Introduction

- Patient diagnosis is very relevant for retrospective studies. It allows to evaluate patient streams and treatment patterns by indication. However, many data assets lack diagnosis information which impairs the ability to gain valuable insight into real-world drug utilization.
- In order to derive patient diagnosis heuristic expert model based on medical guidelines was developed. The low accuracy of expert model limits possibilities of its broader use in the respiratory market.

Objective

- The objective of this research stream was to identify missing patient diagnosis in patient level data assets based on patient characteristics and historical treatment patterns as well as prescribed drug and prescriber information.
- We developed and validated an algorithm that quantifies the relationship between diagnosis and patient characteristics for the respiratory market, consisting of Asthma and Chronic Obstructive Pulmonary Disease (COPD), and allows for indication prediction. Diagnosis imputation was based on relationship between diagnosis, product, patient and prescriber characteristics observed in medical data.

Method

Data:
- We used patient level EMR (electronic medical record) data including anonymized patient identifier, patient consultation with diagnosis (ICD10), prescribed drug, and physician and patient specific data dimensions. Data covers Q2/2014-Q1/2015 and is based on sample of 1,860 physicians including General Practitioners and Specialties.

Patient Selection and Classification:
- We focused on the respiratory market, specifically Asthma and COPD. Using ICD10 codes, all patients were classified accordingly as Asthma, COPD or Others (acute bronchitis, bronchitis, bronchiectasis, respiratory failure, cough). Data cleaning was undertaken to exclude patient records with unclassifiable indications in order not to consider confounding co-morbidities. Data for ~285k patients was used for modeling.

Explanatory Variables:
- The algorithm was built based on published treatment guidelines, literature research, and data investigations. The following information were used for model development: patient attributes, prescription attributes, patient historical treatment and co-treatment pattern, and prescriber attributes.

Approach:
- We used CHAID Model (Chi-squared Automatic Interaction Detector) and Random Forest to quantify the relationship between patient diagnosis and patient characteristics, including past visit and therapy patterns.
- In the next step, we used the model to predict indication at patient consultation level. And finally, we compared drug utilization patterns for different diagnosis.

Validation:
- Models were estimated on training sample (30%) and validated on holdout sample (70%). We compare observed and predicted indication on the consultation level. Additionally, we look at aggregated distribution of indications.

Results

- Data of ~285k patients shows high share of Asthma (49%) followed by COPD (22%) and Others (29%). Main treating specialty is Pulmonology with remaining specialties having significantly less patients of interest. Main indication drivers included patient treatment and co-treatment patterns as well as patient characteristics.
- The model produced very high accuracy rates, with Random Forest outperforming CHAID models.
- In case of Machine Learning, we obtained hit rates for Asthma of 93% in training sample and 85% in holdout sample, while COPD hit rates were 90% in training and 80% in holdout sample.
- On aggregated level, predicted diagnosis distribution is also well aligned to observed.

Data Description

- Split of Asthma Patients by Diagnosis
- Split of COPD Patients by Diagnosis

Model Evaluation

- Hit Rates from CHAID Modeling
- Hit Rates from Random Forest

Conclusions

- We developed and validated an algorithm that can model diagnosis versus patient characteristics for Asthma and COPD, and allows for indication prediction
- Our approach combines the strengths of medical expertise with robust analysis using CHAID and Machine Learning to create a new standard for prediction of patients’ indication