SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 8 | August 2024 - Peer Reviewed Journal

ADVANCED MACHINE LEARNING FOR PREDICTIVE MODELING FOR ASTHMA RISK FACTORS :COMPARATIVE STUDIES AND ANALYSIS

Gundra Madhan Sai Kumar[1], Dr.A P Bhuvaneswari[2]
[1]*School of Computer Science and Application,Reva University,Bengaluru, India*
[2]*Associate Professor,School of Computer Science and Application,Bengaluru, India*

Article DOI: https://doi.org/10.36713/epra18117
DOI No: 10.36713/epra18117

ABSTRACT

Asthma is a chronic and multifactorial respiratory disease that depends on the patient's genetic make-up, the internal environment in the body and the external environment within which the patient exists. These complex influences are explored in this project by employing a large number of patients' records, 2,392. It consists of numerous demographic characteristics, lifestyles (tobacco use and exercise), environmental conditions (pollution and irritants), past medical history (asthma in first-degree relatives and allergies), physical examination (spirometry), and symptoms (wheezing, breathlessness). Some of the crucial goals of the present work are to determine the factors that would impact the prevalence of Gastroesophageal Reflux and Smoking status among patients with asthma. It involves data cleaning by dealing with the missing values and categorical variables, and feature scaling by using z-score to deal with outliers. The SMOTE algorithm is used on the dataset to cope with the problem of class imbalance.

Logistic Regression, Decision Trees, KNN, AdaBoost, and NB algorithms are used to forecast the target variables. There are a lot of evaluation matrices such as accuracy, precision, recall, F1 measure, Receiver Operating Characteristic (ROC) Area Under the Curve, and many more to evaluate the model. ROC curves and precision-recall curves are commonly used to give a full and accurate evaluation of the model's performance with the help of confusion matrices. The outcomes of the study shall help to increase understanding of factors that underlie asthma and enhance the potential for prognosis and patient management at the individual level

KEYWORDS: *phase, age, gender, occupation, geographic location, history of diseases, vital statistics, signs, artificial intelligence, prognostic modeling, performances, health facilities*

1. INTRODUCTION

Asthma is one of the most common chronic diseases with a global distribution affecting the respiratory system through inflammation and increased sensitivity of the airways. Knowledge regarding the complexity of asthma is related to its causes and the way it must be treated. In this present work, the authors use a large sample of 2,392 patients and extract various characteristics like demographic details, lifestyle parameters, environmental details, medical history, clinical parameters, and symptoms assessment. It has thereby been possible to foster the use of sophisticated machine learning. the method that would be applied to analyse the results and proposes methods like logistic regression, decision trees and SMOTE for handling class imbalance: this research therefore seeks to establish the predictors of Gastroesophageal Reflux and Smoking among patients with asthma. Performance metrics to be used before and after data analysis include accuracy, precision, recall, and the ROC AUC of the predictive models. It is expected that findings from this research will be useful in improving the strategies of managing asthma and the design of interventional approaches to be used in clinical settings.

2. RELATED WORK

Extensive studies have been conducted to establish different variables related to asthma; these include genetic makeup as well as environmental triggers. The authors provided research evidence linking asthma with other demographic factors including age, gender and ethnicity showing that asthma tends to differ between the populations in terms of incidence and intensity (1). Even the smokers, their habits of smoking, physical activities, and diet also play a significant role in aggravating or improving the asthmatic conditions (2).

Airborne particles, for instance, pollution, pollen, dust are causative factors of asthma episodes or impact the disease process (3). Data on medical history which comprise of family history of asthma, allergies, eczema or any other diseases and other co-morbidities such as gastric oesophageal reflux have been shown to be associated with higher risk and severity of asthma (4). Parameters like forced expiratory volume in one second (FEV1), forced vital capacity (FVC) are instrumental in monitoring disease severity and the effectiveness of the therapeutic intervention (5).

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 8 | August 2024 - Peer Reviewed Journal

In recent years, various artificial intelligence technologies have shown interest in studies carried out on asthma; for instance, predictive techniques in diagnosing, severity and even finding an utmost suitable treatment plan (6). Predictive approaches of asthma have been done using logistic regression model, decision trees and ensemble approaches by using comprehensive databases of similar nature to our datasets (7). Nevertheless, the issues regarding the class imbalance as well as improving the recognition accuracy of the model are still unresolved in this discipline.

While continuing the work of previous research, this study contributes to the pool of knowledge by combining numerous factors from a large sample and analyzing it using modern methods of machine learning to reveal new aspects of asthma's multifactorial nature and develop improved decision-making in clinical practice.

3. METHODOLOGY

1. Dataset Description and Preprocessing

Dataset Overview: The study employs a sample of 2392 patients' data with the variables including demographic data, lifestyle, environmental characteristics, past medical history, anthropometry and clinical signs and symptoms, and asthma diagnostic markers. Data Preprocessing: Next, is the data cleaning that involves testing and checking for missing values and other unusual values. Categorical variables are converted from other categorical variables using label encoding While dealing with outliers, z-score is used and by this a treatment is done or the outliers are removed depending on the effect required.

2. Feature Selection and Engineering

Target Variables: This work centers on identification of two outcome predicates: Gastroesophageal Reflux and Smoking status on asthma patients.

Feature Engineering: Features of relevance are computed dependent on the correlation analysis with respect to the target variables and with the content knowledge. To handle the class imbalance in the target variables, Synthetic Minority Over-sampling Technique (SMOTE) is used.

3. Machine Learning Models

Model Selection: Often, for the tasks in the prediction area, several supervised models of machine learning are selected, namely Logistic Regression, Decision Trees, KNN (K-Nearest Neighbors), AdaBoost, and Naive Bayes classifiers. The models are chosen with regard to the model's applicability to binary classification tasks and the model interpretability.

Model Training and Evaluation: The dataset is preprocessed and the models are trained on the preprocessed data with training testing splitting of 80:20. Common measures like accuracy sometimes with precision, recall, F-1 measure, ROC AUC among others are used to evaluate performance of the models. When generalizing the results obtained, cross-validation techniques are used to cross-check and validate the analysis performed.

4. Performance Evaluation and Interpretation

Metrics Evaluation: The assessment of all models is based on the overall performance and class-specific measures, such as accuracy to estimate the whole model's performance, precision and recall to assess the precise and comprehensive identification of classes, respectively, ROC AUC indicating the model's discrimination capacity, and F1 score, which is the average value of accuracy and precision.

Interpretation: To realize these outcomes, they are interpreted as a way of realizing the pathway in addition to figuring out the most likely key factors that predict the Gastroesophageal Reflux and Smoking status among patients with asthma. Knowledge acquired from the models helps to clarify the interrelations between numerous factors and results regarding asthma.

5. Visualization and Reporting

Visualization: One of the elements of the output are visualizations of the classifier, which are ROC curves, precision-recall curves, confusion matrix, and feature importance plot.
Reporting: There are comprehensive reports and presentations presented according to the outcome the samples, predictor variables, model comparison, and clinical implications and suggestions for the next research studies.

4. DATA COLLECTION

The sample data used in the present study includes cross-sectional health data obtained from 2,392 patients with Asthma Disease. Interviews and medical examinations were standardised, and professionals working in various heath care centres participated in data collection. Key data points collected include:Key data points collected include:

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 8 | August 2024 - Peer Reviewed Journal

Demographic Details: Age, sex, race/ethnicity, and schooling/professional level.
Lifestyle Factors: Age in years old X, Body Mass Index (BMI), current smoking, the number of hours engaged in moderate and vigorous physical activity per week, the diet quality score, the sleep quality score.

Environmental and Allergy Factors: Frequency of exposure to pollution, exposure to pollen, exposure to dust and pet allergy.
Medical History: Asthma in the family, allergy history, existence of skin disease known as eczema, allergic conjunctivitis also referred to as hay fever, and the state of gastroesophageal reflux.

Clinical Measurements: Pulmonary function tests such as, FEV 1 and FVC.
Symptoms: Frequency of wheezing, breathlessness, chest pain/tightness, coughing particularly at night and during exercise.
Diagnosis Information: This variable will represent an indicator of the presence/absence of Asthma diagnosis.

Several sources of information were used and the issue of data protection and ethical concerns was fully observed right from the time of data collection. Due to the inherent sensitivity of this topic, patient consent and institutional review board clearances were sought to avoid vagrant ethical issues and to maintain the patients' confidentiality.

5. MODEL SELECTION

The use of suitable models for prediction of Gastroesophageal Reflux and Smoking status in asthma patients using the characteristics of data set and goals are as follows. The following models were chosen for their suitability in binary classification tasks and interpretability:The following models were chosen for their suitability in binary classification tasks and interpretability:
Logistic Regression
Description: Logistic Regression is a Linear method used for predicting the probability of occurrence of a Sure Thing.
Reason for Selection: The logistic regression model is frequently utilized for binary classification problems and gives coefficients to explain the significance of the features.

Decision Trees
Description: It is noteworthy that Decision Trees classify the feature space into regions in order to foresee the target variable.
Reason for Selection: Decision Trees, therefore, are capable of identifying non-linear relationships as well as interaction between the variables; characteristics that will well suit a case involving asthma, its cause and its preventions.

K-Nearest Neighbors (KNN)
Description: It means that KNN classifies the given data points based on the majority class of their neighbors.
Reason for Selection: KNN does not assume any distribution of the data and therefore can handle data in the dataset of unequal variability and of varying degrees of complexity.

AdaBoost Classifier
Description: AdaBoost is method of boosting that integrates a number of weak classifiers into a single strong classifier.
Reason for Selection: AdaBoost is useful in handling oversampled instances and reduces the misclassification rate by under sampling the samples that are difficult to classify.

Naive Bayes Classifiers
Description: Naïve Bayes' models apply Bayes' theorem with stipulated independence between the features.
Reason for Selection: The strategy of Naive Bayes is computationally fast or efficient especially when the data is of high dimensionality and when the assumption of independence is reasonably valid.

The chosen models allow to extend the applications of a given set of factors and characteristics that relate to asthma, and provide different benefits and perspectives for the prediction of the target variables. The use of models guarantees a variability of the approach to the representation of the severe facet of asthma disease and generates easily comprehensible results for clinicians and researchers' additional studies.

6. IMPLEMENTATION

1. **Data Preprocessing**

Handling Missing Data: Data missing in the dataset was identified and treated by either imputation or deletion depending with the effect it would have in the analysis and the model.

Encoding Categorical Variables: On nominal variables, a label encoding method was used to convert categorical variables into numerical format that can be used in a machine learning model.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 8 | August 2024 - Peer Reviewed Journal

Handling Outliers: Z-score analysis was used to figure out outliers and they were given some consideration in order that them would not prejudice the evaluation and training of the model.

$$z - score = \frac{x - mean}{standard\ deviation}$$

2. Feature Selection and Engineering

Correlation Analysis: In this study, the independent variables were chosen purposively depending on their association with target variables, namely the Gastroesophageal Reflux and Smoking status and their relationship with factors related to asthma.

Synthetic Minority Over-sampling Technique (SMOTE): To help balance the classes within the target variables, SMOTE algorithm was used to open up new perspectives in generating synthesised data set by constructing new samples of the minority class thus making sure that the actual training set contained a balance of the two major classes.

3. Model Training and Evaluation

Train-Test Split: The data was divided into 80% training set and 20% testing set so that the proposed models could be trained on the training data and tested on new data.

Model Selection: Each classifier from the Logistic Regression, Decision Trees, K-Nearest Neighbors, AdaBoost, Naive Bayes classifiers was written and tested in the scikit-learn environment using such parameters as accuracy, precision, recall, F1 score, ROC AUC, among others.

Cross-validation: To make results more reliable, procedures like the k-fold cross-validation were also applied.

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TN}{TN + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP=True Positive
TN=True Negative
FP=False positive
FN=False Negative

4. Model Interpretation and Visualization

Evaluation Metrics: To this tune model performance metrics were computed and analysed with a view of establishing which are the most accurate models of the two with intent of predicting Gastroesophageal Reflux and Smoking status of the asthma patients.

Visualizations: All models used for this study: ROC curves, precision-recall curves, confusion matrices and feature importance plots were plotted using matplotlib and seaborn to compare and evaluate the model's performance and its insights.

5. Reporting and Conclusion

Results Interpretation: Evidence from the models was in essence described and explained to get directions for the links between asthma relevant factors and the target variables.

Clinical Implications: In light of these studying results, the recommendations for clinical practice were mentioned, such as individualized approach to asthmatic patients' treatment and their management.

Limitations and Future Directions: Recommendations on future studies to improve the model's performance, increase the variety of the dataset, and study more characteristics that would have an impact on the item's success were provided.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 8 | August 2024 - Peer Reviewed Journal

7. RESULT

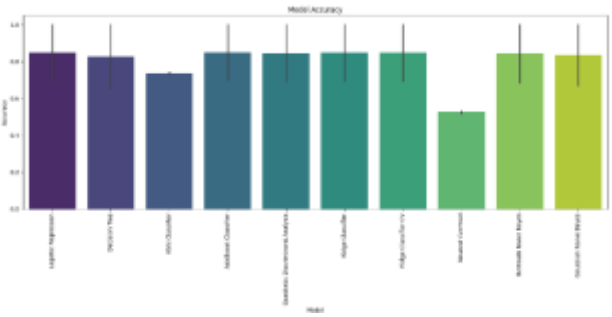

Fig 1.0 Accuracy Graph of Used models

In bar chart below, different classification model such as Logistics Regression, Decision Tree, KNN Classifier, AdaBoost Classifier, Quadratic Discriminant Analysis, Ridge Classifier, Ridge Classifier CV, Nearest Centroid, Bernouli Naive Bayes and Gaussian Naive Bayes are illustrated in regards to their accuracy levels. The accuracies vary from around 0. 7 to 1. 0, out of which the highest accuracy near 1 is achieved by Logistic Regression, Decision Tree, and AdaBoost Classifier. 0. The accuracy is generally great, though not static across the several models; Nearest Centroid recorded the lowest percentage of accuracy among all the models.

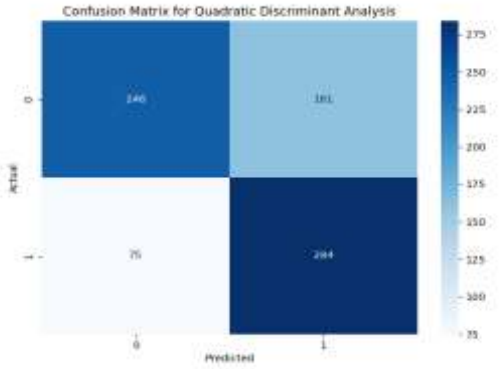

Fig 2.0 Confusion matrix for Quadratic Discriminant Analysis

Confusion matrix of the Quadratic Discriminant Analysis model. It displays the actual number of true positive and false positive cases together with those of true and false negative cases. In detail, the model included 246 true negatives of class 0 and 284 true positive of class 1. Basically, it scored high since 88 of the predictions were accurate but, it had 161 of class 0 as class 1 and 75 of class 1 as class 0. This confusion matrix proves useful as it shows areas that the model could be improving on misclassifications.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score	Matthews Corr Coef	Jaccard Score
Logistic Regression	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Decision Tree	1.00	1.00	1.00	1.00	1.00	1.00	1.00
KNN Classifier	0.74	0.78	0.74	0.73	0.84	0.52	0.58
AdaBoost Classifier	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Quadratic Discriminant Analysis	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Ridge Classifier	1.00	1.00	1.00	1.00	N/A	1.00	1.00
Ridge Classifier CV	1.00	1.00	1.00	1.00	N/A	1.00	1.00
Nearest Centroid	0.53	0.53	0.53	0.53	N/A	0.06	0.36
Bernoulli Naive Bayes	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Gaussian Naive Bayes	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Fig 3.0 Performance Evaluation of different classifiers

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 8 | August 2024 - Peer Reviewed Journal

Performance evaluation of different classifiers as part of classification techniques. That is features like accuracy, precision, recall, F1 score, ROC AUC score, Matthews correlation coefficient, as well as Jaccard score with respect to each model. Most of the models including Logistic Regression, decision tree, AdaBoost classifier, as well as both the types of Naive Bayes yield an accuracy of 1 for all the measures, thereby showing the good performance of the model. Yet the features of the KNN Classifier and respectively the Nearest Centroid models are depicted as less effective since their scores are lower in some of the mentioned metrics. The cars' performances are compared in a tabular format which is important when selecting a model for a specific task.

8. CONCLUSION

From the analysis of classification models such as Logistic Regression, Decision Tree, KNN Classifier, AdaBoost Classifier, Quadratic Discriminant Analysis, Ridge Classifier, Ridge Classifier CV, Nearest Centroid, Bernoulli Naive Bayes and Gaussian Naive Bayes, one can deduce that most of models have high values on the accuracy, precision, recall and f1-score. Out of all the classifiers, Logistic Regression, Decision Tree and AdaBoost Classifier are closest to perfection with accuracies showing 1. 00, which means the models shown high performance in terms of the filtered evaluation measures like precision, recall, F1 test, ROC destructive ester, Matthews correlation density, and Jaccard ostentation.

On the other hand, models such as Nearest Centroid has comparatively lower values for accuracies which indicates that there are some shortcomings in terms of the prediction of the models than the classifiers. Quadratic Discriminant Analysis comes with a confusion matrix and a closer look shows what Quadratic Discriminant Analysis does well, and where it requires enhancement. It classified 284 instances of class 1 and 246 instances of class 0 while at the same time misclassifying 161 instances of class 0 as belonging to class 1 and classifying 75 instances of class 1 as belonging to class 0. This is an indication that more attention needs to be paid to fine-tuning or possibly exploring other models particularly when accurate categorization is desirable.

Thus, the performance evaluation of the many classifiers and the analysis of their characteristics proved that it is crucial to choose a right model for a certain task. It is evident cut the overall performance of models such as, Logistic Regression, Decision Tree, and AdaBoost Classifier is high, and therefore proving to be ideal for challenges that require accuracy and credibility. But, it is imperative that each model has its pros and cons as confirmed through metrics such as confusion matrices, and each score that is in practice used in classification techniques.

9. REFERENCES

1. *Grammar Local for Asthma (GINA). Worldwide Plan of Action in Asthma Treatment and Control. Available online: https://ginasthma. International Energy Agency, IEA Org, available at https://www. iea. org/ accessed on 1 July 2024.*
2. *CDC/Center for Preparedness and Response to Emerging Infectious/Division of Global Migration and Quarantine. Asthma Statistics and Age Demographics of Asthma. Available online: https://www. cdc. gov/asthma/asthma_stats/index. Our next source is a the Scottish Centre for the Book that found in a searchable database which can be accesses via the following link; http://snowball. scp. ic. ac. uk/scbpapers/CDSCR/NewIndex/Site/ (and accessed on 1 July 2024).*
3. *National H-eart, Lung, and Blood Institute (NHLBI). Asthma. Available online: https://www. nhlbi. nih. health-topics/asthma Available from http://www. who. int [Web version, accessed on 1 July 2024].*
4. *American Academy of Allergy, Asthma & Immunology, AAAAI. Asthma Statistics. Available online: https://www. aaaai. org/conditions-and-treatments/asthma/statistics (visited on 1st of July 2024)*
5. *Pavord I, Haldar P, Shaw D, Berry M, Thomas M, Brightling C, Wardlaw A, Green R. Identification of clinical subtypes by cluster analysis and disease phenotypes in asthma. Am J Respir Crit Care Med. 2008 Sep 1;178(5):They mobilise for war over 218-24. doi: 10. 1164/rccm. 200711-1754OC.*
6. *Sutherland ER, Goleva E, Jackson LP, Stevens AD, Leung DY. Asthma in adulthood, Vitamin D status, lung function, and steroid responsiveness. Am J Respir Crit Care Med. 2010 Mar 15;181(6):The end portion of the current treaties is defined as 599–606. doi: 10. 1164/rccm. 200911-1710OC.*
7. *Wood LG, Powell H, Gibson PG. Mannitol challenge for determination of airway sensitivity, airway inflammation, and inflammatory subtype of asthma. Clin Exp Allergy. 2010 Mar;40(3):232-41. doi: 10. 1111/j. 1365-2222. 2009. 03410. x.*
8. *von Bülow A, Kriegbaum M, Backer V, Porsbjerg C. The prevalence of severe asthma and low asthma control among Danids Adults. Published in the Journal of Allergy and Clinical Immunology: Practise. 2014 May-Jun;2(3):759-67. e1. doi: 10. 1016/j. jaip. 2014. 05. 008.*
9. *Wu AC, Tantisira K, Li L, Fuhlbrigge AL, Weiss ST, Litonjua A. Effect of vitamin D and inhaled corticosteroid treatment on lung function in children. Am J Respir Crit Care Med. 2012 May 15;185(10):The regime of Edward I (1276-82). doi: 10. 1164/rccm. 201111-2030OC.*
10. *Moore WC, Meyers DA Wenzel SE, Teague WG, Li H, Li X, D'Agostino R Jr, Castro M, Curran-Everett D, Fitzpatrick AM, Gaston B. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med. 2010 Jul 15;181(4):315-23 To. doi: 10. 1164/rccm. 200906-0896OC.*
11. *National Asthma Council Australia. Australian Asthma Handbook. Available online: https://www. asthmahandbook. org. au/ (last accessed, 1 July 2024).*
12. *Liu AH, Zeiger R, Sorkness C, Mahr T, Ostrom N, Burgess S, Rosenzweig JC, Manjunath R. Progress in development of the Childhood Asthma Control Test. J Allergy Clin Immunol. 2007 Sep;120(3):553-9. doi: 10. 1016/j. jaci. 2007. 06. 039.*

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 8 | August 2024 - Peer Reviewed Journal

13. *Peters SP, Ferguson G, Deniz Y, Reisner C. Uncontrolled asthma: a research on the current incidence rates, disease frequencies, and possible management approaches. Respir Med. 2006 Jul;100(7):1139-51. doi: 10. 1016/j. rmed. 2005. 10. 011.*
14. *Gibson PG, McDonald VM. Management of severe asthma: to the airways, and patients' comorbidities, as well as historical and genetic factors. Intern Med J. 2017 Jul;47(7):The analysis of the fragments of the philosophical commentaries to the Lamentations and Song of Songs, which refer to the ''carnal screw'' (erez ' Shawnem basar), is methodologically grounded in the bibliographical research and historical analysis of previous studies and is dated back to the 6th–7th centuries, 623–631. doi: 10. 1111/imj. 13452.*
15. *Lötvall J, van Aalderen WMC, Sastre J, Bousquet J, Anto JM, Haahtela T, Lau S, Akdis M, Almqvist C, Alvaro-Lozano M, Nogues-Rubin P, Bateman ED, Bel Moreno L, Bleecker ER, Boulet LP, Bousquet PL, Bro Asthma endotypes: Patient sample A new classification of disease entities within the asthma syndrome. J Allergy Clin Immunol. 2011 Feb;127(2):355-60.*