Challenges and Approaches in Predictive Modelling: An Asthma/COPD Algorithm

Lorena Cirneanu[1] (lorena.cirneanu@iqvia.com), Eleanor Ralphs[1], Fiona Grimson[1], Tomas Radovich[1], Benjamin Bray[1], Joseph Kim[1,2]

[1]NEMEA Centre of Excellence for Retrospective Studies, Real-World Insights, IQVIA, London, UK
[2]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine

IMS Health & Quintiles are now IQVIA™

Key Takeaways

The aim of this study is to reduce misclassification between asthma and COPD, by comparing the performance of four different approaches to predict the diagnosis of patients based on a number of covariates.

One point worth mentioning is that the patients' diagnosis is known in our study, thus predicting the diagnosis is meant for learning purposes. This way, the best model chosen will be used to reduce the misclassification error for future instances where diagnosis is unknown.

The four different approaches used are assuming all patients have asthma (null model), using clinical rules to diagnose patients with asthma or COPD (deterministic model), logistic regression and AdaBoost modelling, a machine learning method.

Performance of the four models was assessed by comparing the sensitivities of the four models, so the percentage of patients correctly diagnosed with a given disease.

Based on improved classification of asthma (98%) and COPD (90%) patients, we can conclude that the AdaBoost model is the best choice for assigning diagnoses in this study population.

Methods

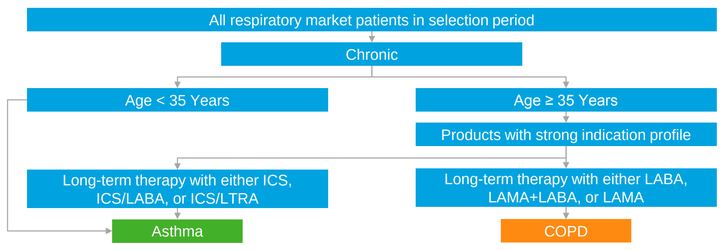
The four methods used to predict the asthma or COPD diagnosis of patients are described in *Table 1*. Drawbacks of these approaches were identified and the best model was determined. Statistical analysis was conducted in SAS Enterprise Guide 7.1 and R 3.5.3.

Table 1: Models and Variables Used to Predict the Asthma or COPD Diagnosis of Patients

Methods	Description of Model	LPD Variables Included (10,000 patients)	IMRD Variables Included (8,653 patients)
Null Model	Assumes the whole study population is diagnosed with asthma and no patient is diagnosed with COPD.	None	None
Deterministic Model	Uses clinical rules to diagnose patients with asthma or COPD, based on medical knowledge as in *Figure 1.*	• Age • New indication (Acute/Chronic) • Strong Product Indicator • Class (Therapy combinations)	None
Logistic Regression Model a) One model with LPD variables only b) One model with LPD + IMRD variables	Splits the dataset into train data and test data, where the training data is used for modelling purposes, while the test data is used for assessing the performance of the model.	• Age • Gender • Class (Therapy combinations) • Molecule • Brand • Duration of prescription • Strong product indicator (Y/N) • New indication • Reliability measure of duration (Y/N)	• Diabetes • Smoking status • Spirometry
AdaBoost Machine Learning Model a) One model with LPD variables only b) One model with LPD + IMRD variables	Checks whether the misclassification rate of asthma and COPD diagnoses can be decreased by converting a set of weak classifiers into a strong one.		

Deterministic Model

Figure 1: The Deterministic Model. Uses Clinical Rules to Diagnose Patients with Asthma or COPD, Based on Medical knowledge



Logistic Regression Model

Predictive modelling was used to fit a model with diagnosis binary outcome. The model was adjusted for the variables present in *Table 1*. The data was split in 50% training and 50% test dataset. The training data was used to train the algorithm on common and different characteristics for both diagnoses in order to see how well the diagnoses were predicted once specific predictors were considered.

AdaBoost Machine Learning Model

AdaBoost modelling was used to assign variable importance and to generate a classification tree with the scope of improving the asthma and COPD deterministic algorithm. AdaBoost modelling works by fitting weak classifiers to the dataset, and selects the variable with the lowest weighted classification error. Subsequently, it calculates the weight for the n^{th} weak variable and updates the weight for each data point. After n iterations, the final prediction is obtained by summing up the weighted prediction of each classifier[3]. This produces a decision tree for boosting, which displays the final predicted variables.

The improved classification tree obtained in *Figure 5* was compared to the deterministic model. The purpose of this approach was to compare a data-driven approach to a clinical-knowledge approach.

Model Performance

The performance of the prediction was assessed by a confusion matrix, for all four models and by a ROC curve for the logistic model. The main comparison mechanism was the sensitivity of each model.

Background

An estimated 5.4 million people are living with asthma and 1.2 million with diagnosed COPD in the UK[1]. Due to the similarities of symptoms and medications prescribed, both conditions are prone to disease misclassification in electronic medical record databases.[2]

The aim of the study is to reduce possible misclassification between asthma and COPD, by using four approaches. To assess the performance of the four methods, records of patients with diagnosis of asthma or COPD from Longitudinal Patient Data (LPD) and IQVIA Medical Research Data (IMRD) were used. LPD is a commercially available database which provides information on product and therapy indicators, as well as patient level data information. LPD is a subset of observations of IMRD. IMRD is a primary care electronic medical record database which represents around 6% of the UK population. IMRD includes data on comorbidities, clinical characteristics and tests, which will help inform some of the predictive approaches.

The inclusion criteria in both LPD and IMRD* for this study are:
- Patients with age ≥ 18 years;
- Patients with either asthma or COPD diagnosis.

The exclusion criteria are:
- Patients with urticaria diagnosis;
- Patients with < 2 exacerbations/prescriptions per year.

The study period is 1st June 2013 to 31st May 2018.

*IMRD is a registered trademark of Cegedim SA in the United Kingdom and other countries. Reference made to the IMRD database is intended to be descriptive of the data asset licensed by IQVIA; This work uses de-identified data provided by patients as a part of their routine primary care.

Results

After applying variable selection for the logistic models and generating the variable importance ranking for the AdaBoost models, the IMRD variables were dropped due to having a significance level of p > 0.10.

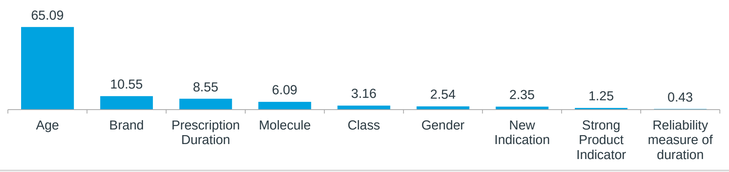
Only the models fitted using the LPD dataset will be presented in the upcoming results. The variables kept in each model are displayed in *Table 2*.

Table 2: Final Variables for Logistic Regression and AdaBoost Models

Models Using LPD Data	Final Variables	
Logistic Model	• Age • Gender • Strong Product Indicator • New Indication • Duration of Prescription	• Class • Molecule • Brand • Reliability measure of duration
AdaBoost Model	• Age • New Indication • Gender • Class	• Molecule • Brand • Duration of prescriptions

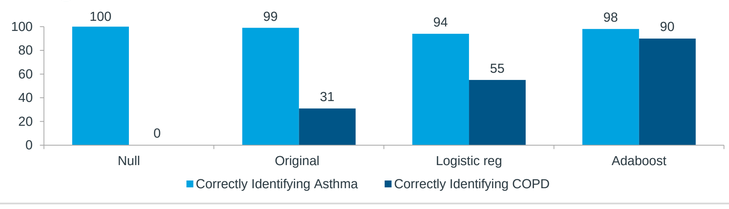
The relative importance of each variable, obtained from the AdaBoost model, is displayed in *Figure 2*. The condition to qualify for the AdaBoost model requires the variables to have a relative importance >2. Strong Product indicator and Reliability measure of duration variables did not qualify, therefore they were not included in the final AdaBoost model.

Figure 2: Relative Variable Importance Selection Mechanism and Relative Importance Values from the AdaBoost Model



After fitting all models, *Figure 3* gives a comparison of the percentage of patients correctly diagnosed with asthma and COPD.

Figure 3: Sensitivities Across all Models Fitted Using LPD Data



Conclusion

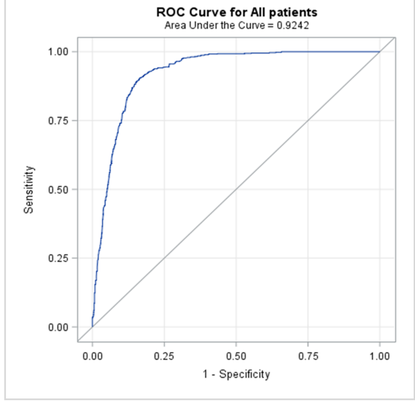
Based on the improved classification of asthma (98%) and COPD (90%) patients and the implicit reduction in the misclassification error, we can conclude that the AdaBoost model is the best choice for assigning diagnoses, in this study population. Moreover, while the probability of correctly identifying asthma across all patients remains constant for all models (between 94% and 100%), the probability of correctly identifying COPD patients drastically increases from 0% to 90%.

Finally, the classification tree generated by the AdaBoost model is more efficient than the deterministic model, as it is composed from only two predictors: age, which was split differently to the original deterministic model, and brand. These variables had the highest relative importance score. Whereas, the deterministic model contained four predictors: new indication, age, strong product indicator and class, of which strong product indicator and new indication did not qualify for the final model.

How Did the Models Perform?

Performance of the logistic model was assessed by calculating the sensitivities of asthma and COPD based on the ROC curve provided in *Figure 4*. The sensitivities for both diagnoses across all models are given in *Figure 3*.

Figure 4: ROC Curve for Logistic Regression Model



The AdaBoost model correctly classifies asthma patients 98% of the time, similarly to the null (100%) and deterministic model (99%), while the logistic regression model has a relatively lower classification rate for asthma (94%).

However patients with COPD have a lower predictive rate overall compared to the asthma patients, which is expected, due to COPD being a less common condition than asthma and harder to diagnose due to common symptoms with asthma and common age-targeted populations[2].

AdaBoost model provides the highest correct classification of COPD with a 90% sensitivity, compared to the logistic regression model (55%), deterministic model (31%) and null model (0%). The null model result is as expected as we assumed no patient has COPD in the population.

Why are the Models Not Performing Well?

1. Null Model (Everyone has Asthma)

- As anticipated, we see a better sensitivity for this model because the number of asthma patients is generally larger than that of the COPD patients, so once we assume that everyone has asthma, we have a higher chance (99%) of having asthma patients classified correctly.
- For equal sample sizes and under the same assumption that everyone has asthma, the sensitivity may decrease for asthma patients and may increase for COPD patients.
- Thus, taking into consideration all the above and the fact that no predictors were considered for this model, a better model alternative is needed.

2. Deterministic Model (Using the Clinical Deterministic Fitting of Rules)

- One drawback of this model is that it decreases the model's discriminating ability by classifying patients as having asthma or COPD based on predictors such as age, therapy combinations, new indication and strong product indicator, while disregarding the idea of training the algorithm on a dataset and then testing it on different data.

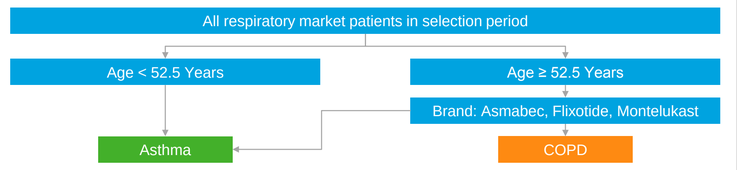
3. Logistic Model

- A disadvantage of this model is that it focuses more on the predictors than on the loss function. A loss function is high if the algorithm does not model the data well. Compared to the logistic model, the AdaBoost model finds combinations of classifiers to minimise misclassification error, thus minimising the loss function. In order to improve the classification rate for the logistic model, a LogitBoost approach could be used instead[4].

How can we improve the Deterministic Model?

This data-driven decision tree generated by the AdaBoost model in *Figure 5* can be used to improve a clinical-knowledge approach to determining the asthma or COPD status. The AdaBoost tree was reduced to two important classifiers, age and brand, compared to the deterministic model in *Figure 1* which has as predictors age, new indication, strong product indicator and class.

Figure 5: Final Classification Tree Produced by the AdaBoost Model



Limitations

Some study and model limitations include:

- Due to limited memory size in the R software, only the first 10,000 patient were used of the LPD dataset for all models.
- High levels of missingness in the IMRD variables (diabetes, smoking status, spirometry) led to the variables not being selected in the final models.
- The results may vary for different datasets.
- When linking and merging the two datasets, some observations are lost due to a number of patients in LPD not having corresponding observations in IMRD. This may be due to a different identification number for the same patient in LPD compared to IMRD, which were reconstructed back only partially to match the identification number in IMRD.

References

1. Bloomer A. (2018). Asthma/COPD Overlap: Diagnosis and Management. GM Journal, 48(10)
2. Miravitlles, M. (2017): Diagnosis of asthma-COPD overlap: the five commandments. European Respiratory Journal 49: 1700506; DOI: 10.1183/13993003.00506-2017
3. Schapire R.E. (2013). Explaining AdaBoost. In: Schölkopf B., Luo Z., Vovk V. (eds) Empirical Inference. Springer, Berlin, Heidelberg
4. Zhi-Hua Zhou (2012). Boosting. In: Ensemble Methods: Foundations and Algorithms, Chapman & Hall/CRC, 23-46.