data are collected in the TEST partition. Recall that for training a 1 to 200 positive to negative ratio is used and for testing a 1 to 853 ratio is used. The TRAIN partition is further split into 5 nested folds and 5 logistic regression models are trained using nTRAIN and nTEST (the nested partitions). Each of these 5 models are applied to the data in the TEST partition to produce 5 predicted scores per observation in TEST. To produce a single score for TEST called pr\_TEST, these predicted scores are averaged across the 5 models. The predicted scores across TRAIN are collected from the nested models and called pr\_TRAIN.

The next step is to establish threshold to select patients who will enter the stage two model. The threshold is selected to retain the top 25% of positive patients and this is selected by only considering the positive patients in TRAIN and therefore the threshold is not using knowledge of TEST. The threshold is defined as the 75th percentile of the predicted scores of the positive patients in the TRAIN partition. All patients above this threshold in TRAIN are passed to TRAIN2S (i.e. the second stage training data) and all patients above this threshold in TEST are passed to TEST2S (i.e. the second stage test data). The exclusion criteria for variables is re-assessed at this point to ensure that variables meet the criteria on this subsample of the data set. A logistic regression model is then trained on TRAIN2S and tested on TEST2S. This procedure for the outer loop is repeated 5 times and the predicted scores for the second stage test data are collected and assessed using the chosen performance metrics. As before, overfitting was assessed by comparing the performance on the training data (TRAIN2S) and the test data (TEST2s).

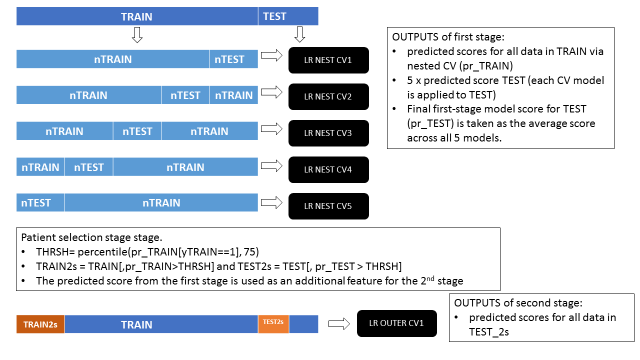


Figure 2 Scheme of a single loop of the outer cross validation scheme for the two-stage modeling approch.

#### Code for two-stage modelling

* Add interaction term and sequence variable:
  + **Code**: “.\002\_Code\Spark\_code\ Two\_stage\_model\06\_2stage\_outerCV\_topcode.py”
  + **Data files** are in S3 bucket “s3://emr-rwes-pa-spark-dev-datastore/BI\_IPF\_2016/01\_data/”:
    - *"* *all\_features\_pos.csv.csv"*: IPF cohort
    - *"* *all\_features\_neg.csv"*: Non-IPF cohort
    - *“all\_features\_score.csv”*: Representative sample
  + **Excluded variable list**: “/home/hjin/BI\_IPF\_2016/01\_data/vars\_to\_exclude\_list1.csv” on master node
  + **Sample strategy**: “.\001\_Reference\_file\Sample\_strategy\_Two\_Stage\_model\_LR.xlsx”
  + **PR curve codes**:
    - “.\002\_Code\R\_code\PR\_curve\two\_stage\pr\_curve\_2stage\_topcode.R”
  + **Results** are in “.\003\_Results\Two\_stage\_model\_results\two\_stage\_topcode”
    - *“\PR\_curve\_2s\Compare\_performance\_2\_stage\_modelling\_topcoding.csv”:* Performance comparison table for two-stage model
    - *“PR\_curve\_2s\FP\_count\_topcoding.csv”*: FP table for two-stage model
    - *“\pred\_score\_1stcombtr\_sim\*\”* : original prediction on training data for each simulation
    - *“\pred\_socre\_1stavgts\_sim\*\*”: original prediction on test data for each simulation
    - *“\pred\_socre\_1stavgrepts\_sim\*\”*: original prediction on representative sample for each simulation
    - *“\pred\_score\_2ndsubtr\_sim\*\”* : original prediction on sub-training data for each simulation
    - *“\pred\_socre\_2ndsubts\_sim\*\*”: original prediction on sub-test data for each simulation
    - *“\pred\_socre\_2ndsubrepts\_sim\*\”*: original prediction on sub-representative sample for each simulation
    - *“\PR\_curve\_2s\pr\_grp\_1stavgts.csv”*: grouped PR curve for test data in 1st stage
    - *“\PR\_curve\_2s\pr\_grp\_1stavgrepts.csv”*: grouped PR curve for representative sample in 1st stage
    - *“\PR\_curve\_2s\pr\_grp\_2ndsubts.csv”*: grouped PR curve for sub-test data in 2nd stage
    - *“\PR\_curve\_2s\pr\_grp\_2ndsubrepts.csv”*: grouped PR curve for sub-representative sample in 2nd stage
    - *“\PR\_curve\_2s\pr\_avgsim\_1stcombtr.csv”*: average PR curve for training data in 1st stage across simulations
    - *“\PR\_curve\_2s\pr\_avgsim\_1stavgts.csv”*: average PR curve for test data in 1st stage across simulations
    - *“\PR\_curve\_2s\pr\_avgsim\_1stavgrepts.csv”*: average PR curve for representative sample in 1st stage across simulations
    - *“\PR\_curve\_2s\pr\_avgsim\_2ndsubtr.csv”*: average PR curve for sub-training data in 2nd stage across simulations
    - *“\PR\_curve\_2s\pr\_avgsim\_2ndsubts.csv”*: average PR curve for sub-test data in 2nd stage across simulations
    - *“\PR\_curve\_2s\pr\_avgsim\_2ndsubrepts.csv”*: average PR curve for sub-representative sample in 2nd stage across simulations
    - *“\PR\_curve\_2s\FP\_table\_1st.csv”: FP table for 1st stage model*
    - *“\PR\_curve\_2s\FP\_table\_2nd.csv”: FP table for 2nd stage model*
* Just rerun 2nd stage model with new excluded variable list and new way of top-coding
  + **Code**: “.\002\_Code\Spark\_code\ Two\_stage\_model\06\_2stage\_just2nd\_topcode.py”
  + **Data files** are in S3 bucket “s3://emr-rwes-pa-spark-dev-datastore/BI\_IPF\_2016/01\_data/” and also in ./04\_Summary/004\_data/
    - *"* *all\_features\_pos.csv.csv"*: IPF cohort
    - *"* *all\_features\_neg.csv"*: Non-IPF cohort
    - *“all\_features\_score.csv”*: Representative sample
  + **Excluded variable list**: “/home/hjin/BI\_IPF\_2016/01\_data/vars\_to\_exclude\_list1.csv” and “/home/hjin/BI\_IPF\_2016/01\_data/vars\_to\_exclude\_list2.csv”on master node
  + **Sample strategy**: “.\001\_Reference\_file\Sample\_strategy\_Two\_Stage\_model\_LR.xlsx”
  + **PR curve codes**:
    - “.\002\_Code\R\_code\PR\_curve\two\_stage\pr\_curve\_2stage\_just2nd.R”
  + **Results** are in “.\003\_Results\Two\_stage\_model\_results\just\_2nd”
    - *“\PR\_curve\_2s\FP\_table\_2nd.csv”: FP table for 2nd stage model*
    - *“\PR\_curve\_2s\FP\_table\_2nd\_wtthres.csv”: FP table for 2nd stage model with thresholds*
    - *“\pred\_score\_2ndsubtr\_sim\*\”* : original prediction on sub-training data for each simulation
    - *“\pred\_socre\_2ndsubts\_sim\*\*”: original prediction on sub-test data for each simulation
    - *“\pred\_socre\_2ndsubrepts\_sim\*\”*: original prediction on sub-representative sample for each simulation
    - *“\PR\_curve\_2s\pr\_grp\_2ndsubtr.csv”*: grouped PR curve for sub-training data in 2nd stage
    - *“\PR\_curve\_2s\pr\_grp\_2ndsubts.csv”*: grouped PR curve for sub-test data in 2nd stage
    - *“\PR\_curve\_2s\pr\_grp\_2ndsubrepts.csv”*: grouped PR curve for sub-representative sample in 2nd stage
    - *“\PR\_curve\_2s\pr\_avgsim\_2ndsubtr.csv”*: average PR curve for sub-training data in 2nd stage across simulations
    - *“\PR\_curve\_2s\pr\_avgsim\_2ndsubts.csv”*: average PR curve for sub-test data in 2nd stage across simulations
    - *“\PR\_curve\_2s\pr\_avgsim\_2ndsubrepts.csv”*: average PR curve for sub-representative sample in 2nd stage across simulations

# Application of the Two-stage Model to the Scoring Sample

To apply the two stage model the scoring sample – the cross validation scheme can be altered to ensure that the maximal amount of the data are available for training. Particularly, the outer loop of cross validation is no longer necessary as the model is only required to make predictions on the independent scoring sample which at no point are used to train the model. Therefore, the scheme in *Figure 3* is implemented. In this case, the logistic regression model is trained and tested using 5 fold cross validation on the full set of positive patients and 200 of their matched negative patient which is referred to as the COHORT sample. These 5 models are then tested on the complete scoring sample (SCORING SAMPLE) and the predicted scores are averaged to produce a single score per patient in the scoring sample.

To selected patients for the second stage model the threshold on the predicted score is defined as the 75th percentile of the positive patients in COHORT. Both the COHORT and SCORING SAMPLE sets are then restricted to patients with a predicted score above this threshold. A second stage model is then trained on the sub-selected COHORT data (COHORT2s) and tested on the sub-selected SCORING SAMPLE.

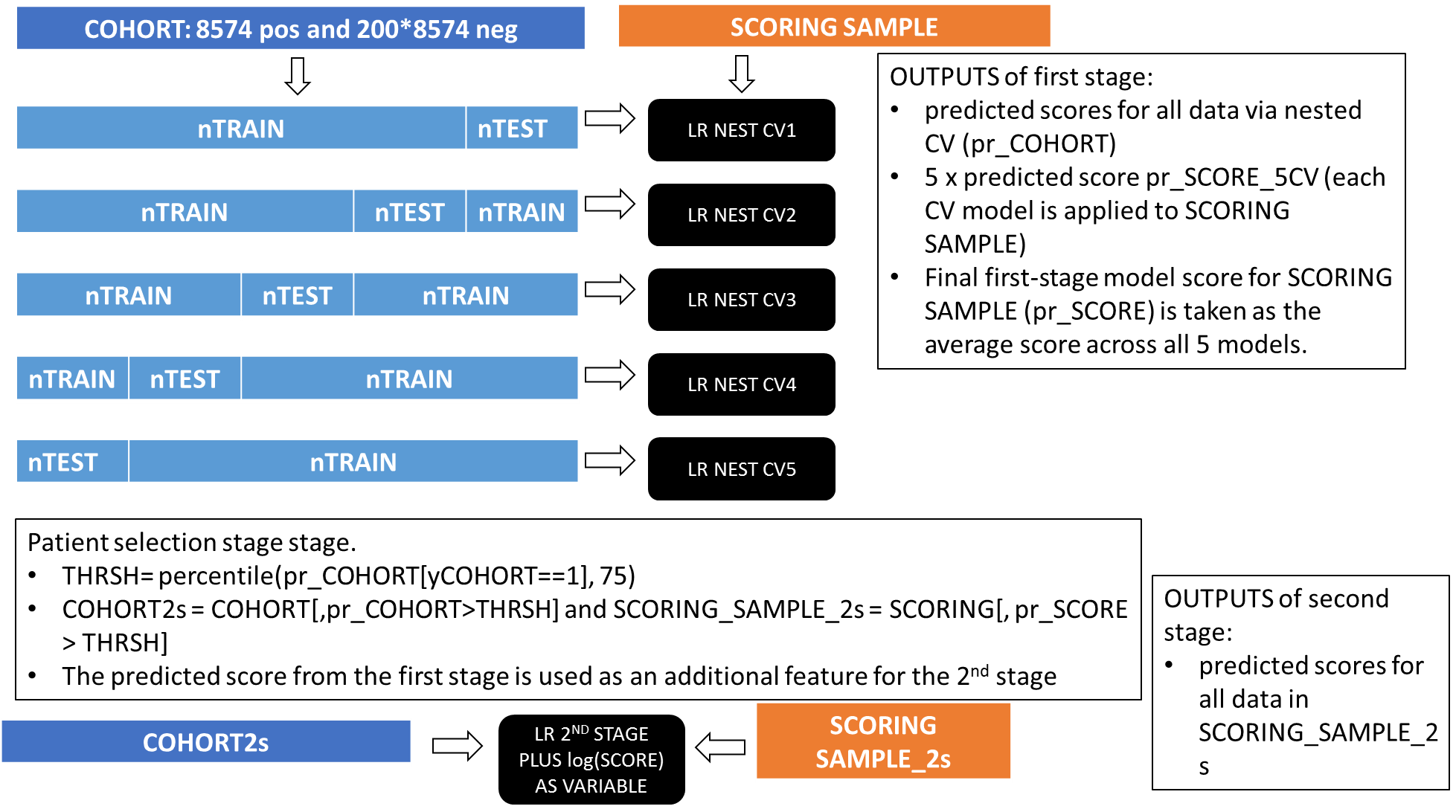


Figure 3 Scheme for application of two-stage model to the scoring sample.

## Code for applying two-stage model on scoring sample

* **Code**: “.\002\_Code\Spark\_code\ Two\_stage\_model\06\_2stage\_newss26m\_topcode.py”
* **Data files** are in S3 bucket “s3://emr-rwes-pa-spark-dev-datastore/BI\_IPF\_2016/01\_data/”:
  + *"* *all\_features\_pos.csv.csv"*: IPF cohort
  + *"* *all\_features\_neg.csv"*: Non-IPF cohort
  + *“Ss\_26m\_218\_features.csv”*: Scoring sample
* **Excluded variable list**: “/home/hjin/BI\_IPF\_2016/01\_data/vars\_to\_exclude\_list1.csv” and “/home/hjin/BI\_IPF\_2016/01\_data/vars\_to\_exclude\_list2.csv”on master node
* **PR curve codes**:
  + “.\002\_Code\R\_code\PR\_curve\two\_stage\pred\_score\_2stage\_newss.R”
* **FP table with thresholds:** ““.\003\_Results\Two\_stage\_model\_results\just\_2nd\PR\_curve\_2s\*FP\_table\_2nd\_wtthres.csv*”
* **Results** are in “.\003\_Results\Two\_stage\_model\_results\applying\_on\_scoring\_sample”
  + *“\pred\_score\_2ndsubtr\”* : original prediction on sub-training data
  + *“\pred\_socre\_2ndsubss\_sim\*\*”: original prediction on sub-scoring sample
  + *“\pred\_score\_all\pr\_grp\_subtrcsv”*: grouped PR curve for sub-training data in 2nd stage
  + *“\pred\_score\_all\pr\_grp\_subtrss.csv”*: grouped PR curve for sub-training plus sub-scoring sample in 2nd stage
  + *“\pred\_score\_all\pred\_score\_2ndsubtr.csv”*: prediction for sub-training data
  + *“\pred\_score\_all\pred\_score\_2ndsubss.csv”*: prediction on sub-scoring sample
  + *“\pred\_score\_all\subdata\_5%recall.csv”*: patient ID with predicted probabilities for 5% recall patients
  + *“\pred\_score\_all\subdata\_10%recall.csv”*: patient ID with predicted probabilities for 10% recall patients
  + *“\pred\_score\_all\subdata\_20%recall.csv”*: patient ID with predicted probabilities for 20% recall patients

## Code for extract subset patients from original scoring sample

* **Code**: “.\002\_Code\SAS\_code\06\_extract\_predicted\_patients.sas”
* **Input data**: “.\003\_Results\Two\_stage\_model\_results\applying\_on\_scoring\_sample\pred\_score\_all”
  + *“subdata\_5%recall.csv”*
  + *“subdata\_10%recall.csv”*
  + *“subdata\_20%recall.csv”*
* **Output data:** “.\04\_Summary\003\_Results\Extraction\_predicted\_patients”
  + *“subdata\_all\_5rec\_v2.csv”*
  + *“subdata\_all\_10rec\_v2.csv”*
  + *“subdata\_all\_20rec\_v2.csv”*

# Performance Metrics

The performance of the model will be assessed using precision-recall curves whereby the precision and recall are plotted across a range of thresholds applied to the predictions. Precision is defined as the number of true positives divided by the sum of the true positives and false positives. The recall is defined as the number of true positive divided by the sum of the true positives and false negatives.

For the two stage model, a reduced number of patients are considered. Nonetheless, the goal is to determine the rates of positive patients found by the algorithm in comparison to the original cohort of 8574 positive patients. Therefore, in this case recall is calculated as the number of true positives identified divided by 8574 in all cases and precision is calculated as above.

# Model Interpretation

To provide insight into variable importance that regression coefficients from the logistic regression model were extracted and multiplied by the range of the corresponding variable. The absolute value was computed and these were ranked in ascending order with the variables with higher magnitude implying high importance.

To provide insight into the directionality and importance of a variable in relation to the predictive power of the model then the relative risk in terms of the predicted score was computed.

Binary variables

For binary variables, this was computed using the level coded as 0 as the baseline and the level coded as 1 as the level of interest. All patients were stratified into two groups, those with the flag set 1 and those with the flag set to 0. The mean predicted score from the first stage modelling approach was calculated per group. To relative risk was computed for the group with the flag set to one by dividing the mean score for this group by the mean score of the group with the flag set to zero.

Age

Patients were collected by age in the following five groups 40 to 49, 50 to 59, 60 to 69, 70 to 79 and >80. The baseline group was defined as the 40 to 49 group. The mean predicted score was computed for each group by considering all patients that fall within the age range of the group. The relative risk four groups where age > 50 by divided the mean predicted score for that group by the mean predicted score for patients who are between 40 and 49 years old.

Frequency variables

Patients were grouped by the frequency (in years) at which an event occurred in the following bins, 0 (i.e. the event was not observed for that patient), >0 and <=1, > 1 and <=2 and > 2. The baseline group is defined as the group of patients with a frequency of 0. The relative risk for the remaining groups was calculated by dividing the mean predicted score for that group by the mean predicted score for the baseline group.

**Code**: “.\002\_Code\R\_code\Model\_interpretation\descriptive\_score\_all\_vars.R”

**Input data**:

Predicted scores: “F:/orla/projects2016/BI\_IPF/results/two\_stage/pred\_score\_1stavgts\_tc\_v1.csv”

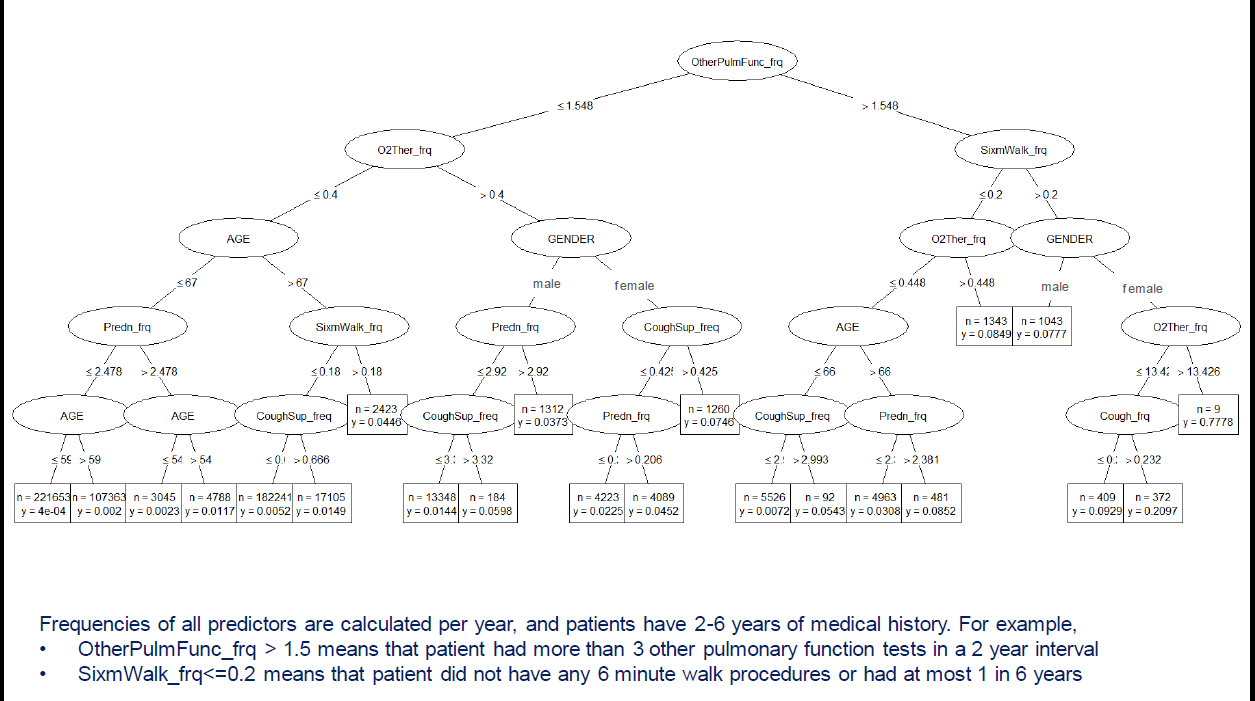
IPF cohort: “F:/orla/projects2016/BI\_IPF/data/model\_data\_175/all\_features\_pos\_for\_orla\_CNT\_xform.csv”

Non IPF cohort (200 matched patients for each positive IPF patient): “F:/orla/projects2016/BI\_IPF/data/model\_data\_175/all\_features\_neg\_for\_orla\_CNT\_xform.csv”

**Output data:**  “.\003\_Results\Two\_stage\_model\_results\Model\_interpretation\” on server 101

* Average predictive scores and relative risks for binary variables: “flag\_vars\_RR.csv”
* Average predicted scores and relative risks for continuous variables: “freq\_vars\_RR.csv”

# Appendix 1





# Appendix 2

|  |
| --- |
| LVL3\_ANTI\_BIO\_LRX\_M\_S\_DT\_LVL3\_SOB\_DYSP\_M\_S\_DT\_BEFORE |
| LVL3\_ANTI\_BIO\_LRX\_M\_S\_DT\_LVL3\_O\_PULM\_F\_T\_M\_S\_DT\_BEFORE |
| LVL3\_ANTI\_BIO\_LRX\_M\_S\_DT\_LVL3\_O2\_THER\_M\_S\_DT\_BEFORE |
| LVL3\_ABN\_CHST\_XRAY\_M\_S\_DT\_LVL3\_PRED\_LRX\_M\_S\_DT\_BEFORE |
| LVL3\_COUGH\_SUP\_LRX\_M\_S\_DT\_LVL3\_HYPOXEM\_M\_S\_DT\_BEFORE |
| LVL3\_ANTI\_BIO\_LRX\_M\_S\_DT\_LVL3\_O\_PULM\_F\_T\_M\_S\_DT\_AFTER |
| LVL3\_PRED\_LRX\_M\_S\_DT\_LVL3\_SOB\_DYSP\_M\_S\_DT\_AFTER |
| LVL3\_ACUT\_RESP\_INF\_M\_S\_DT\_LVL3\_O\_PULM\_F\_T\_M\_S\_DT\_AFTER |
| LVL3\_ACUT\_RESP\_INF\_M\_S\_DT\_LVL3\_HYPOXEM\_M\_S\_DT\_BEFORE |
| LVL3\_GERD\_LRX\_M\_S\_DT\_LVL3\_O2\_SATUR\_M\_S\_DT\_BEFORE |
| LVL3\_HYPOXEM\_M\_S\_DT\_LVL3\_O\_PULM\_F\_T\_M\_S\_DT\_AFTER |
| LVL3\_O2\_SATUR\_M\_S\_DT\_LVL3\_PNEU\_M\_S\_DT\_AFTER |
| LVL3\_COUGH\_M\_S\_DT\_LVL3\_O2\_SATUR\_M\_S\_DT\_BEFORE |
| LVL3\_HYPOXEM\_M\_S\_DT\_LVL3\_PNEU\_M\_S\_DT\_AFTER |
| LVL3\_O2\_THER\_M\_S\_DT\_LVL3\_PRED\_LRX\_M\_S\_DT\_AFTER |
| LVL3\_ACUT\_RESP\_INF\_M\_S\_DT\_LVL3\_O2\_SATUR\_M\_S\_DT\_BEFORE |
| LVL3\_ANTI\_BIO\_LRX\_M\_S\_DT\_LVL3\_RESP\_FAIL\_M\_S\_DT\_AFTER |
| LVL3\_GERD\_LRX\_M\_S\_DT\_LVL3\_O2\_THER\_M\_S\_DT\_BEFORE |
| LVL3\_O2\_THER\_M\_S\_DT\_LVL3\_PNEU\_M\_S\_DT\_AFTER |
| LVL3\_O2\_THER\_M\_S\_DT\_LVL3\_SOB\_DYSP\_M\_S\_DT\_AFTER |
| LVL3\_O2\_THER\_M\_S\_DT\_LVL3\_O\_PULM\_F\_T\_M\_S\_DT\_AFTER |
| LVL3\_COUGH\_M\_S\_DT\_LVL3\_O2\_THER\_M\_S\_DT\_BEFORE |
| LVL3\_MAL\_FATIG\_M\_S\_DT\_LVL3\_O\_PULM\_F\_T\_M\_S\_DT\_BEFORE |
| LVL3\_CHEST\_PN\_M\_S\_DT\_LVL3\_RESP\_FAIL\_M\_S\_DT\_BEFORE |
| LVL3\_HYPOXEM\_M\_S\_DT\_LVL3\_SOB\_DYSP\_M\_S\_DT\_AFTER |
| LVL3\_CHEST\_PN\_M\_S\_DT\_LVL3\_GERD\_LRX\_M\_S\_DT\_BEFORE |
| LVL3\_CHEST\_PN\_M\_S\_DT\_LVL3\_O2\_THER\_M\_S\_DT\_BEFORE |
| LVL3\_CHEST\_XRAY\_M\_S\_DT\_LVL3\_RESP\_FAIL\_M\_S\_DT\_BEFORE |