Identifying Subgroups of Early and Late Onset Asthma and Determining Their Corresponding Risk Factors

CIDs: 02112494, 01613484, 02183162
Word Count: 3909

School of Public Health, Imperial College London, England

Abstract

Objectives

This study aims to identify subgroups of individuals with early and late-onset asthma then explore the aetiology and the associated risk factors with each group. This will be achieved by looking at two questions.

1. What are the risk factors and comorbidities associated with early and late-onset asthma?
2. Are there any distinguishable asthma subgroups within individuals with early and late-onset asthma, and what are their associated risk factors?

The study has important implications for understanding the aetiology and the pathology of early and late-onset asthma and will benefit targeted interventions and screening for specific subgroups of people with early and late asthma.

Methods

The analysis takes place on data collected for the UK Biobank (UK Biobank, 2022). It focuses on the 15,491 individuals with doctor-diagnosed asthma. Individuals were classed as early-onset asthma if they received a diagnosis before the age of 12, and those diagnosed over the age of 12 were deemed as late-onset asthma. Lasso regression was performed with stability correction on different asthma datasets to screen key predictor variables, which were validated with random forests. Clustering was performed with K-Prototype on individuals with both early and late-onset asthma. Three distinct subgroups were identified within each. Regressions and random forests were then carried out to assess the importance of predictors in differentiating between the clusters. The comorbidity analysis was done using Chi-square test on disease prevalence.

Results

Several risk factors of early and late-onset asthma were identified. In early-onset asthma, 6 risk factors were identified, the most impacting were peak expiratory flow (PEV), forced expiratory volume in 1 second (FEV1), sex and the severity of asthma (severity). In late-onset asthma, 45 predictors were identified, the most impacting were those in early-onset asthma as well as the individuals sex.

Co-morbidity analysis revealed common chronic diseases, cancer, autoimmune diseases, anaemia and digestive disorders as unique or shared comorbidities for early and late asthma. Late asthma is more complex and also has more unique comorbidities such as hypertension, obesity and hypothyroidism.

Within both early and late-onset asthma, three different subgroups were identified. For early-onset asthma: Group 1 was made up of those with good lung function, mild asthma, little to no medication and were predominantly male; Group 2 had an intermediate level of lung function but still mild asthma, with some history of asthma medication usage; Group 3 were slightly older on average, had severe asthma and were slightly more likely to be female.

For late-onset asthma, the three subgroups found are as follows: Group 1 had moderate lung function measurements and mostly has mild asthma; they were primarily female and had higher levels of income; Group 2 had good lung function and mild asthma; they were primarily male and had lower levels of income and chemical exposures; Group 3 had the worst lung function measures with severe asthma, which is predominantly older women with a history of smoking.

Conclusion

The main risk factors of early and late-onset asthma are the same, with more unique acquired contributing factors and comorbidities being identified in late-onset asthma. Three subgroups can be identified within early and late-onset asthma, which are mainly differentiable by lung function, severity and gender. Late asthma is the main form of asthma with more complex comorbidities and environmental risk factors identified, and the sub-types are more practical to targeted intervention and treatment.

Keywords: Asthma, Severity, Onset

1. Introduction

Asthma is a major noncommunicable disease (NCD) of the lungs, commonly affecting both children and adults. Main symptoms include inflammation and narrowing of the small airways in the lungs, which can cause breathing difficulties, it affects people of all ages and often starts in childhood (World Health Organisation, 2021; National Health Service, 2021). There are certain 'triggers' that can make the symptoms worse. Triggers vary between individuals but can include viral infections and diseases, dust, smoke, fumes/chemicals, weather changes or location.

In the UK, 8 million individuals have been diagnosed with asthma making it more prevalent than all other lung diseases combined. However, many children diagnosed with asthma will have no symptoms as adults, meaning not everyone diagnosed actively receives treatment, with 5.4 million people currently receiving treatment for the disease in the UK. While incidence rates of asthma in the UK have decreased in recent years, around 160,000 a year are diagnosed (British Lung Foundation, 2012).

Several previous papers have investigated the differences between early and late-onset asthma, and many reach similar conclusions. Early-onset asthma is more severe than late-onset asthma, and late-onset asthma is predominately among females. However, many of these studies might not be reliable for use in the UK due to the use of small or old datasets or being based exclusively on populations outside of Europe.

This study has two aims. Firstly to identify risk factors and comorbidities of early and late-onset asthma. Secondly, to identify subgroups of individuals within the UK with early and late-onset asthma and explore the aetiology and the associated risk factors with each group. This will aid in understanding and treating asthma by providing more individualised risk factors for each individual and allowing for more targeted screening and personalised treatment.

2. Methods

2.1 Data Processing

The data used was from the UK Biobank, which contains information on 501,635 individuals aged 40-79.

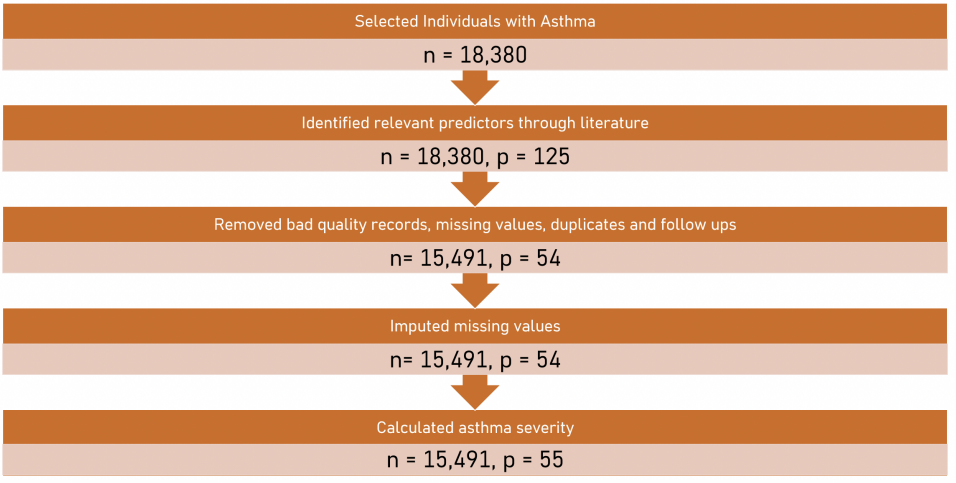

Figure 1. The process of data being selected and cleaned

From the dataset, 18,380 had a doctor's diagnosis of asthma. One hundred twenty-five relevant variables were identified through a literature search and selected for this analysis. These include lung function metrics, blood pressure readings, demographics, granulocytes, diagnoses and others. Only the readings taken at baseline were used, and variables with high levels of missingness or containing duplicate information were removed (71 columns). Around 1000 values for PEF and FEV were imputed using multiple imputations, and the average was taken. The distribution of values in the rows containing missingness were analysed and it was deemed that the values were missing at random. Asthma severity was then classified for each individual using measures of lung function, in line with previous research and the generally accepted academic standard (POLLART, 2009); this classifies values into 'Mild Intermittent' (n = 9733), 'Mild Persistent' (n = 730), 'Severe Intermittent' (n = 4405) and 'Severe Persistent' (n = 623) based upon cut-off values and lung function measurements. Due to a lack of results in the 'Mild Persistent' group and the 'Severe Persistent' group, the two mild groups merged to form the category 'Mild', and the two severe groups merged to form the category 'Severe'.

2.2 Exploratory Data Analysis

Figure two shows the distribution of individuals by asthma onset and severity. The majority of participants (11901) have late-onset asthma, while the remaining 3590 have early-onset asthma. Of those with early-onset asthma, 2256 have mild asthma while 1334 have severe asthma, while out of those with late-onset asthma, 8207 have mild asthma and the reaming 3864 have severe asthma. After preprocessing, the dataset contained 15,491 rows and 55 predictors.

The distributions of normalised Tiffeneau-Pinelli index by asthma onset show a fairly similar distribution. The Tiffeneau-Pinelli index is FFE1 to the full forced vital capacity (FVC). The lower the value, the higher the chance an obstruction to the lungs is present, preventing air from escaping. The mean for those with early-onset asthma is 1.044 with a standard deviation of 1.0976, while those with late-onset asthma have a mean of 0.715 and a standard deviation of 1.056. A t-test was conducted and showed that the means are not equivalent with a p-value, $p < 2.2 \times 10^{-16}$. This

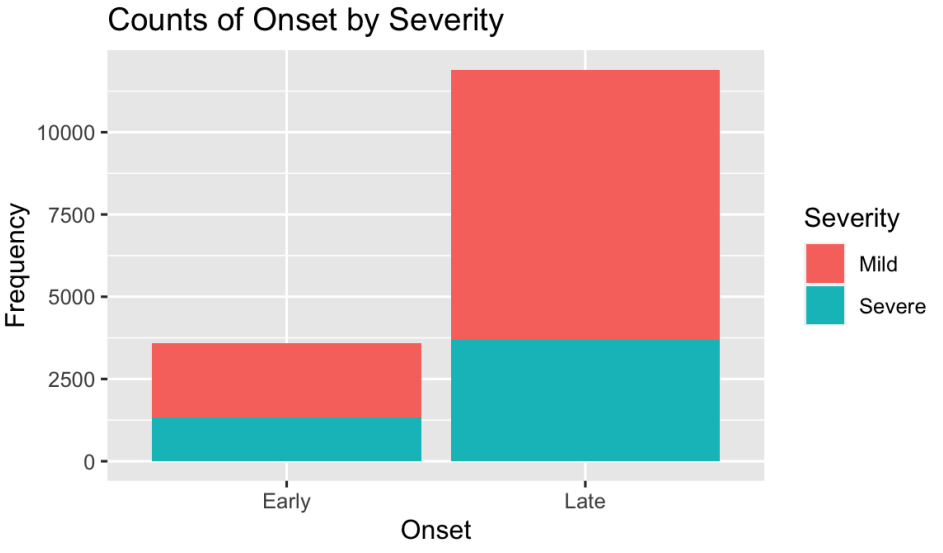

Figure 2. The distribution of onsets by severity

suggests that those with late-onset asthma are likely to have more obstruction to their lungs than the early-onset group on average.

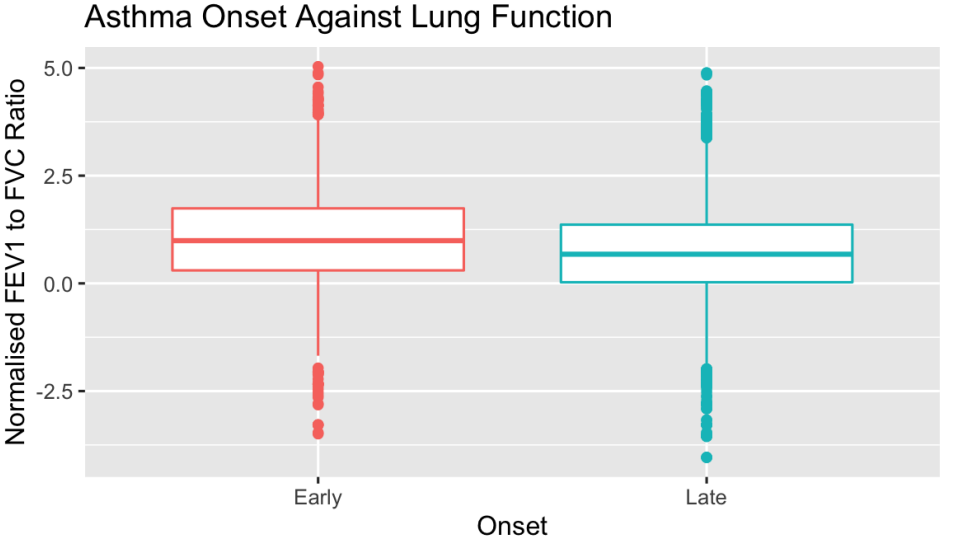

Figure 3. The distribution of onsets by severity

Several variable selection methods were carried out to identify which of the 55 variables should be used for the analysis. A Lasso regression showed the most important variables in both early and late asthma onset using the number of clusters as the predictor. This was taken further with stability selection using a 5-fold cross-validation logistic Lasso model and a random forest as a sensitivity test.

2.3 Clustering

Cluster analysis was performed three times: once on the whole dataset, once on those with early-onset asthma, and then on those with late-onset asthma. This facilitates the ability to determine subgroups within early and late-onset asthma and to show their unique characteristics. The K-prototype algorithm, which aims at combining continuous and categorical variables together to predict clusters, was used to facilitate categorical predictors such as gender and ethnic background. K-Means clustering was also used for the sensitivity analysis and provided similar results. The Elbow and Silhouette methods were applied to select the optimal number of clusters. The Silhouette suggested that 2 clusters are optimal, and 3 clusters are almost as good. The Elbow plot suggested that 3 is the best choice, therefore cluster number k=3 has been chosen as a balance of the two methods (Figure 16 in the Appendix). Variables that distinguished the characteristics of the different clusters ($p<0.05$) were screened as class identifiers. Random forests were then run to determine the main differentiating predictors of the clusters.

2.4 Comorbidity Analysis

We screened for valid ICD-10 diagnostic records corresponding to asthmatic patients in the UK Biobank, and mapped ICD-10 codes to valid Phecodes to homogenise the outcomes at the phenotypic level. We then created two substs within early and late asthmatics: those diagnosed according to ICD-10, and those undiagnosed asthmatics. The Phecode datasets represent a comorbidity corpus, and the high-frequency words will obey a power-law distribution, with a few high-frequency words representing the overall profile of comorbidity information. We selected the top 15 high-frequency co-occurring phenotypes from each of the two subgroups, after which their intersection was taken and the prevalences of these phenotypes were counted. Differences in prevalence were verified by a FDR-corrected Chi-square test at 0.05 level.

3. Results

3.1 What are the risk factors of early and late asthma?

Table 1 shows Lasso models with stability selection on the imputed datasets. This was done to select models and provide a measure of variable importance. A complete case analysis was also run, which provided similar results when using stability selection, strengthening these results.

The first model run was to predict asthma among the 15,491 with asthma and 15,491 without asthma. The controls were selected randomly from those without asthma in the UKBiobank and processed similarly to those with asthma. The primary demographics of the two groups were not matched but were checked and were equivalent. The Lasso selected 28 variables and indicated that the most significant were asthma, number of medications taken, FEV1, FVC ratio, sex, neutrophill percentage, eosinophil count and current tobacco smoking. The random forest selected 44 predictors and achieved 69.95% accuracy. The important predictors indicated were FEV1 to FVC ratio, eosinophil percentage, and doctor-diagnosed hay fever or allergic rhinitis.

The next model produced was to predict asthma onset among the 15,491 asthmatics. The Lasso selected 11 variables which were lung function measures, systolic blood pressure, sex, smoking habits and location lived. The random forest selected 41 variables and achieved an accuracy of 61.97%. The most important predictors were various measures of lung functions. Certain other predictors such as eosinophil count and BMI were also flagged as important.

Table 1. Lasso with Stability Selection and Random Forests on data with imputations

	Lasso		Random Forests	
Data	AUC	Num. of predictors	AUC	Num. of Predictors
Predicting Asthma in cases and controls	79.40	28	69.95	44
Predicting Onset in asthmatics	82.5	11	61.97	41
Severity in Early Onset Asthma	78.5	6	*55.72	40
Severity in Late Onset Asthma	76.5	45	*53.51	52

*Models ran without lung function measures

Two more Lasso models and random forests were run to predict severity, one for early-onset asthma and the other for late-onset asthma. For early-onset asthma, the risk factors identified as important by the Lasso were found to be PEF, best FEV1 from 3 back to back readings, FEV1 pred, minimum PEF from 3 back to back readings, severity and sex. In total, 6 predictors were identified with an accuracy of 78.5%. The random forest was run without the measures of lung function to explore further non-causal (severity is defined by lung function) associated factors. The random forest achieved an accuracy of 55.72% and suggested that the main predictors were age and systolic blood pressure. For late-onset asthma, 45 predictors were identified, with the major risk factors being lung function metrics, age, neutrophil, severity and sex with an accuracy of 76.5%. The random forest accuracy was 53.51% and selected housing and eosinophil count. For the full variable importance of each model, see the Appendix.

Comorbiditiy Analysis

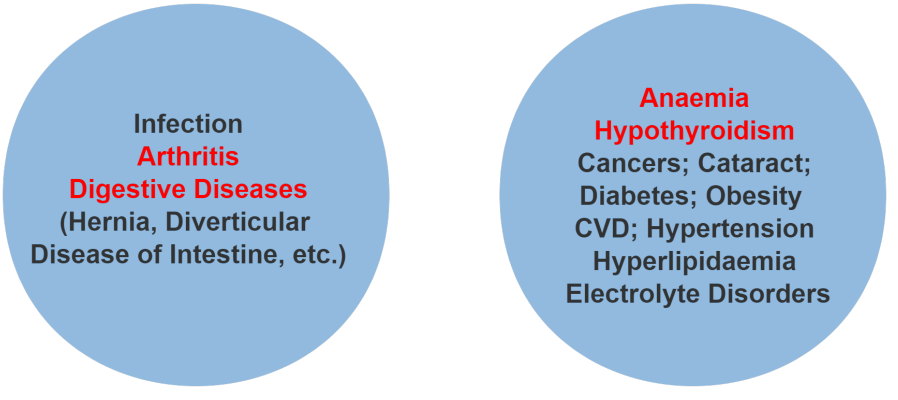

Figure 4. Unique Comorbidities in Diagnosed (Left) and Undiagnosed (Right) Early Asthmatics

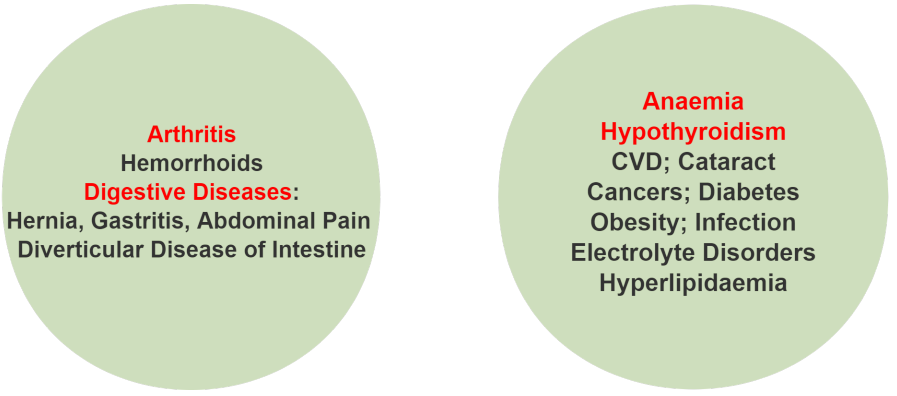

Figure 5. Unique Comorbidities in Diagnosed (Left) and Undiagnosed (Right) Late Asthmatics

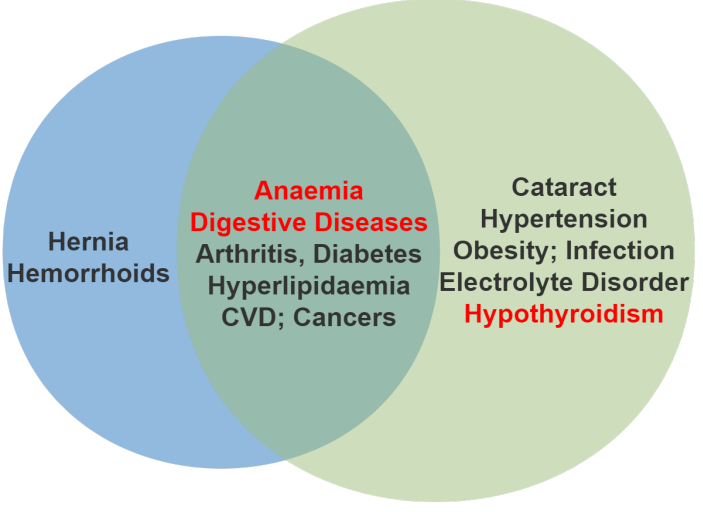

Figure 6. Unique and Common Comorbidities of Early (Left) and Late (Right) Asthmatics

Figure 4 (and Table 8 in Appendix) shows that the early asthma patients with no documented asthma diagnosis by ICD-10 codes had a higher prevalence of common chronic diseases. This is potentially because diseases associated with the asthma phenotype, such as hypothyroidism, hypertension, hyperlipidaemia, obesity, diabetes and anaemia, were preferentially identified in the undiagnosed group. The undiagnosed group is also a more subnormal and unhealthy group with a higher prevalence of cataracts, cardiovascular disease and cancers. In contrast, in the diagnosed group, a higher prevalence of arthritis is observed and high comorbidity of multiple digestive system disorders. Both asthma and arthritis are likely to involve the immune system, and they both inhibit physical activity and adversely affect life quality, making their co-occurrence plausible. The previously reported relationship between digestive system disorders and asthma, for example, the high prevalence of acid reflux disease in patients with severe asthma (Leggett, 2005), was further confirmed by our study. It

could be that the relationship between asthma and digestive disorders may be bidirectional. On the one hand, digestive disorders can contribute to the onset and progression of asthma through both neurological and humoral pathways (airway constriction, acidic irritation, mucosal destruction); on the other hand, the increased digestive burden of asthma (pressure changes from cough and oedema) can also trigger the development of digestive symptoms. This instructs to increase screening for asthma in sub-healthy groups and people with digestive and immune system disorders.

Figure 5 (and Table 9 in Appendix) presents similar findings: among patients with late-onset asthma, the undiagnosed group had a higher prevalence of common chronic diseases, and the diagnosed group had a higher prevalence of digestive diseases. Additionally, it was noted that the prevalence of both general anaemia and iron deficiency anaemia was higher in the undiagnosed group, suggesting that advanced asthma is likely to be associated with red blood cell abnormalities. As previously reported (Morris, 2009), the co-occurrence between asthma and red blood cell diseases such as sickle cell disease is reflected as the anaemic phenotype after Phecode mapping.

Figure 6 shows common and different comorbidities in early and late asthma. The prevalence of hypothyroidism showed significant differences in early and late asthma, regardless of ICD-10 grouping (Table 13 in Appendix). It is critical that Hypothyroidism had a higher prevalence in late asthmatics. Since most hypothyroid patients have Hashimoto's Thyroiditis and have abnormal autoimmune functions, this may reveal a potential aetiology in late-onset asthma: late-onset asthma is more often autoimmune related (Bingyan, 2019). Since late asthmatics experienced more adulthood exposures and their onset was more related to environmental factors, the prevalence of general disorders such as obesity and electrolyte and fluid balance disorders also showed reasonable differences in early and late-onset patients. Table 5 shows that the common comorbidities found in the analysis in Tables 2 and 3 also had common overlaps in patients with early and late-onset asthma. Table 10-12 in Appendix then detailedly summarises asthma comorbidities, and their correlations with the presence or absence of the ICD-10 asthma diagnosis. Common chronic diseases, cancer, autoimmune diseases, anaemia and digestive disorders are all comorbidities shared by patients with early and late asthma.

3.2 Are there any distinguishable asthma subgroups within individuals with early and late-onset asthma, and what are their associated risk factors?

Figure 7 illustrated the three subtypes of early-onset asthma and the corresponding critical variables. All identifiers were used and shown in Figure 19 in Appendix: Lung function indicators (FVC, FEV1 and PEF), participant age and age at onset of asthma, gender, medication, type of work and asthma severity are important exposure variables that distinguish early asthma subtypes. Based on this, three subgroups of early asthma can tentatively be defined: the first subgroup of early asthmatics has an intermediate level of lung function, age, and medication history among all early asthmatics. Their asthma severity level is mostly mild. The second subgroup is the youngest subgroup, with the highest level of lung function, lowest medication level, and a high proportion of mild asthma, and is mostly male. The third subgroup is the oldest, with the poorest lung function measures and the most severe asthma symptoms, with a slightly higher proportion of females.

Similarly, Figure 8 summarises the key exposure variables that distinguish the subtypes of late asthma. In addition to the indicators found in early asthma, acquired exposure variables were observed, such as chemical exposure in the work environment, average household income, and past tobacco smoking (as shown in Figure 20 in the Appendix). These lifestyle and environmental factors may predispose to late asthma. Three subgroups of late asthma were then tentatively defined: the first subgroup is of moderate age and lung function level, having mostly mild asthma and is mostly

female, and having a higher income level. The second subgroup is mostly male, with lower income than the other two subgroups and works with chemical exposure. They have the highest level of lung function and have mostly developed mild asthma. The third subgroup, the most severe asthma subgroup, is mostly elder women with severely impaired lung function. They almost all had a history of smoking. As late-onset asthma is the predominant type of asthma and environmental factors are more relevant to its onset, these findings can guide targeted screening and prevention of asthma and point out the sensitive population. For example, men with chemical exposures in the work environment and older women with a smoking history should be well-targeted and taken care of. Figure 17-18 in the Appendix showed the sensitivity analysis on the merged data and returned similar results. Then an adjusted multivariate-multiple logistic regression was fitted to predict asthma clusters using these key predictor variables. The model's R-square values were both greater than 0.80. As these predictor variables do not all meet the logistic regression assumptions, random forests were used to validate these variables.

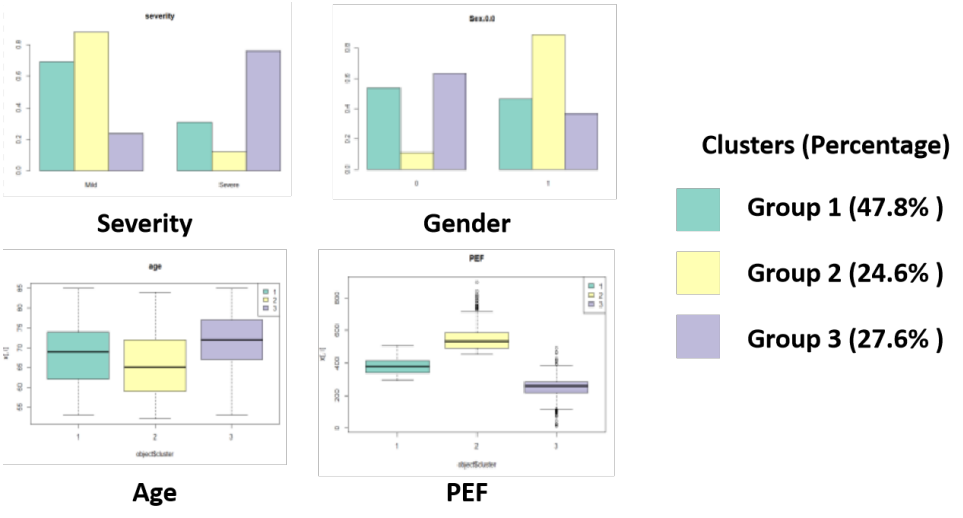

Figure 7. Key Variables in K-prototype Clustering for Early Onset Asthma Patients

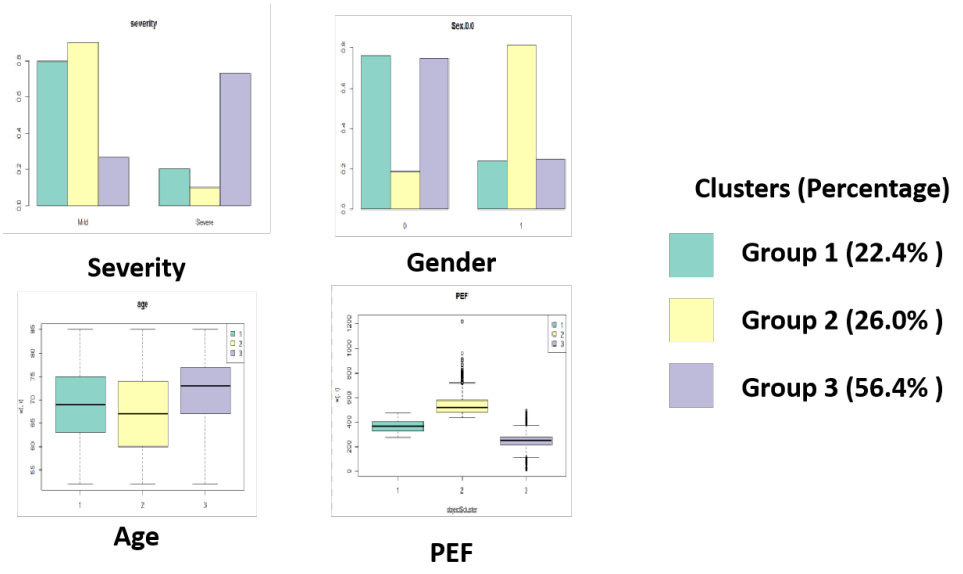

Figure 8. Key Variables in K-prototype Clustering for Late Onset Asthma Patients

Sensitivity Analysis

A random forest was conducted to corroborate further the variables that differentiate between the clusters of early and late-onset asthma.

Table 2. Random Forests run for a sensitivity analysis

		Reference		
	Early Onset	Cluster 1	Cluster 2	Cluster 3
Prediction	Cluster 1	674	3	1
	Cluster 2	10	709	0
	Cluster 3	22	0	703

	Late Onset	Cluster 1	Cluster 2	Cluster 3
Prediction	Cluster 1	675	6	2
	Cluster 2	3	684	0
	Cluster 3	21	0	706

For early-onset asthma, the model trained had a classification accuracy of 0.9847 (95% CI 0.9785, 0.9895), implying high accuracy. The model dictates that the most impactful predictors to differentiate between the clusters are the lung function measures: Severity, PEF, FEV and FVC. Variables including gender were flagged as relevant but had a much smaller impact. Since the lung function measures could be masking the impact of other variables, a second random forest was run without any measures of lung function (including severity), which achieved an accuracy of 0.5875 (95% CI 0.5661, 0.6087). Sex was deemed highly important, followed by current age. Many other predictors, including white blood cell and neutrophil count, had a small impact. These findings coincide with the previous results.

Late-onset asthma saw similar results. The model had a classification accuracy of 0.983 (CI 0.9766, 0.9881). The model dictates that the most impacting predictors are the lung function measures: Severity, PEF, FEV and FVC. Again, a model was made without lung function measures which had an accuracy of 0.5834 (95% CI 0.5621, 0.6045). Sex was deemed the most important predictor again. Again, these coincide with the results. However, they failed to identify smoking as an important predictor.

4. Conclusion

Several risk factors of early and late-onset asthma were identified. In early-onset asthma, PEF, FEV1, the severity of asthma were important, while late-onset asthma saw the same significant variables with the addition of age.

Through comorbidity studies, we have further identified the life burden brought about by asthma in early and late onsets. Late-onset asthma is a more complex disease, and the adulthood exposure experienced by patients may be associated with multiple co-occuring chronic diseases. Comorbidity analysis has also suggested potential pathological studies. The association of asthma with arthritis and hypothyroidism may reveal a potential role for the immune response in the comorbidity. The co-occurrence of asthma and anaemia, and asthma and digestive disorders also suggest that red blood cell abnormalities, and digestive stress may have crosstalks with the aetiology of asthma.

Within both early and late-onset asthma, three different subgroups were identified.

For early-onset asthma: Group 1 was made up of those with good lung function, mild asthma, little to no medication and were predominantly male; Group 2 had an intermediate level of lung function but still mild asthma, with some history of asthma medication usage; Group 3 were slightly older on average, had severe asthma and were slightly more likely to be female.

For late-onset asthma, the three subgroups found are as follows: Group 1 have moderate lung function measurements and mostly has mild asthma; they are primarily female and have a higher level of income; Group 2 have good lung function and mild asthma; they are primarily male and have a lower level of income, they have a workplace with chemical exposures; Group 3 have the worst lung function measures with severe asthma, it is predominantly older women who have a history of smoking.

5. Discussion

This paper has succeeded in its goals of studying the risk factors and identifying unique subgroups within early and late-onset asthma. We have seen that early and late asthma have different comorbidities and risk factors, and late asthma is the main asthma subtype and also the more complex one: it could be associated with immune responses and acquired exposures as well as environmental factors. It was widely known that early-onset asthma was more severe than late-onset asthma, our study then narrowed this down to subtype level - only one subtype each of early and late-onset asthmatics is severe, so the previous view is inaccurate.

Further, we achieved the goal of phenotyping early and late-onset asthma and classified them into three subtypes accordingly. Based on these subtypes, we can suggest some useful indicators for targeted asthma management: e.g. gender, age, severity (and income level for late onset). They have all been validated in penalised regression, clustering and random forest.

This study enlightens us that asthma screenings should be considered in groups of sub-healthy

individuals, particularly those with digestive or immune disorders. This would allow for the potential of an earlier diagnosis, allowing symptoms to be treated before a flair up. This would also significantly improve the diagnostic accuracy of asthma. Also, targeted treatment of men with chemical exposure, or elder women with smoking history may also mitigate the burden experienced by the severely ill subgroups.

5.1 Limitations

There were some drawbacks to the analysis. Firstly, the dataset used was from the UK Biobank, which is prone to selection (Fry, 2017). This stems from 9.2 million people being invited to the study but a response rate of just 5.5% enrolling 501,635 individuals; sampling from the UK based cohort may also challenge the external validity of the study. Furthering this, many individuals missed questions, leading to missingness within the dataset. This meant that slightly more than 10% of all variables with missing values were imputed. While values appeared to be missing at random, this still introduces additional bias into the dataset.

We used 12 years of age as an indicator for the classification of early and late-onset asthma, which may have resulted in differences from existing studies. Furthermore, we used Phecode as an indicator for co-morbidity analysis, which ensures homogeneity of phenotypes, but some diseases may not have corresponding phecodes, causing potential bias. In the future, we could also use more comorbidity information (e.g. analysing multiple pairs or combinations of co-occurring diseases in addition to asthma), add more exposure variables, and use more means of sensitivity analysis (e.g. using multiple clustering methods) to validate our results. We also need further disease mechanism studies to confirm our findings on comorbidities and asthma subtypes.

This study also suffers from common drawbacks. There is a limited sample size of just 15,491 individuals, which is a very small proportion of asthma patients. The participants are all from the UK and currently aged 40-75, meaning that the results might not be generalised to those outside of England, Wales and Scotland and those not within that age bracket.

References

Bingyan, W., Z. Dong. 2019. Impact of thyroid hormones on asthma in older adults. *Journal of International Medical Research,* no. 47(9), 4114–4125.

British Lung Foundation. 2012. Asthma statistics. (accessed: 28.03.2022). https://statistics.blf.org.uk/asthma.

Fry, Anna. 2017. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *Oxford University Press,* no. 186, 1026–1034.

Leggett, J. et al. 2005. Prevalence of gastroesophageal reflux in difficult asthma. *Chest,* no. 127(4), 1227–1231.

Morris, C. R. 2009. Asthma management: reinventing the wheel in sickle cell disease. *American journal of hematology,* no. 84(4), 234–241.

National Health Service. 2021. Asthma statistics. (accessed: 28.03.2022). https://www.nhs.uk/conditions/asthma.

POLLART, SUSAN M. 2009. Overview of changes to asthma guidelines: diagnosis and screening. *American Family Physician,* no. 79, 761–767.

UK Biobank. 2022. (accessed: 28.03.2022). https://www.ukbiobank.ac.uk.

World Health Organisation. 2021. Asthma fact sheet. (accessed: 28.03.2022). https://www.who.int/news-room/fact-sheets/detail/asthma.

Appendix

Lasso Regression Graphs

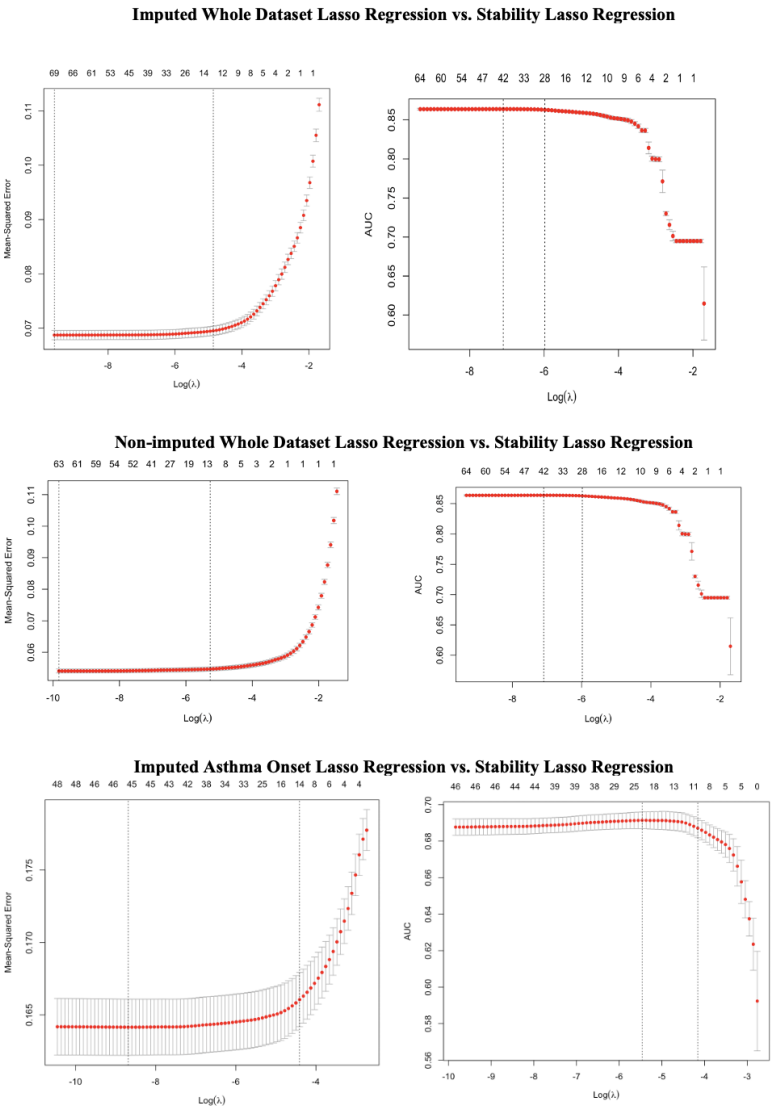

Figure 9. Lasso Regression With and Without Stability Analysis on Merged Asthma Dataset (with sensitivity analysis on imputation).

Co-morbidity Tables and Clustering Graphs

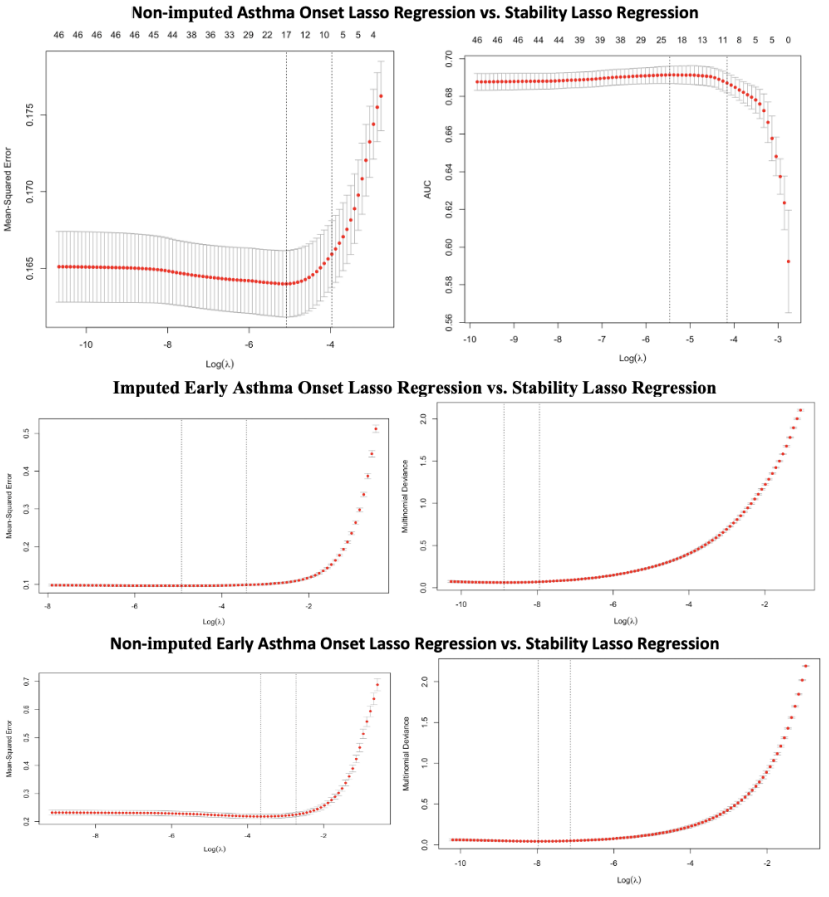

Figure 10. Lasso Regression With and Without Stability Analysis on Early Asthma Dataset (with sensitivity analysis on imputation).

Table 3. Lasso with Stability Selection and Random Forests on data without imputations

	Lasso		Random Forests	
Data	AUC	Num. of predictors	AUC	Num. of Predictors
Predicting Asthma in cases and controls	80.14	28	71.64	43
Predicting Onset in asthmatics	83.47	11	62.82	48
Severity in Early Onset Asthma	74.56	28	58.27	45
Severity in Late Onset Asthma	74.55	28	53.48	51

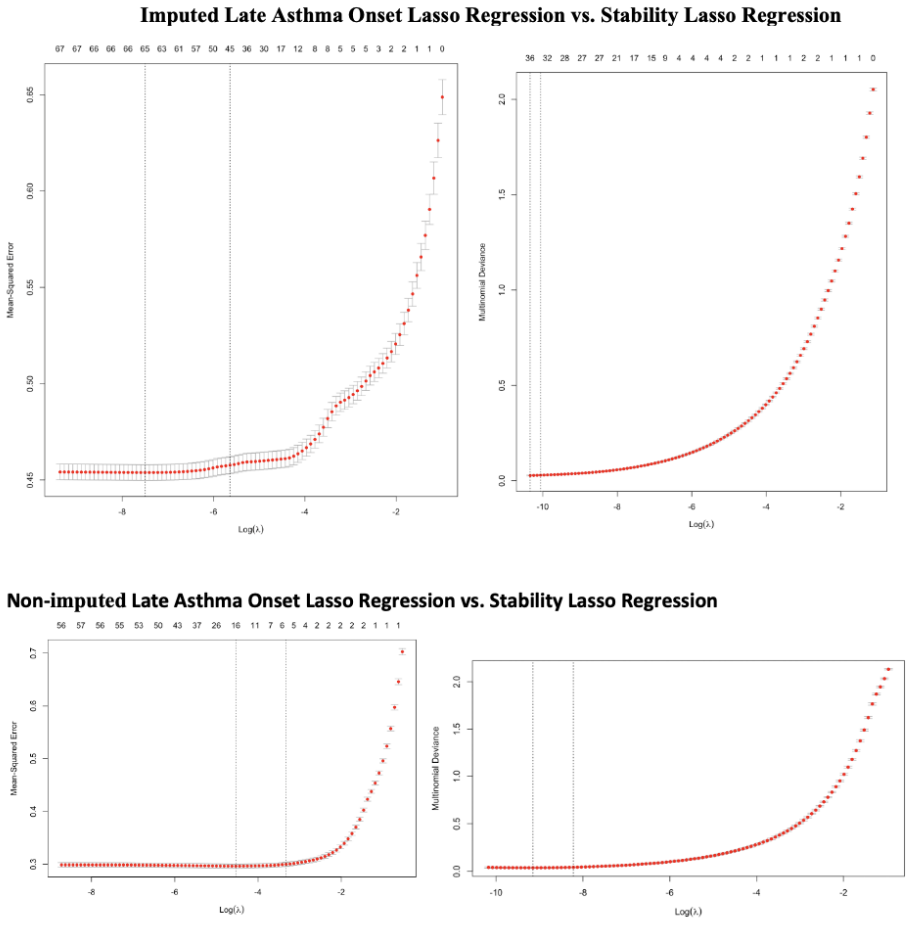

Figure 11. Lasso Regression With and Without Stability Analysis on Late Asthma Dataset (with sensitivity analysis on imputation).

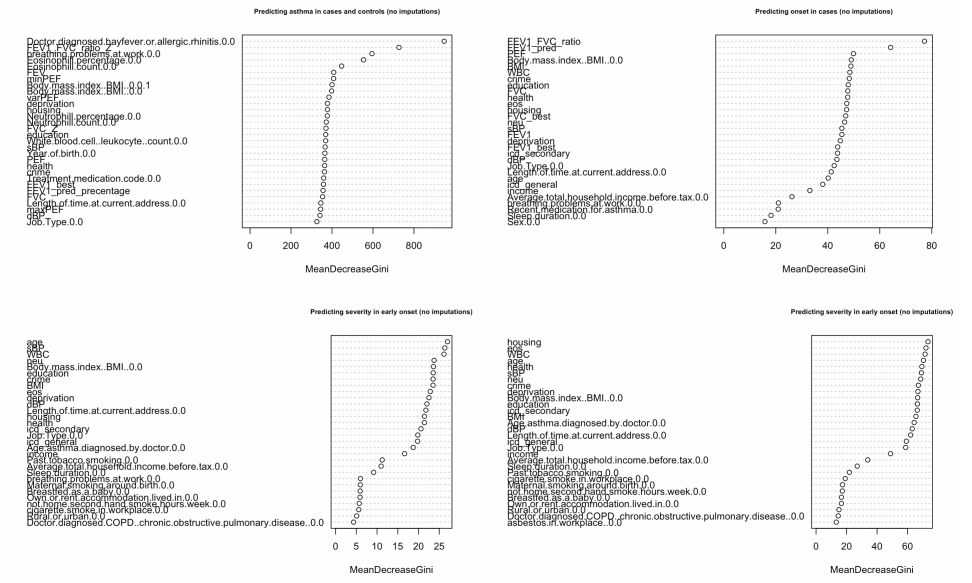

Figure 12. Important Predictors identified With Random forests on the complete cases

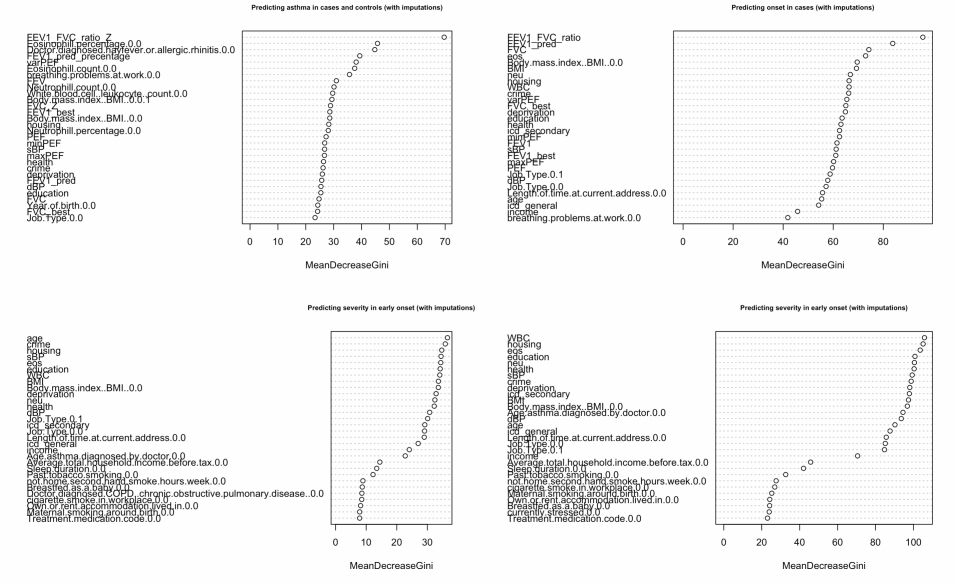

Figure 13. Important Predictors identified With Random forests on the imputed dataset

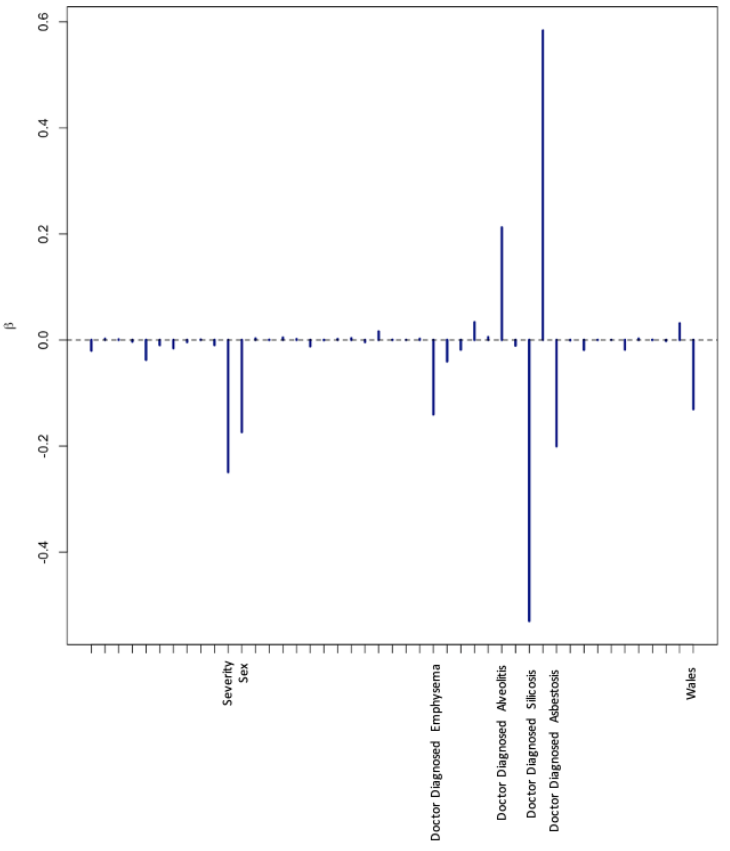

Figure 14. 30 Important Predictors identified With Lasso Regression on Late Asthma Imputed Dataset (Only important predictors shown, all other predictors identified in Table 8).

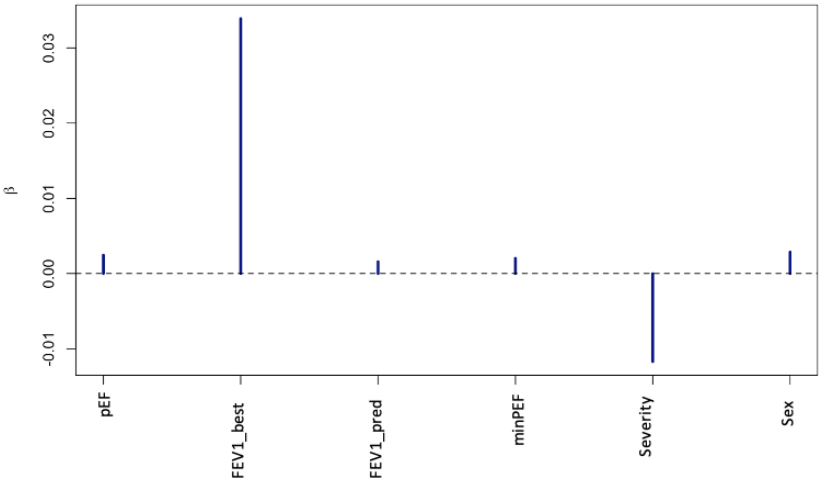

Figure 15. 6 Important Predictors identified With Lasso Regression on Early Vs. Late Asthma Imputed Dataset.

Table 4. 28 Predictors Identified from Predicting Asthma Onset in Cases and Controls With and Without Imputations (Using 5-fold Cross Validation Logistic Lasso Stability Selection)

Imputed and Non-imputed
asthma
Number of treatments medications taken
White blood cell leukocyte count
Eosinophil count
Neutrophil percentage
FVC
FEV
PEF
varPEF
FEV1_pred
FVC_Z
Sex
Own or rent accommodation lived in
Length of time at current address
Sleep duration
Current tobacco smoking
Genetic ethnic grouping
Doctor diagnosed emphysema
Doctor diagnosed chronic bronchitis
Doctor diagnosed sarcoidosis
Job Type
abestosos in workplace
health
education
housing
Scotland
Wales
icd asthma general

Table 5. Predictors identified from the Early and Late Onset Asthmatics with Imputed Dataset.

Late Onset	Early Onset
FEV1	pEF
Neu	FEV1_best
Eos	FEV1_pred
Age	minPEF
FEV1_best	severity
FEV1_pred	Sex
FEV1_FVC_ratio	
minPEF	
maxPEF	
varPEF	
Severity	
Sex	
Own or rent accommodation lived in	
Length of time at current address	
Average total household income before tax	
Sleep duration	
Current tobacco smoking	
Past tobacco smoking	
Home second hand smoke hours week	
Not home second hand smoke hours week	
Breastfed as a baby	
Rural or urban	
Ever stressed	
Ethnic background	
Doctor diagnosed hayfever or allergic rhinitis	
Doctor diagnosed emphysema	
Doctor diagnosed chronic bronchitis	
Doctor diagnosed COPD	
Doctor diagnosed sarcoidosis	
Doctor diagnosed bronchiectasis	
Doctor diagnosed fibrosing alveolitis unspecified	
Doctor diagnosed tuberculosis	
Doctor diagnosed asbestosis	
Doctor diagnosed lung cancer not mesothelioma	
Age asthma diagnosed by doctor	
Recent medication for asthma	
Cigarette smoke in workplace	
Asbestos in workplace	
Breathing problems at work	
Deprivation	
Health	
Education	
Scotland	
Wales	

Table 6. 11 Predictors Identified for Predicting onset in Asthmatics with and without imputations.

With imputations	Without Imputations
sBP	sBP
FVC_best	FVC_best
FEV1_pred	FEV1_pred
minPEF	severity
Sex	Length of time at current address
Past tobacco smoking	Breastfed as a baby
Breastfed as a baby	Ethnic background
Maternal smoking around birth	Body mass index BMI
Rural or urban	Job Type
Ever stressed	Cigarette smoke in workplace
Ethnic background	Asbestos in workplace

Table 7. 28 Predictors Identified in Early and Late Onset Asthmatics using Non-Imputed Dataset

Early Onset Without Imputations	Late Onset Without Imputations
FVC	FVC
FEV1	FEV1
Eos	eos
dBP	dBP
sBP	sBP
FEV1_best	FEV1_best
FVC_best	FVC_best
Sex	Sex
Sleep duration	Sleep duration
Current tobacco smoking	Current tobacco smoking
Past tobacco smoking	Breastfed as a baby
Breastfed as a baby	Maternal smoking around birth
Maternal smoking around birth	Doctor diagnosed hayfever or allergic rhinitis
Doctor diagnosed hayfever or allergic rhinitis	Doctor diagnosed asthma
Doctor diagnosed emphysema	Doctor diagnosed COPD
Doctor diagnosed cystic fibrosis	Doctor diagnosed cystic fibrosis
Doctor diagnosed alpha 1 antitrypsin deficiency	Doctor diagnosed asbestosis
Doctor diagnosed lung cancer not mesothelioma	Doctor diagnosed mesothelioma of the lung
Age asthma diagnosed by doctor	Recent medication for asthma
Recent medication for asthma	Asbestos in workplace
Asbestos in workplace	Breathing problems at work
Breathing problems at work	England
England	Wales
Scotland	Icd asthma general
Wales	Icd asthma secondary
Icd asthma general	Cluster
Icd asthma secondary	Scotland
cluster	Past tobacco smoking

Table 8. Top Comorbidity on Early Onset Patients With and Without ICD-10 Asthma Diagnosis (Only within patients with valid ICD-10 records; Significant differences are labeled as red)

Phecode - Phenotype	Word Frequency Count with ICD Asthma Diagnosis (Percentage) N=1165	Prevalence in Patients with ICD Asthma Diagnosis N=1165	Word Frequency Count without ICD Asthma Diagnosis (Percentage) N=1397	Prevalence in Patients without ICD Asthma Diagnosis N=1397	Prevalence in Early Onset Patient N=2562
495 - Asthma	1165 (12.1%)	100%	-	-	45.5%
401 - Hypertension	269 (3.3%)	23.1%	338 (3.3%)	29%	23.7%
272 - Hyperlipidaemia	99 (1.2%)	8.2%	239 (2.3%)	19.1%	12.4%
411 - Angina Pectoris & Myocardial Infarction	82 (1%)	4.9%	210 (2%)	9.6%	6.6%
366 - Cataract	117 (1.4%)	8.4%	196 (1.9%)	14.3%	10.3%
250 - Diabetes	47 (0.6%)	3.9%	148 (1.4%)	10.6%	6.6%
208 - Benign Neoplasm of Colon	77 (0.9%)	6.6%	146 (1.4%)	12.5%	8.7%
530 - Digestive System Diseases	172 (2.1%)	11.8%	123 (1.2%)	8.2%	9.1%
550 - Hernia	143 (1.7%)	11.8%	107 (1%)	8.7%	9.2%
198 - Malignant Neoplasm	28 (0.3%)	1.8%	119 (1.2%)	6.5%	3.8%
278 - Obesity	46 (0.6%)	3.9%	118 (1.1%)	10.1%	6.4%
041 - Bacterial Infection NOS	26 (0.3%)	12.4%	111 (1.1%)	8.1%	4.7%
276 - Disorders of Electrolyte and Fluid Balance	13 (0.2%)	0.9%	88 (0.9%)	6.5%	3.4%
285 - Anaemia	36 (0.4%)	3.1%	85 (0.8%)	7.1%	4.6%
244 - Hypothyroidism	44 (0.5%)	3.7%	81 (0.8%)	6.9%	4.8%
535 - Gastritis and Duodenitis	71 (0.9%)	5.5%	51 (0.7%)	4.5%	4.6%
562 - Diverticular Disease of Intestine	95 (1.2%)	8.2%	59 (0.6%)	5.1%	6.0%
172 - Skin malignant Neoplasm	65 (0.8%)	5.5%	105 (1%)	8.2%	6.2%
455 – Hemorrhoids	55 (0.7%)	4.7%	72 (0.7%)	6.2%	5.0%
716 – Arthritis	62 (0.8%)	5.3%	32 (0.3%)	2.7%	3.7%
785 – Abdominal Pain	61 (0.7%)	5.2%	52 (0.5%)	4.5%	4.4%
174 - Malignant Neoplasm of Breast	45 (0.5%)	3.7%	62 (0.6%)	4.2%	3.9%
280 - Iron Deficiency Anaemia	39 (0.5%)	3.3%	73 (0.7%)	5.2%	4.4%

Table 9. Top Comorbidity on Late Onset Patients With and Without ICD-10 Asthma Diagnosis (Only within patients with valid ICD-10 records; Significant differences are labeled as red)

Phecode - Phenotype	Word Frequency with ICD Diagnosis Count (Percentage) N=4221	Prevalence in Patients with ICD Asthma Diagnosis	Word Frequency without ICD Diagnosis Count (Percentage) N=4981	Prevalence in Patients without ICD Asthma Diagnosis	Prevalence in Early Onset Patient N=9202
495 - Asthma	4221 (13.6%)	100%	-	-	45.9%
401 - Hypertension	1045 (3.4%)	24.7%	1277 (3.3%)	25.4%	25.1%
272 - Hyperlipidaemia	389 (1.3%)	8.8%	962 (2.5%)	17.7%	13.6%
411 - Angina Pectoris & Myocardial Infarction	230 (0.7%)	3.9%	836 (2.2%)	10%	7.2%
366 -Cataract	437 (1.4%)	8.9%	678 (1.8%)	11.3%	10.2%
250 - Diabetes	194 (0.6%)	4.3%	665 (1.7%)	10.7%	7.8%
208 - Benign Neoplasm of Colon	305 (1%)	7.2%	544 (1.4%)	10.9%	9.2%
530 - Digestive System Diseases	698 (2.2%)	14.2%	392 (1%)	6.5%	10%
550 - Hernia	506 (1.6%)	11.3%	271 (0.7%)	5.3%	8.1%
198 - Malignant Neoplasm	67 (0.2%)	1.4%	403 (1%)	5.2%	3.5%
278 - Obesity	271 (0.9%)	6.4%	516 (1.3%)	10.4%	8.5%
041 - Bacterial Infection NOS	109 (0.4%)	2.5%	445 (1.2%)	7.5%	5.2%
276 - Disorders of Electrolyte and Fluid Balance	62 (0.2%)	1.4%	258 (0.7%)	4.5%	3.1%
285 - Anaemia	118 (0.4%)	2.8%	374 (1%)	7.4%	5.3%
244 - Hypothyroidism	301 (1%)	6.9%	486 (1.3%)	9.2%	8.1%
535 - Gastritis	314 (1%)	6.9%	252 (0.7%)	4.4%	5.6%
562 - Diverticular Disease of Intestine	351 (1.1%)	8.3%	234 (0.6%)	4.7%	6.4%
172 - Skin Malignant Neoplasm	217 (0.7%)	4.9%	385 (1%)	7.2%	6.2%
455 – Hemorrhoids	293 (0.9%)	6.9%	254 (0.7%)	5.1%	5.9%
716 – Arthritis	217 (0.7%)	5.1%	115 (0.3%)	2.3%	3.6%
785 – Abdominal pain	242 (0.8%)	5.7%	206 (0.5%)	4.1%	4.9%
174 - Malignant Neoplasm of Breast	223 (0.7%)	5%	284 (0.7%)	5.1%	5%
280 - Iron Deficiency Anaemia	138 (0.4%)	3.2%	292 (0.8%)	5.8%	4.6%

Table 10. Unique Co-morbidity of Early and Late Asthma

Differential co-morbid features of early and late asthma within the ICD-no-asthma subset	Differential co-morbid features of early and late asthma within the ICD-asthma subset	Differential co-morbid features of early and late asthma (ICD subgroup merged)
401 - Hypertension		
366 - Cataract		
550 - Hernia		
	455 – Hemorrhoids	
	041 - Bacterial infection NOS	
	278 - Obesity	
276 - Disorders of electrolyte and fluid balance		
244 - Hypothyroidism		

Table 11. Common Co-morbidity of Early and Late Asthma

Common co-morbid features of early and late asthma within the ICD-no-asthma subset	Common co-morbid features of early and late asthma within the ICD-asthma subset	Common co-morbid features of early and late asthma (ICD subgroup merged)
	278 - Obesity	
		276 - Disorders of electrolyte and fluid balance
401 - Hypertension		401 - Hypertension
366 - Cataract		366 - Cataract
550 - Hernia		550 - Hernia
	041 - Bacterial infection NOS	
	455 – Hemorrhoids	
174 - Malignant neoplasm of breast		
785 – Abdominal pain		
280 - Iron deficiency anaemia		
272 - Hyperlipidaemia		
411 - Angina pectoris & myocardial infarction		
250 - Diabetes		
208 - Benign neoplasm of colon		
530 - Digestive system diseases		
198 - Malignant Neoplasm		
172 - Skin malignant neoplasm		
562 - Diverticular disease of intestine		
285 - Anaemia		
716 – Arthritis		
535 - Gastritis and duodenitis		

Table 12. Common Co-morbidity of Asthma

Co-morbidities of early asthma not related to ICD asthma diagnosis	Co-morbidities of late asthma not related to ICD asthma diagnosis	Co-morbidities of early asthma related to ICD asthma diagnosis	Co-morbidities of late asthma related to ICD asthma diagnosis
	401 - Hypertension		
455 –Hemorrhoids			455 – Hemorrhoids
535 - Gastritis and duodenitis			535 - Gastritis and duodenitis
785 – Abdominal pain			785 – Abdominal pain
280 - Iron deficiency anaemia			280 - Iron deficiency anaemia
174 - Malignant neoplasm of breast		272 - Hyperlipidaemia	
		411 - Angina pectoris & myocardial infarction	
		366 -Cataract	
		250 - Diabetes	
		208 - Benign neoplasm of colon	
		530 - Digestive system diseases	
		550 - hernia	
		198 - Malignant Neoplasm	
		278 - Obesity	
		041 - Bacterial infection NOS	
		276 - Disorders of electrolyte and fluid balance	
		285 - Anaemia	
		244 - Hypothyroidism	
		172 - Skin malignant neoplasm	
		562 - Diverticular disease of intestine	
		716 – Arthritis	

Table 13. FDR Corrected P-values of Chi-Square Tests on Top Comorbidity of Early and Late Onset Patients (With and Without ICD Asthma Diagnosis; Red ones are FDR 5% significant hits)

Phecode - Phenotype	P-Value of Chi-squared test between ICD and no ICD group in earlyonset	P-Value of Chi-squared test between ICD and no ICD group in late onset	P-Value of Chi-square test between early and late onset ICD group	P-Value of Chi-square test between early and late onset No ICD group	P-Value of Chi-square test between early and late onset merged group
495 - Asthma	-	-	-	-	8.17E-01
401 - Hypertension	2.52E-03	6.21E-01	4.27E-01	1.87E-02	2.59E-01
272 - Hyperlipidaemia	4.63E-14	2.03E-33	6.82E-01	3.83E-01	2.12E-01
411 - Angina pectoris & myocardial infarction	3.83E-05	5.86E-28	2.58E-01	7.87E-01	4.72E-01
366 -Cataract	2.08E-05	6.19E-04	7.42E-01	6.98E-03	9.64E-01
250 - Diabetes	1.87E-09	1.27E-28	7.13E-01	9.78E-01	9.52E-02
208 - Benign neoplasm of colon	3.95E-06	7.77E-09	6.56E-01	1.89E-01	6.21E-01
530 - Digestive system diseases	7.29E-03	9.40E-33	8.21E-02	6.57E-02	3.07E-01
550 - hernia	2.61E-02	1.34E-24	7.69E-01	1.54E-05	1.56E-01
198 - Malignant Neoplasm	6.55E-08	6.69E-22	5.61E-01	1.36E-01	6.47E-01
278 - Obesity	1.78E-08	7.99E-11	4.61E-03	8.52E-01	1.91E-03
041 - Bacterial infection NOS	1.32E-03	1.67E-25	2.38E-43	6.34E-01	4.86E-01
276 - Disorders of electrolyte and fluid balance	1.03E-16	1.63E-16	1.06E-03	7.29E-03	6.34E-01
285 - Anaemia	3.93E-05	1.75E-21	7.60E-01	8.21E-01	2.83E-01
244 - Hypothyroidism	1.62E-03	2.63E-04	3.03E-04	1.94E-02	1.08E-07
535 - Gastritis and duodenitis	4.33E-01	1.15E-06	1.89E-01	9.73E-01	1.07E-01
562 - Diverticular disease of intestine	5.47E-03	1.85E-11	9.78E-01	7.01E-01	6.34E-01
172 - Skin malignant neoplasm	2.19E-02	2.47E-05	6.21E-01	3.68E-01	1.00E+00
455 – Hemorrhoids	2.09E-01	1.07E-03	1.94E-02	2.12E-01	1.72E-01
716 – Arthritis	2.85E-03	7.79E-12	9.08E-01	6.21E-01	9.15E-01
785 – Abdominal pain	6.22E-01	1.38E-03	6.82E-01	6.82E-01	4.72E-01
174 - Malignant neoplasm of breast	7.01E-01	1.00E+00	1.48E-01	3.83E-01	5.27E-02
280 - Iron deficiency anaemia	5.27E-02	2.42E-08	9.73E-01	6.11E-01	7.93E-01

Table 14. Random Forests run for a sensitivity analysis without using lung function measurements

		Reference		
	Early Onset	Cluster 1	Cluster 2	Cluster 3
Prediction	Cluster 1	318	91	194
	Cluster 2	157	548	148
	Cluster 3	224	51	336

	Late Onset	Cluster 1	Cluster 2	Cluster 4
Prediction	Cluster 1	292	91	198
	Cluster 2	168	564	124
	Cluster 3	246	57	282

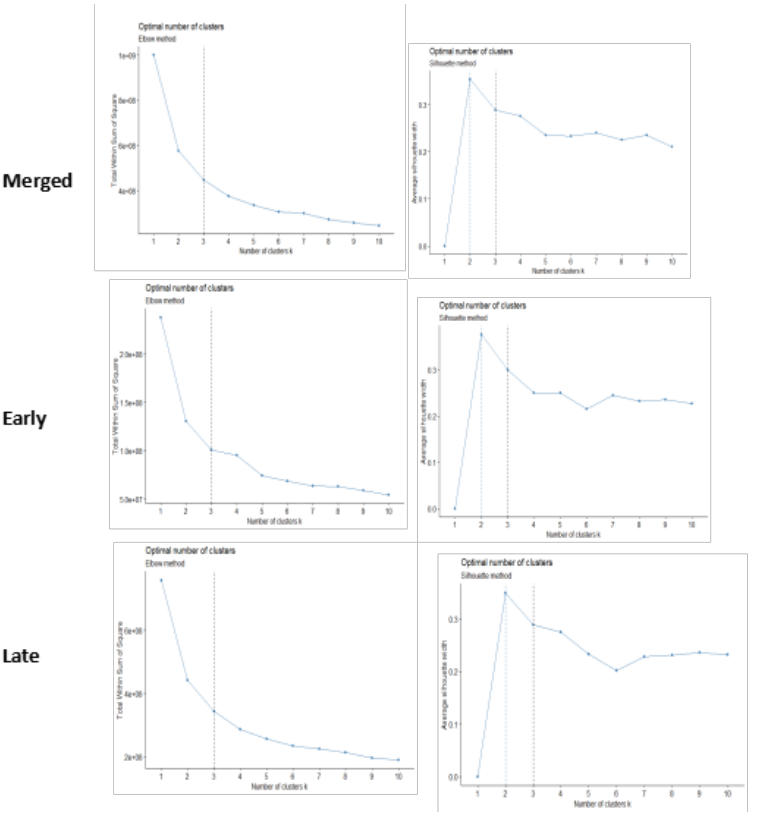

Figure 16. Cluster number selection on merged, early and late asthma onset data

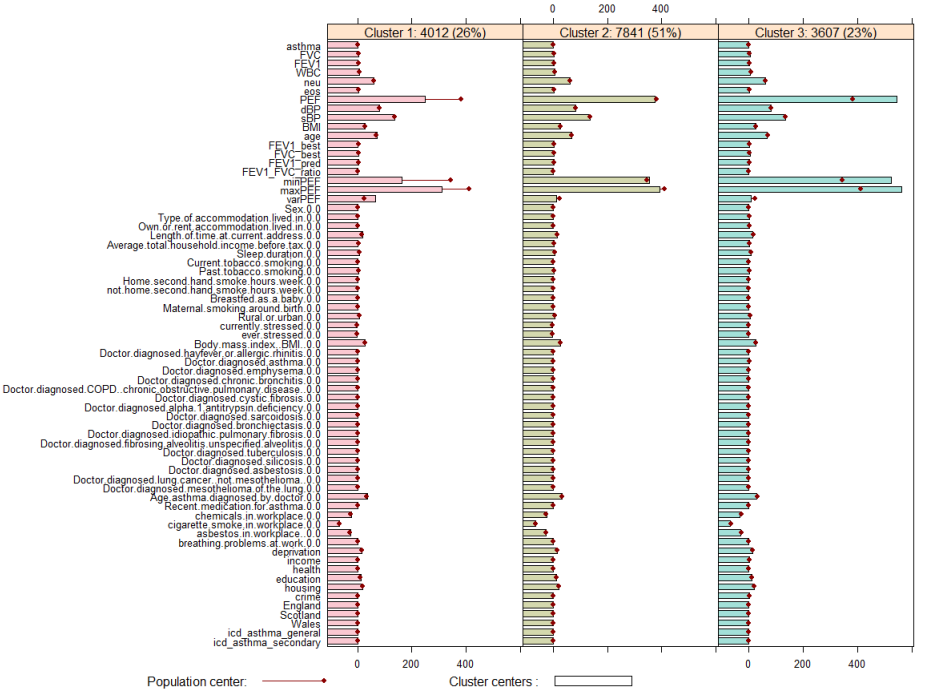

Figure 17. K-means Clustering of Merged Asthma Dataset (no difference found in asthma onset)

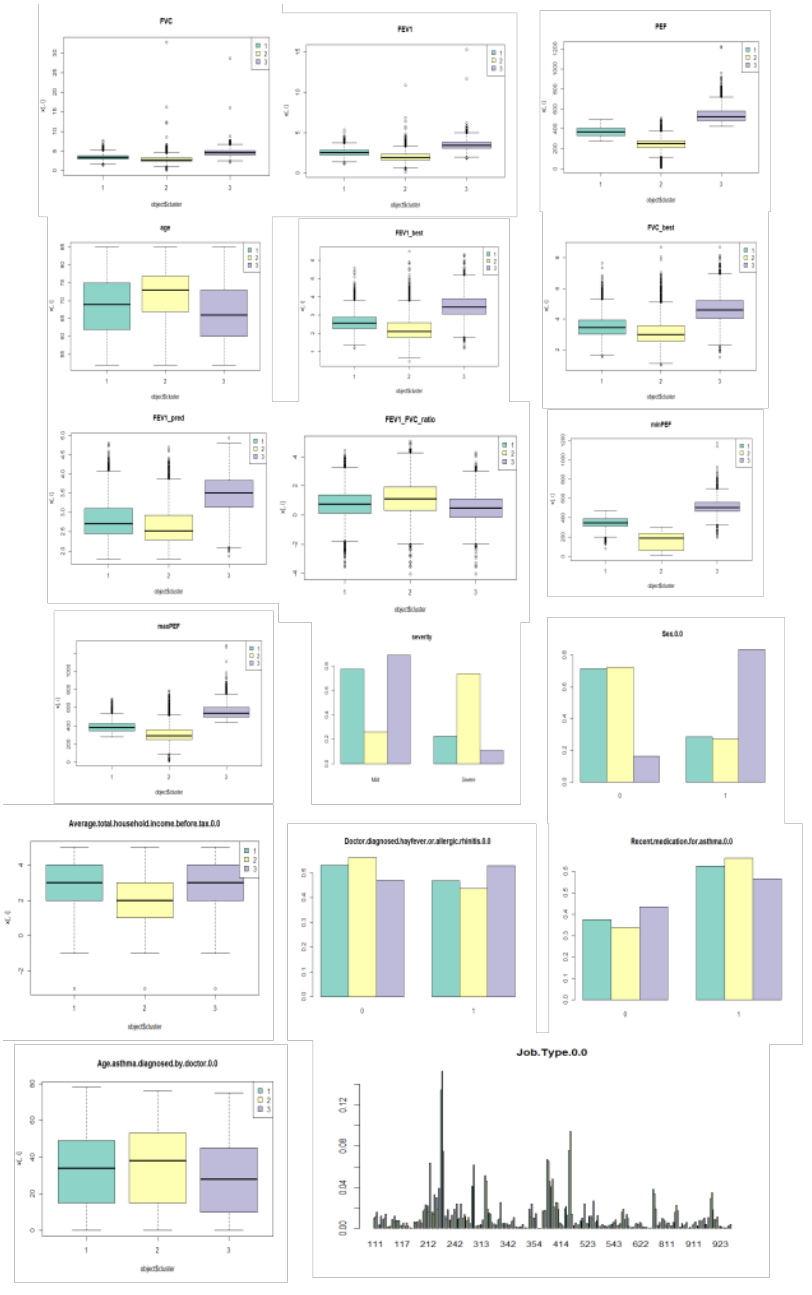

Figure 18. K-prototype Clustering of Merged Asthma Dataset (no difference found in asthma onset)

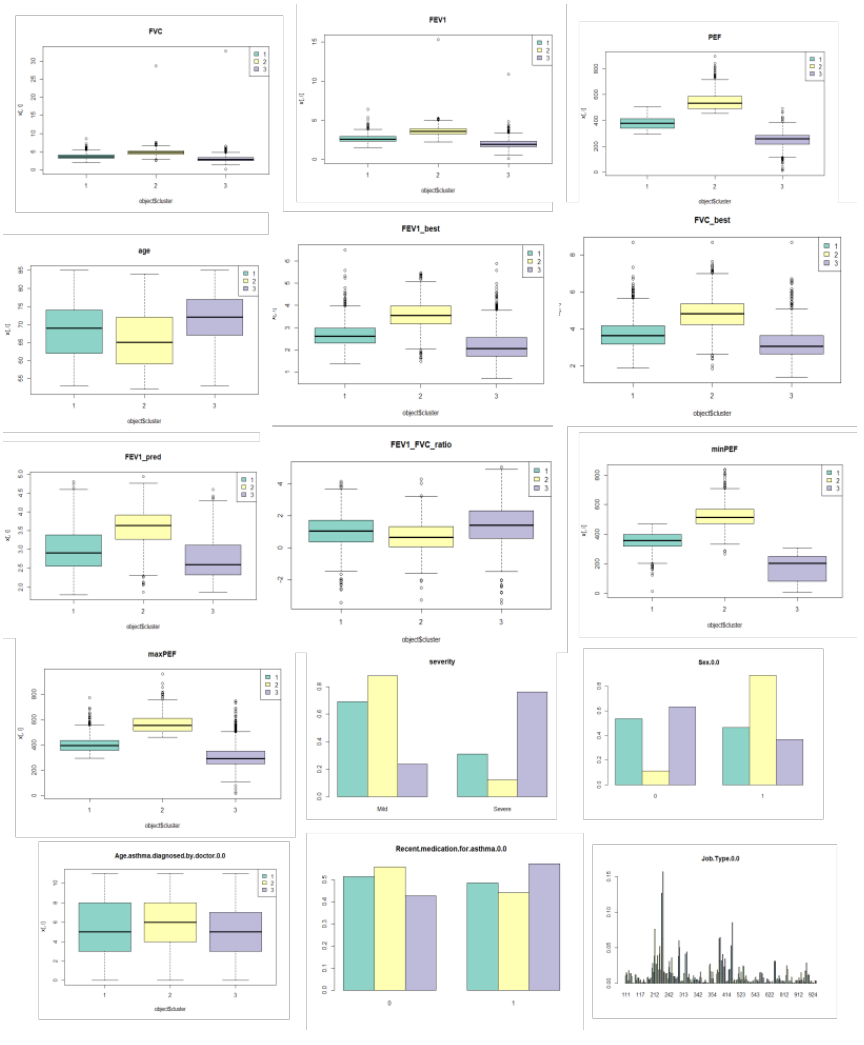

Figure 19. Key Variables in K-prototype Clustering for Early Onset Asthma Patients

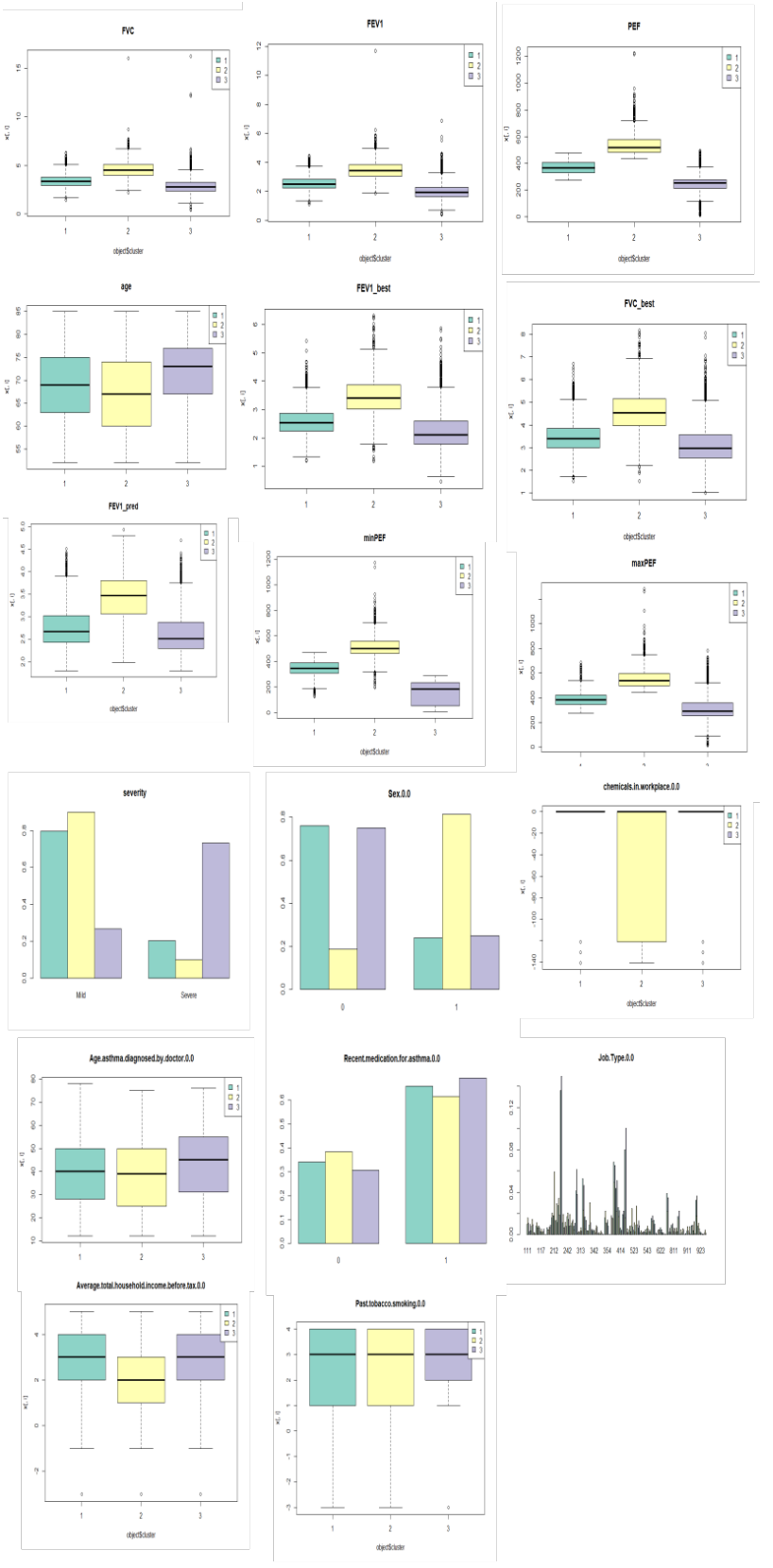

Figure 20. Key Variables in K-prototype Clustering for Late Onset Asthma Patients